PLSC 30500: Introduction to quantitative social science

Autumn term 2021, University of Chicago
Last updated September 23, 2021

Instructors:
Professor Andy Eggers (aeggers@uchicago.edu)
Professor Molly Offer-Westort (mollyow@uchicago.edu)
Teaching assistant: Steven Boyd (stevenboyd@uchicago.edu)

Class meetings:

- "Lecture" meets 9:30-10:50 Tues Thurs. (Foster Hall, Room 505)
- "Lab" meets Thurs 11-11:50 (Pick Hall, Room 506) or 1:30-2:20 (Cobb Hall 430)

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 for one-on-one 15 minute meetings Mondays 10-11 AM (sign up https://calendly.com/mollyow/office-hours), and group 30 minute meetings Thursdays 5-6 PM (sign up https://calendly.com/mollyow/office-hours-1). Sign up for a one-on-one meeting if you have a question that is specific to your project or your research, sign up for a group meeting if you have a general question or want to hear other students' general questions. Sign up at least one hour in advance of the meeting time. Office hour availability may change during the quarter.

Logistics

- Course materials are posted online on the course GitHub repository at https://github.com/UChicago-pol-methods/IntroQSS-F21.
- 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 Fridays 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 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 Linear Models in the winter and Causal Inference 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 new course.

Perhaps you are already pretty good at tidy R programming. In that case the first three 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.

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 make frequent reference to *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 will use *Foundations of Agnostic Statistics* (Aronow & Miller). You may want to purchase a physical copy of the book.

- Aronow, Peter M. and Benjamin T. Miller. *Foundations of Agnostic Statistics.* Cambridge University Press, 2019.

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 60%
Class participation 10%

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

Final project and presentation 30%

- Over the course of this class, we will ask you to apply the skills and analytical tools that you learn to analyze a dataset. Your final project will be documentation of your analysis. We have some suggestions for datasets that you may use, but you are also welcome to select your own. 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.

Accomodations

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

1.1 Sept 28: Introduction

Class objectives, policies, and logistics (including software).

Reading: none

For reference:

- "The Plain Person's Guide to Plain Text Social Science", by Kieran Healy: https://plain-text.co/ Mostly for introduction, skim rest.

1.2 Sept 30: Data visualization

Reading:

- R4DS Chapters 3 & 4
- "The basics" primers (1.1 and 1.2): https://rstudio.cloud/learn/primers/1

For reference:

- 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.)

Homework 1 due Monday 10/4 5pm
(note this is not the usual problem set deadline)

2.1 Oct 5: Data wrangling 1

Reading:

- R4DS Chapters 5 & 6
- RStudio cloud primers 2 ("Work with data", https://rstudio.cloud/learn/primers/2, which includes tibble, select, filter, arrange, the pipe, summarize, group_by, mutate) and 3 ("Visualize data", https://rstudio.cloud/learn/primers/3, which should be a review of ggplot from week 1)

2.2 Oct 7: Data wrangling 2

Reading:

- R4DS Chapters 9-13

- RStudioCloud primer 4 (“Tidy your data”, https://rstudio.cloud/learn/primers/4: 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 2 due Friday 10/8 5pm (this IS the usual problem set deadline)

3.1 Oct 12: Probability/statistics foundations (1)

Reading:
- Aronow & Miller
    - 1.1 Random events
    - 1.2: Random variables
    - 1.3 Bivariate relationships

For reference:
- Wasserman 1: Probability

3.2 Oct 14: Probability/statistics foundations (2)

Reading:
- Aronow & Miller
    - 2.1 Summary features of a random variable
    - 2.2. Summary features of joint distributions, only through 2.2.2
- R4DS 19 writing a function (note that notation for variance and skewness is for the *sample*)

For reference:
- Wasserman Chapter 2: Random Variables
- Wasserman Chapter 3: Expectation

Homework 3 due 10/15 5pm

4.1 Oct 19: Causality

Reading:
- Holland, Paul W. "Statistics and causal inference." *Journal of the American Statistical Association* 81, no. 396 (1986): 945-960.

- 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.)

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: Richard McElreath, Lecture 6 of *Statistical Rethinking* lecture series (winter 2019) https://www.youtube.com/watch?v=I_7yIUqWBmE -- start at 17:30

4.2 Oct 21: Experiments

Reading:
- 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:
- Aronow & Miller Chapter 7.1

Homework 4 due 10/22 5pm

5.1 Oct 26: Regression (1)

Reading:
- Fowler and Bueno de Mesquita, draft book chapter: “Regression for describing and forecasting“

For reference:
- Aronow & Miller, section 4.3 “Estimation of nonlinear conditional expectation functions”

5.2 Oct 28: Regression (2)

For reference:
- Frisch and Waugh, “Partial Time Regressions as Compared with Individual Trends”, *Econometrica* 1933.

Homework 5 due 10/29 5pm

6.1 Nov 2: Regression (3)

For reference:

- Cinelli and Hazlett, “Making Sense of Sensitivity: Extending Omitted Variable Bias”, *Journal of the Royal Statistical Society* 2020.

6.2 Nov 4: Regression (4)

Reading:

- *R for Data Science* chapter 14, “Many Models” https://r4ds.had.co.nz/many-models.html

Homework 6 due 11/5 5pm

7.1 Nov 9: Inference (1)

Reading:

- Aronow & Miller
    - 2.2.3 - 2.3: Multivariate generalizations (skim)
    - 3.1 Learning from random samples
    - 3.2 Estimation

7.2. Nov 11: Inference (2)

Reading:

- Aronow & Miller
    - 3.3 The plug-in principle
    - 3.4 Inference (for a random sample)

For reference:

- Wasserman Chapter 8: The Bootstrap

Homework 7 due 11/12 5pm

8.1 Nov 16: Inference (3)

Reading:

- Aronow & Miller
    - 4.2 Inference (for regression)

For reference:

- Wasserman Chapter 13.1-13.3 Regression

8.2 Nov 18: Inference (4)

More on p-values and multiple hypothesis testing.

Reading:

- TBA

For fun:

- Play around with https://projects.fivethirtyeight.com/p-hacking/

Homework 8 due 11/19 5pm

9.1 Nov. 30, and 9.2 Dec. 2

Project presentations and review

Final project due 12/7 5pm