Causal Assignment 2

Hannah Jones

2/9/2021

Causal Inference Assignment 2

Gentzkow and Shapiro

1. Chapter 2-8 Summary Chapter 2 emphasizes how automating the work-flow of a research project is important in yielding reproducible results. By automating the process, you can reduce the number of steps taken to reproduce results, avoid issues associated with moving data around, and any issues that result from changing research teams. A good way of automating is by creating a directory that can instruct how the data/code should be run through.

Chapter 3 emphasizes the importance of version control. The traditional way of adding a date or editor gets very clunky and confusing over time. Doing automated version control keeps track of who changed what, and maintains one authoritative version. Additionally, this process enables you to roll back any unwanted changes.

Chapter 4 walks through the creation and management of directories. This is useful to create separate processes by function (for example, one for cleaning data and one for analysis). Organizing in this way enables collaborators to change analysis without having to re-clean/merge the data.

Chapter 5 emphasizes the importance of a logically designed database. The database should have a key that uniquely identifies each element and can never be missing. The raw data should be preserved in normalized files to maintain the original data. Then, the data should be transformed from the raw data and combined as necessary. Finally, any additional information should be merged and any other transformations done.

Chapter 6 shows the value of abstraction in writing code. Abstraction refers to the process of generalizing a formula and writing it once, then feeding the required set of variables to the formula when needed. This enables clean code and eliminates ‘copy-and-pasting’ for formulas that will be run multiple times on different variables.

Chapter 7 discusses how to determine the right amount of documentation to provide clarity but not introduce errors/confusion. Documentation does not have to be updated for code to run, so it is easy to update code without updating the relevant notes. To avoid this issue, the author suggests using as little documentation as possible, and writing clear code that is easily updated to substitute for the flat documentation.

Chapter 8 introduces a better method for task management through using a task management system as opposed to emails and other one-off comment methods. Task management specifically associated with a task and assigned to a collaborator is a better way of keeping track of the status of tasks.

2. Why do Genztkow and Shapiro think these elements of modern empirical work are so important? What problems does each element solve The authors think these elements are so important to eliminate mistakes, wasted time and confusion in empirical analysis. Automation solves the problem of manually running code and the mistakes that may come from understanding what to run and in what order. Version control eliminates the clunky-ness of using file names to update and keep track of edits. Directories enable the team to have separate chunks of data manipulation/analysis for different functions, which eliminates the need to run the data through the whole gamut of code for every little change to the data or code– it only has to be re-run through the relevant directory. Keys help generalize the code and create a logical flow to a database which solves issue with confusing and missing values. Abstraction enables cleaner code and prevents calculation issues associated with running the same calculations on multiple variables. Effective documentation prevents contradicting documentation associated with updating code without updating notes. Finally, effective task management through specific task-management software eliminates the ambiguity of one-off email-centric task management.

3. Give an example of the sort of problem that could arise in the course of an empirical project if someone were to fail to adopt these principles. If someone were to fail to adopt the principles of automation and version control, the team may end up with a cluttered repository of relevant project files, with no clear flow or relationship. The team may run analysis on the wrong or incomplete datasets, or the code may not work due to naming conventions. Additionally without effective documentation and task-management, the code may become confusing and contradictory, without a clear understanding of whose responsibility it is to keep it clean.

4. How do you plan to incorporate these solutions into your own work? I will use the Git extension in R studio to update and maintain my workflow. I will aim to let go of old habits and let this new, more efficient and effective way of managing my files become the operating norm.

Git and GitHub

5. Git and github are used for interacting with code and other project files. Git is a distributed version control system. Github works with git and hosts files and eases the interaction with git.
6. Using git for empirical research enables easier collaboration and organization as you go through many versions of data and code. It also eases reproducibility. If you don't use git, it is easy to drown in irregular and confusing version control that over-complicates making necessary changes and improvements over the life of a project.
7. Git is great! So far, I have not figured out how to collaborate with others on github. I have pushed my own changes through git, but would like to learn more about collaboration. To do this, I will work with a classmate to explore how it is done.
8. The four main git operations are: Stage, commit, pull and push. Stage tells Git that you want to add changes to the repo history. Commit tells git to make these changes a part of the repo history. Pull retrieves any changes made on the GitHub repo (either by me or other collaborators). Push tells to push any committed changes to the repo.
9. Link to "Titanic" Repository with a Read Me: https://github.com/hannahjonesut/Titanic
10. Link to "Titanic" Repository: https://github.com/hannahjonesut/Titanic
11. Causal Inference Class clone found : https://github.com/hannahjonesut/causal-inference-class