Data Analysis for Policy Research Using R

Columbia | SIPA

Fall 2020

Instructor: Harold Stolper
Pronouns: he/they
E-mail: hbs2103@columbia.edu
OH: Group/walk-in Tues 5:30-6:30pm | book indiv. appt.

Class: Tues 2:10-4pm, online
Recitation: Thurs 6:10-8pm, hybrid
TA: Kevin Wibisono (kw2870)
TA OH: TBD

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 on your own. 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 cases; racial bias in NYPD subway fare evasion enforcement; the distribution of Village Fund grants in Indonesia; US police shootings; wage gaps by gender/race; and student projects on topics of your choosing.

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.

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 R Studio. Instructions on installing R and R studio can be found at https://stat545.com/install.html. Please install both R and R studio on your laptop prior to our first class session.

Course Structure and Approach

Course Structure

This course will primarily consist of:

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. The idea is to introduce key concepts and syntax in R, as well as methodological issues, to prepare for in-class discussion and data exercises. This asynchronous content will often take the form of web-based lessons (html files) including sample code and output that students can try out on their own as they go. In some weeks asynchronous materials will also focus on policy-relevant examples of data analysis tools and challenges, and may include short readings. Each class will begin with a short Zoom quiz on this asynchronous content to encourage engagement in advance of class. This will be followed by some in-class time for discussion to engage with the material, often using Zoom breakout rooms and Poll Everywhere. Pre-class lecture content 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 most weeks. We'll be working through R code together using R Studio to prepare and explore data for analysis.
3. **Four 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 may 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 an appropriate model. In the latter half of the class, student groups are ***required*** to sign-up for three individual meetings meet with the instructor and TA to discuss project progress.
5. **A course discussion board** where students can ask homework 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. When asking questions on Courseworks, 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 you are all encouraged to share your thoughts/answers on posts by your classmates. Writing out explanations to student questions will improve your own knowledge and benefit your classmates. Thoughtful contributions will also count towards your overall class participation grade.

6. **Recitation and office hours.** During recitation time, the TA will review student questions about the material introduced that week and hold group office hours. The TA will also hold individual office hours each week, with a particular focus on assisting students with their projects in the second half of the semester.

The instructor will hold both group/walk-in office hours, and individual office hours by appointment.

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 each week, but learning how to find and employ answers from both within R Studio and using Google will be among your most valuable resources.

When applicable, chapters from the following open-source resources will be listed as supplementary learning resources:

- 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.
- 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

Four Weekly Assignments (Data Memos) (40% - 4 x 10)

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-).

Attendance, Zoom Quizes, and Participation (10%)

Students are required to attend weekly class sessions, which will being with a short Zoom quiz of the material assigned for the day. You're expected to participate in the weekly class sessions and discussion, share responses to Poll Everywhere questions when appropriate, and participate on the Courseworks discussion board. Zoom quizzes will account for 5% of your total grade, all other participation 5%.

Course Policies

Virtual Classroom Environment

We all have a responsibility to ensure that every member of the class feels valued and safe, and to express our ideas in a way that doesn't make people feel excluded. 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 (if there is interest, we can make time for more discussion!). But 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.

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 Courseworks 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, 9/8/2020: Introduction, R Basics and Workflow

- In-class data: gapminder
- *Assignment 1 posted after class: due by midnight on Monday, 9/14/2020*

Week 2, 9/15/2020: Data Types & Structures, R Markdown, Intro to the Tidyverse

- In-class data: country-level COVID-19 case data
- *Assignment 2 posted after class: due by midnight on Monday, 9/21/2020*

Week 3, 9/22/2020: Importing, Cleaning & Summarizing Data, Intro to ggplot

- In-class data: Brooklyn subway fare evasion arrest data
- *Assignment 3 posted after class: due by midnight on Monday, 9/28/2020*

Week 4, 9/29/2020: Joins, Data Structure & Research Design, Inference & Regression

- In-class data: Brooklyn subway fare evasion arrest data
- *Assignment 4 posted after class: due by midnight on Monday, 10/5/2020*

Week 5, 10/6/2020: Data Validation, Survey Data, Weighting, Exploratory Data Analysis

- In-class data: Wage gaps in the American Community Survey via IPUMS
- *Project deliverable #1: Submit 2 possible research questions by 10/16 at 5pm.*

Week 6, 10/13/2020: Data Visualization with ggplot

- In-class data: US police shootings
- *Required meeting #1 with instructor*

Week 7, 10/20/2020: Panel Data Methods

- In-class data: Detroit water shutoffs and public health

Week 8, 10/27/2020: More Data Cleaning, Working with String & Date Variables

- In-class data: Indonesian Village Fund data
- *Project deliverable #2: Mini-proposal with summary statistics due by midnight on 10/30.*

Week 9, 11/3/2020: NO CLASS - ELECTION DAY

Week 10, 11/10/2020: Looping

- In-class data: TBD
- *Required meeting #2 with Instructor*

Week 11, 11/17/2020: Mapping with R

- In-class data: TBD
- *Required meeting #3 with TA*

Week 12, 11/24/2020: Data Visualization Principles, Data Project Workshop

Week 13, 12/01/2020: Final Presentations

Week 14, 12/08/2020: Final Presentations

Final papers due 12/18/2020 (tentative)