DIGITAL WORKFLOWS

PhD Breakfast, 27/2/2024

Johannes Boehm

Sciences Po

SOME VIEWS ON MANAGING YOUR (DIGITAL) WORKFLOW

First, three guiding principles on (digital) workflow

- Very subjective
- What works for me may not work for you (and vice versa)
- Some are in contradiction to each other

Then, some rambling on other topics

- ML/LLMs
- Open Source software
- My setup
- …

1. USE WHATEVER SOFTWARE GETS THE JOB DONE

- Slides: If you have a brilliant paper and you do you slides in PowerPoint, that's ok

  Also: if you have amazing slides but a bad paper, you won't get any points

- Stata/R/Python/Julia/whatever:

  Your main job as a researcher is to produce research.

  If you're quicker cleaning your data in Excel, go for it

  If you're quicker cleaning your data in Stata, go for it

  If you're quicker writing your estimation code in Fortran, go for it

- These slides are made with PowerPoint

2. BE FORWARD-LOOKING

- Learning new tools expands your PPF

  You’ll be more productive in the future

  You’ll be able to do things that others cannot do

- Not being able to use some tools can prevent collaboration

  People are picky with their tools

- Some tools force you work “cleanly”

  Replicability

  Projects take long. You’ll want to understand what you did in 6 years’ time.

3. AS PROJECTS GET LARGER, ORGANIZATION MATTERS MORE

- Github for managing code & assignment of tasks to RAs

  Ideally, use Continuous Integration to avoid breaking stuff somewhere else

- Email is not ideal, but at least your messages won't disappear (looking at you, Slack!)

- Optimal organization depends on people

  People that are not up to speed on technology won't contribute

  In my view, regular (oral) conversations among co-authors very important

MY SETUP

Stata for data cleaning & linear regressions, Julia for structural stuff

- Stata has very convenient syntax
- Julia is fast (and elegant), and will only get better

Ideally would like to do everything from Julia

- Perhaps soon: Tidier.jl (Tidyverse syntax), Douglass.jl (Stata syntax)

OPEN SOURCE SOFTWARE

Nice way to give back to the community while learning stuff

Much shorter development cycle than papers: days to weeks (satisfying!)

Don’t think “I can’t do that”. Most scientific software is written by people like us.

RegressionTables.jl, Douglass.jl, GLFixedEffectModels.jl, Fastcluster.jl, KeplerGL.jl

Contributions welcome!

SOME THOUGHTS ON “MACHINE LEARNING”

Conceptually, ML nothing new

It’s just semiparametric statistical models

But: amazing progress on engineering: hardware (GPU) and software (autodiff)

Relevance for social sciences research: the things you can do with it

Text and image analysis

“Stuff that you would have otherwise had to use many undergrad RAs for”