Causal Inference

Mixtape Session

Roadmap

- Hidden curriculum
  - Background
  - Empirical workflow
  - Hierarchical folder structure
  - Naming conventions
  - Version control
  - Soft skills

Data Scarcity vs Data Abundance

- Think about it: for most of the history of science, and really all the social sciences, **there was practically no data** – only theories about data
- Today we drown in data. Ordinary workers have more data than they know what to do with
- As such causal inference cannot be easily separated from the work itself, which means programming skill – both in the analysis stage, but maybe more importantly the data wrangling and cleaning stage

Code and Software

- For causal inference we need:
  - → data
  - → software for data
  - → understanding of statistics and causality
  - → skill using software, cleaning and analyzing data, applying our models and interpreting our results
- But which software? SAS, SPSS, Eviews, **R, Stata, python**, julia and more
- Language agnostic programming principles that are necessary but not covered in econometrics courses (“hidden curriculum”)

Making mistakes

- Once upon a time there was a boy who wrote a job market paper using the NLSY97.
- This boy presented the findings a half dozen times, spoke to the media a few times, got 17 interviews at the ASSA, 7 flyouts, and an offer from Baylor
- He submitted the job market paper to the *Journal of Human Resources*, a top field journal in labor, and received a “revise and resubmit” request from the editor (woo hoo!)

Coding error

- But then digging into his one directory, he found countless versions of his do file and hundreds of files with random names
- And once he finally was able to get the code running again, he found a critical coding error that when corrected (“destroyed”) his results
- The young boy was devastated and never resubmitted which he does not recommend (but he was sad!)

All competent empirical work is a mousetrap

“Happy families are all alike; every unhappy family is unhappy in its own way.” - Leo Tolstoy, Anna Karenina

“Good empirical work is all alike; every bad empirical work is bad in its own way.” - Scott Cunningham, This slide

Cunningham Empirical Workflow Conjecture

- The cause of most of your errors is **not** due to insufficient knowledge of syntax in your chosen programming language
- The cause of most of your errors is due to a poorly designed empirical workflow

Workflow

Wikipedia definition:

> *“A workflow consists of an orchestrated and repeatable pattern of activity, enabled by the systematic organization of resources into processes that transform materials, provide services, or process information.”*

Dictionary definition:

> *“the sequence of industrial, administrative, or other processes through which a piece of work passes from initiation to completion.”*

Empirical workflow

- Workflow is a fixed set of routines you bind yourself to which when followed identifies the most common errors
  - → Think of it as your morning routine: alarm goes off, go to wash up, make your coffee, check Twitter, repeat *ad infinitum*
- Finding the outlier errors is a different task; empirical workflows catch typical and common errors created by the modal data generating processes

Why do we use checklists?

- Before going on a trip, you use a checklist to make sure you have everything you need
  - → Charger (check), underwear (check), toothbrush (check), passport (oops), . . .
- The empirical checklist is solely referring to the intermediate step between “getting the data” and “analyzing the data”
- It largely focuses on ensuring data quality for the most common, easiest to identify, situations you’ll find yourself in

The Checklist

- Empirical workflows are really just a checklist of actions you will take before analyzing your data
- It is imperative that you do not analyze your data (e.g., “explore the data”, “run some regressions”) until you have checked everything off
- Your checklist should be a few simple, yet non-negotiable, programming commands and exercises to check for coding errors
- These are some of mine – feel free to add your own