U6614: Data Analysis for Policy Research Using R

Data Projects:
Timeline, Overview, Expectations

Spring 2021

Instructor: Harold Stolper

Data project overview

- Students will work in groups of 2 to use readily available data to inform a policy-relevant research question of their choosing, presenting their work and write up findings towards the end of the semester.
    - Topics will be approved by the instructor in a required initial meeting.
    - Ongoing guidance from the teaching team through 2 additional required meetings.
- Projects must focus on **analyzing the effect of at least one independent variable of interest on some relevant outcome variable**.
    - The majority of the work you do will involve data cleaning, manipulation, and exploratory data analysis to inform the choice of appropriate statistical methods.
    - Early exploratory analysis will inform your choice of regression specifications.

Timeline

DAY/WEEK	DELIVERABLES
Today	Find a partner
Fri, Feb 19th by 11:59pm	Project deliverable: Submit 2 possible research questions
Mon, Feb 22nd – Fri, Feb 26th	Required meeting #1 with instructor
Fri, March 12th by 11:59pm	Project deliverable: Mini-proposal with summary statistics
Mon, Mar 15th – Fri, Mar 19th	Required meeting #2 with instructor
Mon, Mar 22nd – Fri, Mar 26th	Required meeting #3 with TA
Mar 30th, April 6th, April 13th	PRESENTATIONS (in-class)
Sun, April 18th	Final Report due

How much data manipulation should be done outside of R?

- **Zero!** 🙅‍♀️

No data manipulation in Excel!

- Why not? With R you can produce a step-by-step record of your work.

Deliverable 1: Submit 2 possible research questions by Friday, Feb 19th

Each mini "proposal" should address the following (<1 page for each). If you have a first choice, list it first.

- **State research question(s).** What policy, program, characteristics or behavior do you want to learn about.
- **Why do we care?** Think about policy relevance and what you hope to learn (i.e. not just "we care about the environment)."
- **Describe potential data source(s):**
  - You'll be using some source of sample variation in X to estimate its effect on Y: describe what information you're looking for and what data sources will give you this information.
  - For your input data: state the unit of observation, representative population, key limitations.
- **Describe your empirical strategy, i.e. outline the analysis you will do to answer your research question(s)**
  - This will evolve, but think about what comparisons you want to explore and what the data will likely support.
    - Think about internal validity and what it means for your project.
  - Example: The public defender arrest data (input data) is a cross-section of individual arrests records that we can aggregate to subway station-level observations (analysis data). We can then exploit variation across subway stations in neighborhood characteristics like racial composition to explore how fare evasion enforcement intensity varies across neighborhoods. This will allow us to estimate racial disparities in enforcement while controlling for other neighborhood characteristics.

Deliverable 2: Proposals due by Fri, March 12th at 11:59pm

Submitted a knitted R Markdown file with 5 sections (including output, and where applicable):

1. **State research question(s)**
2. **Motivation**
   - What do you hope to learn from the proposed analysis? What are the policy implications? What policy context should we know?
   - Are the (potential) mechanisms linking your policy variable(s) of interest and outcomes clear to the reader?
3. **Summary statistics**
   - Define key variables (outcomes, policy/treatment variables) and any control variables that may be important to account for (explain why)
   - Show descriptive stats to summarize the distribution of these variables (using charts and/or tables), e.g. describe variation over time and/or between relevant groups
   - Make sure to describe sample variation in your policy variables of interest and outcomes.
4. **Describe your empirical strategy**
   - Carefully describe the analysis you plan to do. Think about research design and the policy variation you'll investigate in your regressions.
   - Outline key steps to prepare the data for analysis (data cleaning, recoding, merging, appending, aggregation, etc.).
   - Highlight key limitations you will need to address – be specific!
5. **Appendix**
   - Include your coding work to-date for importing, cleaning, recoding, restructuring and joining input data sources.

NOTE: use code chunks to generate summary stats, otherwise don't clutter your write-up w/code (you can include more in an Appendix).

Project presentations

- In-class presentations will be 20% of your total course grade
- It's not intended to be a *final* presentation, but a presentation of your analysis to date and chance to get feedback (before written reports are due)
- Presentation plan:
  - *March 30th*: 2 presentations (8 minutes to present, 6 minutes for Q&A)
  - *April 6th*: 6 presentations (7 minutes to present, 6 minutes for Q&A)
  - *April 13th*: 6 presentations (7 minutes to present, 5 minutes for Q&A)
- **Attendance for all of your classmates' presentations is <u>required</u>**

Final reports

- Final reports will be 30% of your total course grade
- More details on submission guidelines and grading will be shared later in the semester

IPUMS: a useful tool for accessing survey data

- IPUMS USA
  - *IPUMS USA collects, preserves and harmonizes U.S. census microdata... Data includes decennial censuses from 1790 to 2010 and American Community Surveys (ACS) from 2000 to the present.*
- IPUMS CPS
  - *IPUMS CPS harmonizes microdata from the monthly U.S. labor force survey, the Current Population Survey (CPS), covering the period 1962 to the present.*
- IPUMS International
  - *IPUMS-International is dedicated to collecting and distributing census data from around the world.*