Data Analysis for Policy Research Using R

Columbia | SIPA

Spring 2021

Instructor: Harold Stolper
Pronouns: he/they
E-mail: hbs2103@columbia.edu
OH: Walk-in (Zoom) Wed 10:30-11:30am | book indiv. appt.
Course website: https://hreplots.github.io/U6614/

Class: Tues 11am-12:50pm, online
Recitation: Thurs 9-10:50am, online
TA: Niyati Malhotra (nm3153)
TA OH: Monday TBA
(sign up via Google Sheets)

Course Description

This course will develop the skills to prepare, analyze, and present data for policy analysis and program evaluation using R. In Quant I and II, students are introduced to probability and statistics, regression analysis and causal inference. In this course we focus on the practical application of these skills to explore data and policy questions. The goal is to help students become effective analysts and policy researchers: given available data, what sort of analysis would best inform our policy questions? How do we prepare data and implement statistical methods using R? How can we begin to draw conclusions about the causal effects of policies, not just correlation?

We'll learn these skills by exploring data on a range of policy topics: COVID-19 country case data; racial bias in NYPD subway fare evasion enforcement; water shutoffs in the city of Detroit; Village Fund grants in Indonesia; and student projects on topics of your choosing.

We'll also set spend time discussing how we can use "data for good", the role of data for policy advocacy, coding identity and decolonizing data.

Course Learning Goals

We will focus on developing skills in the following areas:

- **Research design:** understanding how data structure impacts analysis and causal inference
- **Data management:** cleaning and structuring data for analysis
- **Exploratory analysis:** identifying and analyzing key factors in your analysis
- **Explanatory analysis:** estimating relationships between variables to inform policy
- **Data visualization and presentation:** conveying findings to your target audience
- **Policy writing and interpretation:** translating statistical analysis in accessible terms
- **R programming skills** (these skills support all above the areas)

Prerequisite Requirements

1. Students should have some very basic exposure to R, or a demonstrated aptitude for object-oriented programming languages.
2. Students should have completed both U6500 and U6501 (Quant I and II) or equivalent.

The prerequisites aim to ensure students can keep up with the pace of the course, and have the necessary econometric foundation to inform their work with R. Learning a new programming language requires a steady investment of time and energy: investing in this foundation from the outset allows us to explore interesting and realistic data exercises in (relatively) short order.

Required Software

The course will be taught using R, a free, open-source programming language. R has become the most popular language for statistical analysis in many circles. One advantage to using R is the thousands of open-source "libraries" created by R users. By learning R you'll be able to carry out practically any statistical method and access powerful capabilities for data collection, manipulation, and visualization. It is necessarily more complex than Stata, but far more flexible.

We'll be working with R using RStudio. Instructions on installing R + RStudio can be found at https://stat545.com/install.html. ***Please install both R and RStudio on your laptop prior to our first class session.***

Additionally, ***please activate Piazza notifications in advance of the first class meeting*** by clicking on Piazza in the course site on Courseworks.

Course Structure and Approach to Learning R

Course Structure

1. **Asynchronous (pre-class) lessons** will be shared via the course website with the expectation that students work through them independently in advance of class. Each class meeting will begin with a short CW quiz on this asynchronous content. The idea is to introduce key concepts and syntax in R, as well as methodological issues, to prepare for in-class discussion and coding exercises. This asynchronous content will take the form of web-based lessons (html files) including sample code and output that students can replicate on their own as they go. Asynchronous lessons for Tuesday's class session will be posted by the previous Thursday.
2. **In-class workshop-style instruction using R** will take up the majority of our in-class time together. We'll be working through R code together using RStudio to prepare and explore data for analysis. This will include a mix of reviewing pre-filled code line by line, and short exercises for students to arrive at their own coding solutions. We will also spend considerable time for in-class discussion about how we can use R code to craft and implement a research design with appropriate econometric methods–"why" and "when", not just "how" to work with data using R.
3. **Five weekly data assignments and short write-ups ("data memos")** which will require you to expand on the work we do together in class and write up your work using clear, accessible language. We will introduce R Markdown as a tool for you to write up your work and present code and findings in a single document. Data memos will be due before midnight on Mondays, in advance of Tuesday's class session.

4. **A data project** of students' choosing (with instructor approval) to be conducted in consultation with the teaching team and presented and submitted towards the end of the semester. Students are required to work in ***groups of two***. The project will require you to use R to explore a policy-relevant research question with readily available data. It must focus on analyzing the effect of at least one independent variable of interest on some relevant outcome variable, though the majority of work you do will involve data cleaning, manipulation, and exploratory data analysis to inform the specification of appropriate statistical models. In the latter half of the semester, student groups are ***required*** to sign-up for three individual meetings with the instructor and TA to discuss project progress.
5. **A course discussion board** where students can post reactions to asynchronous lessons and ask questions/comments to share with classmates and the teaching team. If you're stuck or experiencing problems with R more generally, odds are others are too. Posting questions and concerns allows us all to benefit from each others knowledge. We'll be using **Piazza** via Courseworks for our online discussion. When asking R questions on Piazza, please include as many details to replicate the "error" (if applicable), insert code, screenshots, and text to your posts. The teaching team will do our best to reply within 48-72 hours, but we encourage students to reply to other posts with their own insights. Thoughtful Piazza contributions will count towards your overall class participation grade.
6. **Recitation and office hours.** During recitation time, the TA will review code to reinforce the material for the week, typically building on the same data exercises to help prepare students for the assignment. Later in the semester, recitation will transition to group OH focused on support for student projects, in addition to regular TA office hours each week. The instructor will hold group/walk-in OH on Wednesday mornings, and individual office hours by appointment via Calendly.

Approach to Learning R

Our approach will emphasize "learning by doing" by working through R code together in class to explore data. Lecture content will introduce key concepts in advance of class workshop time, to prepare us for the workshop exercise. Assignments will task you with refining and expanding the code from in-class workshop exercises, putting your new knowledge to work.

It will take us some time to build up the skills to effectively explore messy, real-world data. Learning a new programming language can be overwhelming, and this class is only the beginning. The goal of this course is not to become proficient in the sense of memorizing all the commands you think you will need, but rather to understand the basics of R syntax and develop the comfort level to explore new functionality and troubleshoot on your own.

Online resources and coding "cheat sheets" will be shared periodically, but learning how to find and employ answers from both within RStudio and using Google will be some of your most valuable resources.

Virtually all course materials will be generated using RStudio, posted to the course website along with source code for you to consult for examples of useful syntax.

The following open-source textbooks are good supplementary learning resources that we'll rely on throughout the course:

- Bryan, J. (2018). *STAT 545: Data wrangling, exploration, and analysis with R*. Retrieved from https://stat545.com.

- Grolemund, G., & Wickham, H. (2018). *R for Data Science*. Retrieved from http://r4ds.had.co.nz.
- Hanck et al (2020). *Introduction to Econometrics with R*. Retrieved from https://www.econometrics-with-r.org/index.html.
- Xie, Y., Allaire, J. j., & Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Retrieved from https://bookdown.org/yihui/rmarkdown.

Data Community

In-class exercises and discussion are designed to foster a data community where students can interact among themselves and with the teaching team to share ideas. Data and coding obstacles generally feel less overwhelming when you can exchange ideas with others. The Courseworks discussion board will also help us collectively interact around data and coding issues and learn from each other.

Assignments, Grading and Course Requirements

Five Weekly Assignments (Data Memos) (30% of your overall grade – 5 x 6)

Weekly assignments are due by midnight on Monday night. Assignments will be graded on a check plus/minus scale. Late submissions will not receive a grade as we will be discussing solutions during class.

Individual Student Projects and Required Meetings (50%)

Your project grade will include an-class presentation of your work to-date near the end of the semester (20% of your total grade), and a short report (30% of your total grade). The data project will also involve three required meetings with the teaching team for project advising, and include several intermediate deliverables: (1) submitting research ideas; and (2) and a proposal with summary statistics. Intermediate deliverables will not receive their own grade, but late intermediate submissions will result in a one grade deduction from your overall project grade for every day late (e.g. from an A to A-).

Courseworks Quizzes on Asynchronous (Pre-Lecture) Material (10%)

CW quizzes at the beginning of scheduled class meetings will account for 10% of your total grade. These quizzes will consist of multiple choice questions, and are designed to encourage you to engage with the asynchronous (pre-class) lessons in advance of class so we can put new R functionality to work in class and focus on application and discussion.

In-class Participation (10%)

Students are required to attend weekly class sessions, with webcams turned on. You're expected to participate in the weekly class sessions and discussion, and participate in the Piazza discussion board in Courseworks. This component can make the difference between an A and B for your overall course grade, for example, so please come to class prepared and ready to participate.

Attendance

Attending synchronous class sessions in Zoom is required, **with your webcam turned on**. Multiple unexcused absences may result in additional deductions to your overall course grade beyond any deductions for forgone participation.

That said, we encourage you to communicate with the teaching team about any extenuating personal circumstances you may be dealing with. We're happy to make any accommodations we can to help minimize stress and anxiety as we all navigate the current learning environment.

Course Policies

Virtual Classroom Environment

SIPA's greatest asset is the diversity of students, but it also means being mindful that what we say affects others in ways we may not fully understand. Learning R and trying to get a handle on unfamiliar data can feel overwhelming at times. It's important that we all help create an environment where students feel comfortable asking questions and talking about what they don't understand.

After registration closes, community guidelines for Zoom participation will be set with student input.

Towards an Anti-Racist Learning Experience

Every class should be an anti-racist class, even when the subject matter is broadly oriented. In this class we'll cover examples that reflect systemic gaps based on race, ethnicity, immigration status, and gender identity, among other aspects of personal identity. Given our focus on statistical methods, we are limited in the time we can spend discussing all of the policy context contributing to these gaps. It is critical to acknowledge that the social and economic marginalization reflected in the data is rooted in systemic oppression that upholds opportunity for some at the expense of others. We should all be thinking about our own role in upholding these systems. Over the course of the semester, we'll engage in discussions to help challenge our own notions about how to use data for good.

Teaching Team Communication and Student Support

Given the large of number of student inquiries over a virtual environment, we ask that you rely on scheduled office hours and the Piazza discussion board as much as possible. The instructor will hold group office hours that are open to all without an appointment, as well as individual appointment slots that you can book in advance at https://helloharold.youcanbook.me. We'll do our best as a teaching team to respond to inquiries within 72 hours.

While late submissions will not be accepted out of fairness, we understand many of us are dealing with a great deal of stress and uncertainty right now. If you are experiencing unexpected challenges that are affecting your ability to meet your course obligations, I encourage you to reach out to the teaching team in advance of any looming deadlines.

Academic Integrity

SIPA does not tolerate cheating or plagiarism in any form. Students who violate the Code of Academic & Professional Conduct will be subject to the Dean's Disciplinary Procedures. Please consult the code of conduct here.

While grading your assignments, if we come across answers to parts of any assignments that are clearly not your own words, all involved parties will receive a zero for those parts and may be referred to Academic Affairs if appropriate.

Disability Accomodations

SIPA is committed to ensuring that students registered with Columbia University's Disability Services (DS) receive the reasonable accommodations necessary to fully participate in their academic programs. The teaching team will work with SIPA's DS liaison to make sure the necessary accommodations are provided. You are encouraged to make an appointment with the instructor to discuss any concerns you have about your accommodations.

Course Schedule

The syllabus is subject to change at the discretion of the instructor with proper notice to the students. Students are likely to have varying levels of statistical knowledge and experience with R. Because it is difficult to anticipate the optimal pace for students in this class, the following schedule should be treated as a guide. Topics may carry-over into the following week(s), and we may end up cutting/adding/re-ordering later topics based on student needs and interest.

Week 1, 1/12/2021: Introduction, R Basics and Workflow

- In-class data: gapminder
- *Assignment 1 posted after class: due before 11:59pm on Monday, 1/18*

Week 2, 1/19/2021: Data Types & Structures, R Markdown, Intro to the Tidyverse

- In-class data: country-level COVID-19 case data
- *Assignment 2 posted after class: due before 11:59pm on Monday, 1/25*

Week 3, 1/26/2021: Importing, Cleaning & Summarizing Data, Intro to ggplot

- In-class data: Brooklyn subway fare evasion arrest data part 1: cleaning microdata
- *Assignment 3 posted after class: due before 11:59pm on Monday, 2/1*

Week 4, 2/2/2021: Joins, Aggregation, Inference & Regression

- In-class data: Brooklyn subway fare evasion arrest data part 2: aggregation to subway station-level observations
- *Assignment 4 posted after class: due before 11:59pm on Friday, 2/12*
- *Assignment: post reaction to Categorizing Identity reading to Piazza by Monday, 2/8*

Week 5, 2/9/2021: Research Design, Regression Analysis, Weighting

- In-class data: Brooklyn subway fare evasion arrest data part 3: poverty, race and enforcement
- *Project deliverable: Find a partner for your data project by Tuesday, 2/16*

Week 6, 2/16/2021: Data Visualization with ggplot

- In-class: review and discussion of Assignment 4, and project timeline
- *Project deliverable: Submit 2 possible research questions by Friday, 2/19*
- *Sign up for required project meeting #1 w/instructor during Week 7 (meet by Friday, 2/27)*

Week 7, 2/23/2021: Working with Panel Data (Part 1)

- In-class data: Detroit water shutoffs, disparate impact by race/income
- *Assignment 5 posted after class: due before 11:59pm on Monday, 3/1*
- *Project deliverable #2: Mini-proposal with summary statistics due after Spring Break, by 11:59pm on Friday, 3/12*

Week 8, 3/2/2021: Spring Break

Week 9, 3/9/2021: Working with Panel Data (Part 2)

- In-class data: Detroit water shutoffs, public health impacts
- *Sign up for required project meeting #2 with instructor (meet by Friday, 3/19)*

Week 10, 3/16/2021: Working with String & Date Variables

- In-class data: Indonesian Village Fund data
- *Sign up for required project meeting #3 with TA (meet by Friday, 3/26)*

Week 11, 3/23/2021: Mapping with R

- In-class data: Indonesian Village Fund data

Week 12, 3/30/2021: Data Visualization Principles, Data Project Workshop

Week 13, 4/6/2021: Final Presentations

Week 14, 4/13/2021: Final Presentations

Final papers due 4/18/2021