Code and Data for Empirical Studies

Zi Wang

HKBU

January 7, 2025

Empirical Studies

- Gentzkow and Shapiro (2014): "empirical studies"
  - Asking good questions.
  - Digging up novel data.
  - Designing statistical analysis.
  - Writing up results.
- What do we do most of the time?
  - Writing and debugging code.
- Most of us do not have formal training in computer science.

Problems

- In trying to replicate the estimates from an early draft of a paper, we discover that the code that produced the estimates no longer works because it calls files that have since been moved. When we finally track down the files and get the code running, the results are different from the earlier ones.

Problems

- In trying to replicate the estimates from an early draft of a paper, we discover that the code that produced the estimates no longer works because it calls files that have since been moved. When we finally track down the files and get the code running, the results are different from the earlier ones.

Problems

- In trying to replicate the estimates from an early draft of a paper, we discover that the code that produced the estimates no longer works because it calls files that have since been moved. When we finally track down the files and get the code running, the results are different from the earlier ones.

- In the middle of a project we realize that the number of observations in one of our regressions is surprisingly low. After much sleuthing, we find that many observations were dropped in a merge because they had missing values for the county identifier we were merging on. When we correct the mistake and include the dropped observations, the results change dramatically.

Problems

- In trying to replicate the estimates from an early draft of a paper, we discover that the code that produced the estimates no longer works because it calls files that have since been moved. When we finally track down the files and get the code running, the results are different from the earlier ones.

- In the middle of a project we realize that the number of observations in one of our regressions is surprisingly low. After much sleuthing, we find that many observations were dropped in a merge because they had missing values for the county identifier we were merging on. When we correct the mistake and include the dropped observations, the results change dramatically.

Problems

- In trying to replicate the estimates from an early draft of a paper, we discover that the code that produced the estimates no longer works because it calls files that have since been moved. When we finally track down the files and get the code running, the results are different from the earlier ones.

- In the middle of a project we realize that the number of observations in one of our regressions is surprisingly low. After much sleuthing, we find that many observations were dropped in a merge because they had missing values for the county identifier we were merging on. When we correct the mistake and include the dropped observations, the results change dramatically.

- A referee suggests changing our sample definition. The code that defines the sample has been copied and pasted throughout our project directory, and making the change requires updating dozens of files. In doing this, we realize that we were actually using different definitions in different places, so some of our results are based on inconsistent samples.

Expertise from CS and Data Science

Much of the time, when you are solving problems with code and data, you are solving problems that have been solved before, better, and on a larger scale.

Automation

- Rules:
  - Automate everything that can be automated.
  - Write a single script that executes all code from beginning to end.

Example

- We wish to test the hypothesis that the introduction of television to the US increased sales of potato chips.

Example

- We wish to test the hypothesis that the introduction of television to the US increased sales of potato chips.
  - Excel sheet 1: “tv” which contains for each county in the US the year that television was first introduced.

Example

- We wish to test the hypothesis that the introduction of television to the US increased sales of potato chips.
  - Excel sheet 1: "tv" which contains for each county in the US the year that television was first introduced.
  - Excel sheet 2: "chips" which contains total sales of potato chips by county by year from 1940 to 1970.

Example

- We wish to test the hypothesis that the introduction of television to the US increased sales of potato chips.
  - Excel sheet 1: "tv" which contains for each county in the US the year that television was first introduced.
  - Excel sheet 2: "chips" which contains total sales of potato chips by county by year from 1940 to 1970.

Example

- We wish to test the hypothesis that the introduction of television to the US increased sales of potato chips.
  - Excel sheet 1: "tv" which contains for each county in the US the year that television was first introduced.
  - Excel sheet 2: "chips" which contains total sales of potato chips by county by year from 1940 to 1970.

- "Interactive" model of research:

Example

- We wish to test the hypothesis that the introduction of television to the US increased sales of potato chips.
  - Excel sheet 1: "tv" which contains for each county in the US the year that television was first introduced.
  - Excel sheet 2: "chips" which contains total sales of potato chips by county by year from 1940 to 1970.

- "Interactive" model of research:
  - Open the file in Excel and use "Save As" to save the worksheets as text files.

Example

- We wish to test the hypothesis that the introduction of television to the US increased sales of potato chips.
  - Excel sheet 1: "tv" which contains for each county in the US the year that television was first introduced.
  - Excel sheet 2: "chips" which contains total sales of potato chips by county by year from 1940 to 1970.

- "Interactive" model of research:
  - Open the file in Excel and use "Save As" to save the worksheets as text files.
  - Open up a statistical program like Stata, and issue the appropriate commands to load, reshape, and merge these text files.

Example

- We wish to test the hypothesis that the introduction of television to the US increased sales of potato chips.
  - Excel sheet 1: "tv" which contains for each county in the US the year that television was first introduced.
  - Excel sheet 2: "chips" which contains total sales of potato chips by county by year from 1940 to 1970.

- "Interactive" model of research:
  - Open the file in Excel and use "Save As" to save the worksheets as text files.
  - Open up a statistical program like Stata, and issue the appropriate commands to load, reshape, and merge these text files.
  - Define a new variable to hold logged chip sales, and issue the command to run the regression.

Example

- We wish to test the hypothesis that the introduction of television to the US increased sales of potato chips.
  - Excel sheet 1: "tv" which contains for each county in the US the year that television was first introduced.
  - Excel sheet 2: "chips" which contains total sales of potato chips by county by year from 1940 to 1970.

- "Interactive" model of research:
  - Open the file in Excel and use "Save As" to save the worksheets as text files.
  - Open up a statistical program like Stata, and issue the appropriate commands to load, reshape, and merge these text files.
  - Define a new variable to hold logged chip sales, and issue the command to run the regression.
  - Open a new MS Word file, copy the output from the results window of the statistical program into a table.

Example

- We wish to test the hypothesis that the introduction of television to the US increased sales of potato chips.
  - Excel sheet 1: "tv" which contains for each county in the US the year that television was first introduced.
  - Excel sheet 2: "chips" which contains total sales of potato chips by county by year from 1940 to 1970.

- "Interactive" model of research:
  - Open the file in Excel and use "Save As" to save the worksheets as text files.
  - Open up a statistical program like Stata, and issue the appropriate commands to load, reshape, and merge these text files.
  - Define a new variable to hold logged chip sales, and issue the command to run the regression.
  - Open a new MS Word file, copy the output from the results window of the statistical program into a table.
  - Write up an exciting discussion of the findings, and save. Submit to a journal.

Why don't we like the interactive mode?

- Replicability.
  - Because there is no record of the precise steps that were taken, there is no authoritative definition of what the numbers in our paper actually are.

- Efficiency.
  - In a real project, there might be a thousand steps from raw data to final results. For each of these, there could be several alternatives, detours, and experiments that were tried and discarded. Each step is typically run hundreds of times as the analysis is developed and refined.

Automation

- Turn every step into a piece of code.

```
chips.csv        mergefiles.do          tv_potato_submission.pdf
cleandata.do     regressions_alt.do     tv_potato.tex
extract0B.xls    regressions_alt.log    tv.csv
fig1.eps         regressions.do         tvdata.dta
fig2.eps         regressions.log
figures.do       tables.txt
```

- Store the information about the order in which the steps are run.

```
---- rundirectory.bat ----
stattransfer export_to_csv.stc
statase -b mergefiles.do
statase -b cleandata.do
statase -b regressions.do
statase -b figures.do
pdflatex tv_potato.tex
```

Replicability

- The rundirectory.bat script works like a roadmap, telling the operating system how to run the directory.
- Unlike a ReadMe file with notes on the steps of the analysis, rundirectory.bat cannot be incomplete, ambiguous, or out of date.
- We can now delete all of the output files in the directory - the .csv files, the .log and .eps files, tables.tex, the .pdf - and reproduce them by running rundirectory.bat.
- A system shell provides a more natural interface for calling commands from multiple software packages, and for operating system commands like moving or renaming files.
- A rule of research is that you will end up running every step more times than you think.

Version Control

- Rules:
  - Store code and data under version control.
  - Run the whole directory before checking it back in.

```
cleandata_022113.do      cleandata_022613.do      regressions.log
cleandata_022113a.do     cleandata_022613_jms.do  regressions_022413.do
chips.csv                tvdata.dta               regressions_022713_mg.do
regressions_022413.log
```

“Data and initial” is poor

- Pain.
- Confusion.

Version Control Software

- Track successive versions of a given piece of code.
  - You set up a "repository" on your PC.
  - Every time you want to modify a directory, you "check it out" of the repository.
  - After you are done changing it, you check it back in.
  - The software remembers every version that was ever checked in.
  - When you change your mind, you ask the software for a history of changes to the directory. The version control software automatically records who authored every change.
  - It maintains a single, authoritative version of the directory at all times.
  - Version control is like an undo command for everything.
  - Example: Git; BitBucket.

Directories

- Rules:
  - Separate directories by function.
  - Separate files into inputs and outputs.
  - Make directories portable.

Separate directories by function

```
---C:/build---
/input
    extract0B.xls

/code
    rundirectory.bat
    export_to_csv.stc
    mergefiles.do

/output
    tvdata.dta

/temp
    chips.csv
    tv.csv
```

```
---C:/analysis---
/input
    tvdata.dta (link to C:/build/output)

/code
    rundirectory.bat
    regressions.do
    regressions_alt.do

/output
    fig1.eps
    fig2.eps
    tables.txt

/temp
    regressions.log
    regressions_alt.log
```

Keys

- Store cleaned data in tables with unique, non-missing keys.

county	state	cnty_pop	state_pop	region
36037	NY	3817735	43320903	1
36038	NY	422999	43320903	1
36039	NY	324920	.	1
36040	.	143432	43320903	1
.	NY	.	43320903	1
37001	VA	3228290	7173000	3
37002	VA	449499	7173000	3
37003	VA	383888	7173000	4
37004	VA	483829	7173000	3

county	state	population
36037	NY	3817735
36038	NY	422999
36039	NY	324920
36040	NY	143432
37001	VA	3228290
37002	VA	449499
37003	VA	383888
37004	VA	483829

state	population	region
NY	43320903	1
VA	7173000	3

Abstraction

- In programming, turning the specific instances of something into a general-purpose tool is known as abstraction.
  - Abstract to eliminate redundancy.
  - Abstract to improve clarity.
  - Otherwise, do not abstract.

Documentation

- Do not write documentation you will not maintain.
- Code should be self-documenting.