PLSC 30500: Introduction to quantitative social science

Autumn term 2022, University of Chicago
Last updated August 14, 2022

Instructors:
Professor Andy Eggers (aeggers@uchicago.edu)
Professor Molly Offer-Westort (mollyow@uchicago.edu)
Teaching assistant: Oscar Cuadros (oscarcuadros@uchicago.edu)

Class meetings:

- "Lecture" meets 8:30-11:20 Fridays, Cobb Hall 302.
- "Lab" meets 12.30 (accelerated version) and 1:30 (standard lab) Fridays, Pick Hall 118.

We put "Lecture" and "Lab" in quotes because we plan to blur the distinction between the two. You will sometimes be doing hands-on data analysis in lecture and you will sometimes be listening to explanations of concepts in the lab.

Office hours:
Molly's office hours will be Wednesdays 4:30-5:30 over Zoom: https://uchicago.zoom.us/j/92934535989?pwd=Y1h5eTkxR0dRSWtJcjhQUkpQYXBQQT09
(You will need to authenticate with a UChicago account to join).

Andy's office hours are on Tuesdays 3:20-4:20, and Fridays 2:10-3:30. Reserve twenty-minute slots (at least 1.5 hours in advance) at https://calendly.com/andyeggers/office-hours.

Logistics

- Course materials are posted online on the course GitHub repository at https://github.com/UChicago-pol-methods/IntroQSS-F22.
- You do not need to use GitHub to access the files online, and we will be updating materials available throughout the quarter.
- Homework will be submitted Tuesdays at 5pm CT on the course Canvas website. Some readings will be posted on Canvas, but we will not use the Canvas website otherwise.
- We will manage questions about the course through a private course Stack Overflow team: https://stackoverflow.com/c/uchicagopolmeth. We encourage you to make your questions public, as asking and answering questions will be part of your participation grade for the class. If you are asking a question about R code, try to provide a minimal working example to help others understand your problem:
    - https://stackoverflow.com/questions/5963269/how-to-make-a-great-r-reproducible-example
    - https://stackoverflow.com/help/minimal-reproducible-example

Course description

This course introduces skills and concepts that will help students understand and produce quantitative social science research.

On completing this course, students should:

- be able to produce beautiful and informative graphics summarizing a dataset
- be able to fluently “wrangle” a new dataset into a form amenable to analysis
- have a good foundation in statistical programming using R, tidyverse, and related tools
- understand basic foundations of probability and statistics that arise in common forms of statistical inference
- understand what statistical inference means and the basics of how to do it
- understand the challenges of causal inference and how we use experiments and regression to address them

The course is designed as the first course in the political science department’s quantitative methods sequence. (It is followed by Causal Inference in the winter and Linear Models in the spring.) Later courses in the sequence build on what we teach, and we will avoid spending lots of time on topics that we know will be covered adequately in those courses. We certainly hope that this course will inspire students to continue on in the sequence. That said, we aim for the course to be useful and enjoyable to students who take no further methods courses or who take methods courses outside of our sequence.

Course philosophy

Researchers who do quantitative social science typically need a mix of different skills, including some combination of substantive expertise (i.e. knowledge of the subject of study), knowledge of statistics, programming ability, and creativity. Assembling the skills you need takes time. One nine-week course is not enough.

Unlike a lot of “intro stats” courses, we will start with what are sometimes called data science skills: visualization and data wrangling. We do this for three main reasons. First, although they are not taught in many methods sequences, these skills are indispensable for conducting social science research; we cover them first to highlight their importance and to give us time to practice them throughout the quarter. Second, we will heavily use these skills in studying the rest of the topics in the course (e.g. in doing statistical inference using the bootstrap), so it makes sense to establish a firm foundation at the beginning. Third, many students find it exciting to work with data, and we hope that the joy of learning these tools will inspire students for the rest of the quarter (including students who may not have thought of themselves as “data people”).

As just noted, we expect and hope that most students will find it exhilarating to learn to work with data in R, and that this exhilaration will motivate you through the rest of the course. The risk with our “data science first” approach is that some students will find the programming to be miserable rather than joyful. If this is you, please come see us for help.

Computing

Students will work with data using the R statistical environment.

Students will learn the basics of R, but we will also make heavy use of packages in the “tidyverse” (particularly ggplot/dplyr/tidyr, but also broom and purrr). Opinions differ on whether students should first master “base R” before moving into the tidyverse. Ideally you master everything, but time is short. We have seen sufficient evidence of the “tidyverse first” approach working well that we wanted to try it out in this course.

Perhaps you are already pretty good at tidy R programming. In that case the first two weeks may be too slow for you. If so, please come see us -- let’s come up with a way for you to work on something more appropriate to your skill level during those weeks.

Also, if you want to use base R only, or you want to use another programming approach entirely, please discuss this with us first.

You will need access to a laptop to use in and out of class. Please let us know if that is an obstacle.

We recommend working in RStudio on your own laptop. You may also use RStudio cloud, which is available for free online. If you don’t like RStudio you may also use the command line to your heart’s content.

Prerequisites

There are no formal prerequisites for the course. Students will certainly benefit from previous exposure to probability, statistics, programming, and regression analysis (and possibly a refresher at Math Camp in September), but we won’t assume much background in any of those. Concepts from calculus will be referenced, but will not be required for assignments.

Materials for the 2022 math prefresher camp for political science are available here:
http://www.tinyurl.com/plscfresh
If you did not attend, it is probably worth reviewing these materials regardless of your background experience.

If you don't have much experience with programming or think you may struggle with that aspect of the course, we recommend spending time beforehand on some of the many excellent tutorials on R/tidyverse.

References and Teaching Materials

We will draw from a range of sources throughout the course.

For programming in R, we will rely mainly on *R for Data Science* (R4DS). This book is available in a free online version.

- Wickham, Hadley and Garrett Grolemund (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media. Online book: https://r4ds.had.co.nz/

For coverage of probability and statistics, we offer students two main options that cover roughly the same material in slightly different ways.

- Our general approach is based on *Foundations of Agnostic Statistics* (Aronow & Miller, 2019, CUP). Some students find this book challenging, in part because its explanations tend to be brief and it provides few examples.
- For a version that includes lengthier explanations and more examples, use instead *Introduction to Probability* (Blitzstein & Hwang, 2019, Taylor & Francis). **Section numbers below are based on the second edition**, which is available for free viewing at probabilitybook.net. The University library also has a digital version of the first edition available; this should also cover the same material, just make sure section headings match.

Additional texts or online resources that may be useful for reference, but which will not be required for the course unless otherwise referenced:

- Data visualization
  - Wickham, Hadley, Danielle Navarro, and Thomas Lin Pedersen (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag. Online book: https://ggplot2-book.org/
  - Chang, W. (2018). *R graphics cookbook: Practical recipes for visualizing data*. O'Reilly Media. Online reference: http://www.cookbook-r.com/ (not as in-depth as the print book version)
- Probability and statistics
  - Wasserman, Larry. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013. https://link.springer.com/book/10.1007/978-0-387-21736-9

    - Goldberger, Arthur Stanley. *A course in econometrics*. Harvard University Press, 1991.
- Statistical learning & Data science
    - Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: Data mining, inference, and prediction.* Springer-Verlag, 2009. https://web.stanford.edu/~hastie/ElemStatLearn/

Assessments

Weekly problem sets 40%
Class participation 10%

- Class participation will include both in-class participation, as well as submission of questions and answers on the class Stack Overflow.

Midterm exam 20%
Final project and presentation 30%

- The final project is a chance to apply things you have learned in this class to a topic and dataset of your choice. It is very important that you *apply things you have learned in this class* and not simply do whatever you would have done before taking this class. Think of this more as a chance to demonstrate new skills and understanding than a chance to answer a research question. Further details will be made available during the course.

Collaboration and academic integrity

We encourage you to work with your classmates to understand the material in the course. (And at our private StackOverflow site we hope you will ask and answer questions.) But you should make sure that you are eventually able to do everything yourself. That means not relying too heavily on classmates.

You can work closely on **problem sets**, but you should do the write-up yourself: write your own code and write your own responses to the questions. For simple coding questions, we expect many students to have similar answers. But you won't learn to code unless you write code yourself, and you won't learn to think and write about data analysis unless you think and write for yourself.

You can confer with instructors and classmates about the **final project**, but the work should be your own. Familiarize yourself with the university's policies on academic dishonesty and plagiarism, e.g. https://studentmanual.uchicago.edu/academic-policies/academic-honesty-plagiarism/. The key idea is that you should give credit to others when you use their language and findings. If you

commit plagiarism, there could be serious consequences, including failing the course and being asked to leave the university.

Accommodations

Please reach out to the instructors directly if you would like to request accommodations for the course to better facilitate your learning. Student Disability Services (disabilities.uchicago.edu) is also available to provide you resources and support, and may provide approval for specific academic accommodations. If you or your household is affected by the ongoing pandemic in a way that affects your ability to participate in or attend class, please reach out to us as well. Informing us in a timely manner will help us to ensure accommodations are met and we are able to implement an appropriate assessment of your learning.

Schedule

Week 1, Sept 30: Introduction and programming

Class objectives, policies, and logistics (including software); data visualization; data wrangling

Reading:

- *R For Data Science (R4DS)* Chapters 3 & 4
- "The basics" primers (1.1 and 1.2; 2): https://rstudio.cloud/learn/primers
- Any additional material you need to catch up (many tutorials available – ask us if you need help finding something)

For reference:

- "The Plain Person's Guide to Plain Text Social Science", by Kieran Healy: https://plain-text.co/ Mostly for introduction, skim rest.
- Kieran Healey's *Data Visualization* (https://socviz.co/lookatdata.html). The first chapter provides useful and sensible advice on visualization, including what not to do and some research on how people perceive scientific figures. The second chapter is a gentle introduction to R.
- Wickham, "A layered grammar of graphics", https://byrneslab.net/classes/biol607/readings/wickham_layered-grammar.pdf Articulates some principles behind the ggplot2 package. (But this is not the best guide for learning ggplot2, partly because the syntax has changed.)

Notes on RStudio cloud primers:

- Primer 1 ("The Basics", https://rstudio.cloud/learn/primers/1) gives a taste of visualization and programming

- Primer 2 (“Work with data”, https://rstudio.cloud/learn/primers/2) includes tibble, select, filter, arrange, the pipe, summarize, group_by, mutate
- Primer 3 (“Visualize data”, https://rstudio.cloud/learn/primers/3) goes further with ggplot than the taster in Primer 1)
- Primer 4 (“Tidy your data”, https://rstudio.cloud/learn/primers/4) includes pivot_longer, pivot_wider, join -- note that the syntax in the instructions is outdated: they tell you to use gather instead of pivot_longer and spread instead of pivot_wider; but it does recognize the newer syntax)

Homework 1 due Wednesday 10/5 5pm (note homeworks are usually due Tuesdays)

Week 2, Oct 7: Probability

Reading for lecture:

- Aronow & Miller: Chapter 1.1-1.3 (random events, random variables, bivariate relationships)
- Blitzstein & Hwang
    - Probability and counting 1.1-1.4; 1.6
    - Conditional probability: 2.1-2.5
    - Random variables and their distributions: 3.1-3.3

Reading for programming work (homework and lab):

- *R For Data Science (R4DS)* Chapters 5 & 6, 9-13
- “The basics” primers (3, 4): https://rstudio.cloud/learn/primers

For reference

- If you haven’t seen probability recently, you can review materials from the political science math prefresher here: https://uchicago-pol-methods.github.io/polisci-math-prefresher-2022/06_probability.html

Homework 2 due Tuesday 10/11 5pm

Week 3, Oct 14: Summarizing distributions

Reading:

- Aronow & Miller: Chapter 2.1-2.2.2 (expected value, variance and standard deviation, MSE, covariance, correlation, independence)
- Blitzstein & Hwang
    - Expectation: 4.1-4.2; 4.5
    - Variance: 4.6

- Mean and median: 6.1
- Joint distributions: 7.1, 7.3

For reference

- Bueno de Mesquita and Anthony Fowler, “Correlation: What is it and what is it good for?”, Chapter 2; “Correlation requires variation”, Chapter 4

Homework 3 due Tuesday 10/18 5pm

Week 4, Oct 21: Identification and Causality

Reading:

- Holland, Paul W. "Statistics and causal inference." *Journal of the American Statistical Association* 81, no. 396 (1986): 945-960. Focus on sections 1-4, 7 & 9.
- Gerber & Green. (2012). *Field experiments : design, analysis, and interpretation*. Chapter 2.
- Lundberg et al. (2021). “What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory.” *American Sociological Review*.

For reference:

- On what can be a cause: Maya Sen and Omar Wasow, “Race as a Bundle of Sticks”, Annual Review of Political Science (2016). https://www.annualreviews.org/doi/pdf/10.1146/annurev-polisci-032015-010015
- On DAGs: Scott Cunningham, “Directed Acyclic Graphs”, Chapter 3 of *Causal Inference: The Mixtape*. https://mixtape.scunning.com/dag.html (Read through section 3.1.3; the rest is for reference.)
- Also on DAGs: Richard McElreath, Lecture 6 of *Statistical Rethinking* lecture series (winter 2019) https://www.youtube.com/watch?v=I_7yIUqWBmE -- start at 17:30
- On random assignment in experiments: Gerber & Green. (2012). *Field experiments : design, analysis, and interpretation*. Chapter 2.
- Bueno de Mesquita, Ethan and Anthony Fowler, “Causation: What is it and what is it good for?”, Chapter 3 from their book, Thinking Clearly with Data: A Guide to Quantitative Reasoning and Analysis, and “Randomized Experiments”, Chapter 11 (first half)

Homework 4 due Tuesday 10/25 5pm

Week 5, Oct 28: Estimation (1)

Note: Midterm on material in weeks 1-4.

Reading:

- Aronow & Miller, 3.1-3.3.1 (Random iid sampling, sample mean, WLLN, bias/unbiasedness, consistency, variance estimators, CLT for sample mean, plug-in principle)
- Blitzstein & Hwang, 6.3 (Sample moments), 10.2 (Law of large numbers), 10.3 (Central limit theorem)

For reference:

- Fowler and Bueno de Mesquita, “Samples, Uncertainty, and Statistical Inference”, chapter 6 of *Thinking clearly with data*.

Homework 5 due Tuesday 11/1 5pm

Week 6, Nov 4: Inference (1)

Reading:

- Aronow & Miller 3.4 (inference: confidence intervals, hypothesis testing, the bootstrap)

For reference:

- Gerber & Green. (2012). *Field experiments : design, analysis, and interpretation*. Chapter 3.
- Wasserman Chapter 8: The Bootstrap

Homework 6 due Tuesday 11/8 5pm

Week 7, Nov 11: Estimation (2)

Reading:

- Aronow & Miller:
    - Conditional expectation: 2.2.3-2.3
    - Regression estimation: 4.1, 4.3 up to 4.3.3 (inclusive)
- Blitzstein & Hwang
    - Conditional expectation: 9.1-9.2 (B&W does not cover regression)

For reference:

- Fowler and Bueno de Mesquita, “Regression for describing and forecasting“, chapter 5 of *Thinking clearly with data*.
- Wasserman Chapter 13.1-13.3 Regression

Homework 7 due Tuesday 11/15 5pm

Week 8, Nov 18: Inference (2)

Reading:

- Aronow & Miller 4.2
- Imbens, Guido. (2021). *Statistical significance, p-values, and the reporting of uncertainty.* Journal of Economic Perspectives.

For reference:

- Gelman and Loken, “The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time.” 2013

Homework 8 due 11/29 5pm

Week 9, Dec 2: Project presentations

Final project due Wednesday 12/7 5pm