Reproducible Data Analysis Workflow

Boyi Guo

Department of Biostatistics
University of Alabama at Birmingham

Thursday, July 16, 2020

Who Am I?

Who Am I?

- Rising 4th-year Ph.D. student in BST @ UAB
- Dissertation: Bayesian high-dimensional data analysis (Genomics)
- Background:
  - B.S. in Computer Science & Stat @ UIUC
  - M.S. in Statistics (Analytic track) @ UIUC
  - Experienced R programmer (8-year)
  - "Ridiculously awesome" commented by a REGARDS collaborator

Who would benefit?

Who would benefit?

- Analysts who use R
- Analysts who use SAS
- Any people who collaborates with analysts

Why Is It Important?

Why Is It Important?

- **Reproducibility**: *exact* replication of analysis results
- **Comprehensibility**: easy for other analysts to understand and apply
- Efficiency: update analysis results faster
- Tracking Changes: know what you have done and changed
- Reducing Copying Error

What Is It About?

What Is It About?

Purpose

- To review some available strategies/tools
- To create a chance for awareness and open discussion

How (in 3 levels)

How (in 3 levels)

- Beginner
  - Naming Conventions
  - Computational reproducibility
- Intermediate
  - Automation
  - Version Control
- Advanced
  - Dynamic Report Generation
  - Containerization

Beginner

Naming Conventions

- File System Structure
  - (prepped) Data
  - Code
  - Manuscript
- File Naming Convention
  - Order as prefix
  - Content
  - Date as suffix (preferably _yymmdd for searchability)
  - e.g.: 00_Data_Prep_200713.SAS
- Coding Style
  - SAS Style
  - R Style

Computational Reproducibility

- Document software version (and package version for R)
  - Save your computational environment (Advanced)
- Set random seeds for the pseudo random number generator when applicable
  - Bootstrapping, Random Forest, non-deterministic algorithms

Intermediate

Automate Table 1

- How to start:
  - SAS: %ggBaseline
  - R: tibbleOne
- Tip:

  Copy HTML tables to Word without losing the format

Version Control System

- What is version control system: softwares that helps record changes to files by keeping a track of modifications done to the code.
  - Cloud file sharing platforms (Box, Dropbox, . . .)
  - Git (GitHub, which is a remote platform for sharing projects)
- Benefits
  - File history with comments
  - Less files: No more dated files
  - Collaboration: No overwriting
- How to start:
  - git -the simple guide

Advanced

Dynamic Report Generation (R user)

- What is dynamic report generation: Integration of manuscript writing and analysis reporting.
  - Imagine word document that runs analytic software
  - R Markdown, R Packages: `officer`, `flextable`
- Benefits:
  - Automation
  - Less chance to make copying error
- How to start:
  - R Markdown

Containerization

- What does containerization means: saving the current computational environment
- Benefits:
  - Cross operational platform
  - Protect against system/software updates
- How to start:
  - Docker with SAS
  - Docker with R

Conclusion

Conclusion

- The importance of reproducible analysis work flow
- Review some of tools
- My experience:
  - An investment of time and effort
    - Hard to set up the first time
    - Benefit long-term
  - Needs practice