U6614: Data Analysis for Policy Research Using R

Data Projects: Overview, Timeline, Expectations

Fall 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 *down the road*.

Timeline

DAY/WEEK	DELIVERABLES
Today	Find a partner by Tuesday, Oct. 19th
Fri, October 22nd by 11:59pm	Project deliverable: Submit 2 possible research questions
Mon, Oct 25th – Fri, Oct 29th	Required meeting #1 with instructor
Fri, Nov. 12th by 11:59pm	Project deliverable: Mini-proposal with summary statistics
Mon, Nov 15th – Fri, Nov 19th	Required meeting #2 with instructor
Mon, Nov 22nd – Mon, Nov 29th	Required meeting #3 with TA
Dec 7th (class) Dec 9th (recitation)	PRESENTATIONS
Thurs, Dec 16th	Policy 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, October 22nd

Each mini "proposal" should address the following (<1 page for each). If you have a 1st 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, likely 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.

Deliverable 1: Submit 2 possible research questions by Friday, October 22nd

- **Describe your empirical strategy, i.e. outline the analysis you will do to answer your research question(s)**
  - 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.

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 schedule:
  - *Dec. 7th*: 6 presentations during class time
  - *Dec. 9th*: 6 presentations during recitation time
- **Attendance for all of your classmates' presentations is <u>required</u>**

Policy 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.*