U6614: Data Analysis for Policy Research Using R

Data Projects: Overview, Timeline, Expectations

Spring 2022

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 *Required Project Meeting #1*.
  - Ongoing guidance from the teaching team through 2 additional required meetings (one with instructor and one with your assigned TA)

Data project overview

- Projects will 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*.
    - Your ultimate goal isn't to estimate regressions, but rather to carry out empirical analysis that informs your research question… regressions are one important tool you will use later on.

Timeline

DAY/WEEK	DELIVERABLES
By next class, Tues. Feb 22nd	Find a partner
Fri, Feb. 25th by 11:59pm	Project deliverable: Submit 2 possible research questions
Mon, Feb 28th – Fri, March 4th	Required meeting #1 with instructor
Fri, March 25th by 11:59pm	Project deliverable: Proposal with summary statistics
Mon, March 28th – Fri, April 1st	Required meeting #2 with instructor
Mon, April 4th – Fri, April 15th	Required meeting #3 with TA
April 19th, 21st & 26th	PRESENTATIONS
May 6th	Policy report due

Finding a partner

- Find 1 project partner *from your own class section* (morning/afternoon).
  - This will make it easier when we set aside class time for project support.
  - Every group will receive project support from the TA assigned to their lecture section (in addition to the instructor).
- How to find a partner:
  - Talk to your classmates!
  - We'll make an *Ed* post for you to your share your project interests later this week.
  - If you're looking for a partner you should post and/or reply to others.
  - For people without partners we'll continue the discussion during class next week
  - Everybody should have a partner *by the end of next Tuesday, Feb. 15th*.
- Once you've settled on a partner, please enter both partner names here.

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.
  - Getting comfortable with R is one of our course goals!

Deliverable 1: Submit 2 possible research questions by Friday, Feb. 25th

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").

Deliverable 1: Submit 2 possible research questions by Friday, Feb. 25th

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 data sources you will use:**
  - You'll be using some source of sample variation in X to estimate the 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.

Deliverable 1: Submit 2 possible research questions by Friday, Feb. 25th

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 data sources you will use:**
  - You'll be using some source of sample variation in X to estimate the 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.
  - i.e. think about internal validity and what it means for your project:
    - Will you be able to distinguish between correlation and causation, or at least uncover policy-relevant correlations? What challenges will you have to address in order to do this?

Tips for an interesting and informative project

- Projects will evolve in unpredictable ways based on findings from your exploratory analysis.
- No matter where you end up, **you'll need meaningful variation in both X and Y variables**.
  - You can't estimate the relationship between X and Y if either one doesn't vary much in your sample!
  - *Exploring sample variation in your candidate X and Y variables is one of the first things to prioritize!*
  - A cautionary example:
    - One recent group used a country-year panel to explore the effect of educational attainment on violence against women.
    - But neither educational attainment or their measures of violence against women varied much over time within countries!
    - This type of question would have been better served by exploiting sharper policy changes (e.g. a big new laws) that impacted the prevalence of violence against women, and more targeted measures of the violence against women that exhibited more variation (along with more disaggregated data, either individual microdata, or sub-national observations).

Deliverable 1: Submit 2 possible research questions by Friday, Feb. 25th

- **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 and poverty rates (X) to explore how fare evasion enforcement intensity (Y) varies across neighborhoods. This will allow us to estimate racial disparities in enforcement while controlling for other neighborhood characteristics.

Deliverable 1: Submit 2 possible research questions by Friday, Feb. 25th

- **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 and poverty rates (X) to explore how fare evasion enforcement intensity (Y) varies across neighborhoods. This will allow us to estimate racial disparities in enforcement while controlling for other neighborhood characteristics.
  - Note that if there isn't much variation across subway station areas in either X or Y, this analysis wouldn't be informative!

Deliverable 1: Submit 2 possible research questions by Friday, Feb. 25th

- **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 and poverty rates (X) to explore how fare evasion enforcement intensity (Y) varies across neighborhoods. This will allow us to estimate racial disparities in enforcement while controlling for other neighborhood characteristics.
  - Note that if there isn't much variation across subway station areas in either X or Y, this analysis wouldn't be informative!
  - Student projects often rely on fixed effects or difference-in-differences, but instrumental variables or regression discontinuity is also fine.

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:
  - *April 19th and April 26th during class*
- **Attendance for 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