Best Practices

Erick F Molina

Summer, 2024

ITAM

Motivation

- We want to conduct reliable research $\Rightarrow$ minimize human errors.
- We need to be organized in everything we do to avoid forgetting things or doing the same task twice.
- We want our research to be replicable for it to have validity.
- If someone wants to continue our research line... could they start from where we left off?

The Goal

- Since researchers make constant changes, it's better to automate as much as possible.
- This will help because we can add new datasets at any time.
- We should always follow a *workflow*.
- Document everything we do when processing the datasets.
- We may likely forget what we did a while ago.
- Assign understandable and consistent names to everything that needs naming.

Folder Organization

- Don't put off for tomorrow what you can do today, delaying tasks has a cost.
- Each project should have its own folder. This way, we'll maintain an organized structure and avoid confusing datasets between different projects.
- Each project should be divided into different subfolders.
- Each subfolder can have sub-subfolders, but remember that everything must be clear and consistent.

Folder Organization

I suggest using a folder structure similar to the following:

1. **data** This folder contains all the datasets.
    - 1.1 **raw** This folder contains datasets exactly as we obtained them.
    - 1.2 **processed** This folder contains intermediate datasets. These are datasets we've manipulated but are not yet ready for final econometric analysis or creating any table or figure.
    - 1.3 **final_data** This folder contains final datasets that will no longer undergo any changes.

Folder Organization

2. **scripts** This folder will contain all the code files that modify datasets or create figures/tables.

   2.1 **dofiles** This folder will contain only Stata dofiles.

   2.2 **rscripts** This folder will store only files with the .R extension.

   - It's important to have a master file for the dofiles and R scripts. A master file is a file that runs all other files.
   - It's also important to name the dofiles and R scripts with numbers to know their natural order. Example:
     - 00_master.do
     - 01_cleaning.do
     - 02_balance_table.do

Folder Organization

3. **results** This folder contains the tables, figures, and numbers produced from analyzing the datasets.
   - 3.1 **tables** This folder contains the regression tables we create.
   - 3.2 **figures** This folder contains the graphs and maps we may create.

More Best Practices

- Creating dictionaries and codebooks is helpful. In Stata, for example, you can add labels to variables explaining what each one means.
- At the start of each dofile or R script, add a header indicating the most relevant information. For example, the file name, author, and purpose.
- Comment on as many lines of code as possible. Even though many lines are understandable just by reading them, the purpose of some may not be clear later.
- Create a ReadMe file inside the project folder.

ReadMe File

- A **ReadMe** file is a file that describes everything in a folder.
- Include a brief description of the project.
- List the folders and files within the project.
- Include a reference for contacting the author in case of questions.

Workflow Organization

It's recommended to follow a certain order to build a project. Specifically, it's very useful to:

- Start by observing the datasets:
  - How does the dataset look in a program like R or Stata?
  - What are the variables, and what do they mean?
  - What is the level of observation?
- Read the ReadMe file, if available.

Workflow Organization

- Identify missing values.
- Clean and create the variables you deem necessary.
- If you're going to merge datasets, ensure it's done correctly.
- Identify the types of variables you will work with.
- You can create preliminary graphs or tables to analyze any anomalies in the data.

How to Seek Help

- First, use the help functions in the software in question.
- If that doesn't help, use Google. It's very likely that someone has already asked the same question you have.
- There are famous websites like R-bloggers, Quick-R, and StackOverflow where programming questions are answered.