Workflow and Research Transparency: Examples in Stata and R

Matt Ingram, Nakissa Jahanbani, and Cesar Rentería
mingram@albany.edu
Thursday Policy Lunch (TPoL)
University at Albany, SUNY
Sep. 7, 2017

What is “workflow”?

- Scott Long: “a coordinated framework for conducting data analysis”
- Consists of your complete work process:
  - From planning, organizing, and documenting your research
  - Through data collection, generation, cleaning, storage, management
  - To analysis, presentation of results, and publication
- Everybody has a workflow
  - You might not think about it very much, or it might be unstructured or ad hoc
  - Goal: become more explicit, mindful, structured
- Note: when researchers talk about “workflow”, they generally mean the process of documenting your entire research process so that it can be reproduced/replicated in future

Matt's workflow (then and now)

Matt 1.0 (grad school)

1) Analysis in Stata
   - clean/reorganize data, generate new variables, rename, run analysis

2) Write text in Word

3) Integrate steps 1-2 by:
   - Copying and pasting
     - tables
     - figures
   - At times, typing by hand

4) If anything happened before I was done (mistake, new theory, new method, new data, etc.), had to repeat steps 1-3, and it was often hard to recreate exactly how I had done something

Matt 2.0

1) Analysis in R/Stata
   - Clean/reorganize data, generate new variables, rename, run analysis

2) Write text in same document

** Integration is automatic (dynamic; i.e., document changes as data/code changes) **

3) If anything happens before I am done (mistake, new theory, new method, new data, etc.), repeat step 1, modifying any writeup, and can recreate exactly how I did something

Matt's workflow (then and now)

Matt 1.0 (grad school)

1) Analysis in Stata
   - clean/reorganize data, generate new variables, rename, run analysis

2) Write text in Word

3) Integrate steps 1-2 by:
   - Copying and pasting
     - tables
     - figures
   - At times, typing by hand

4) If anything happened before I was done (mistake, new theory, new method, new data, etc.), had to repeat steps 1-3, and it was often hard to recreate exactly how I had done something

Matt 2.0

1) Analysis in R/Stata
   - Clean/reorganize data, generate new variables, rename, run analysis

Tools: Stata, R/RStudio (Python, SAS, …)

2) Write text in same document
Tools: LaTeX, markdown, HTML
** Integration is automatic (dynamic; i.e., document changes as data/code changes) **

Tools: MarkDoc in Stata;
knitr or Sweave in R

3) If anything happens before I am done (mistake, new theory, new method, new data, etc.), repeat step 1, modifying any writeup, and can recreate exactly how I did something

Why should you care?

- Recall
  - You can pick up a project days/weeks/months (years?) later and know what you were doing
- Collaboration
  - Good workflow makes it easier to work with others (co-authorship, teamwork)
- Replication
  - Good workflow → transparent, reproducible research (good science)
- Accuracy
  - Get right answers
  - Good workflow makes error detection – and error correction – easier
- Efficiency
  - Good workflow saves time
  - Good workflow → modularity (ability to use portions of one project on a different project), so future projects are easier and more efficient
- Funding
  - Research transparency is increasingly required for grant proposals (e.g., NSF's data management plans), or can make a good proposal a great proposal; some funders require pre-registration of workflow
- Employment
  - Research transparency is increasingly the norm professionally; potential employers notice
- Publication
  - Many journals now require complete replication materials ahead of publication; some even agree to publish based only on a pre-registered workflow prior to data collection

Motivation 1: Values

- Clarity
- Credibility
- Legitimacy
- Honesty
- Transparency
- Scientific community

Motivation 2: Goals and Interests

- Efficiency
- Communication
- Funding
- Employment
- Publication
- Replication (scientific cumulation of knowledge via access to raw materials: data, production methods, and analytic methods)

On replication: only feasible way to "establish and confirm evidence-based claims" in social science

(Bill Jacoby , ICPSR 2017)

DA-RT

From Arthur (Skip) Lupia, ICSPR 2017

- Data Access
  - Current practice: “mine!”
  - Aspiration: default is to provide access unless exceptional circumstances (e.g., vulnerable population)
- Research Transparency
  - Production Transparency
    - How did you produce your data?
      - Including case selection, sampling, questionnaire design, documentation, recording, etc.
  - Analytic Transparency
    - How did you analyze your data?
      - Including software, computing code, models, etc.

Qualitative Controversy

- Some debate about extent to which DA-RT extends to qualitative work
- Growing consensus that applies in modified form
- Overview
  - Diversity of qualitative work requires some flexibility in DA-RT
  - DA-RT should be more “granular”, identifying specific sources of information
  - Scholars should provide “tracking appendix” (TRAX), linking individual sources of information to inferences or claims
- See: https://www.qualtd.net/

DA-RT Resources and Examples

- Resources
  - DA-RT: https://www.dartstatement.org/
- Most stringent examples are of journals that require:
  - Authors to provide all replication materials prior to publication
  - Replication materials generally must follow pre-established guidelines, e.g., codebook, data, computer code
  - Journal must be able to replicate and verify all core results, including graphs and tables
  - That is, readers of these journals have guarantee that all articles have been replicated at least once
- Examples from two top journals
  - American Journal of Political Science
    - Acceptance is contingent upon providing replication materials to Harvard Dataverse
    - Replications conducted by Odum Institute at UNC Chapel Hill and QDR at Syracuse University
    - https://dataverse.harvard.edu/dataverse/ajps
  - Political Analysis
    - Acceptance also requires authors to provide replication materials to Dataverse
    - Replications conducted by journal staff
    - https://dataverse.harvard.edu/dataverse/pan

Frequent Recommendations

From Bill Jacoby, ICPSR 2017:

- Always anticipate reproducibility and replication requirement
- Comprehensive documentation (e.g., lab notebook)
  - In software, do not use pull-down menus; never use any interactive features (i.e., never use any point-and-click features)
  - Write command files (.do files in Stata; .R scripts in R)
  - Comment files extensively; always err on side of over-documenting and over-commenting
  - Develop own set of rules for naming directories, files, and variables (having your own personal rules will make it easier for you to recall)
- **Use integrative tools**, i.e., tools that allow you to integrate writing and any tools you use for data production and analysis (e.g., MarkDoc in Stata, knitr/Sweave in R, plus LaTeX, markdown, HTML)

Best Practices - General

- Directories
  - Set a project directory
  - Within that directory, set a “working” folder
    - Keep all in-progress materials in this working folder
  - Create separate folders/sub-directories for original data, figures, and tables, and final products
- Files
  - Use short names with ISO date, e.g., “filename20170907”, not “filename - Sep 7 2017”
  - Always save data or command file that has been modified under new name
- Command files
  - Header (name, project, date, last update, seed for random variables, etc.)
  - Environment setup (machine type, location, software version, packages, etc.)
  - Set working directory and other file paths
  - Consider using multiple command files
    - At least keep data management and data analysis separate (e.g., one file for data preparation, one for data analysis)
  - Files should be robust, i.e., should generate same results every time, even when run by other people on different computers
  - Avoid ambiguous abbreviations (if abbreviate, comment on it)
  - Comment extensively (document your thought process by making your comments clear and thoughtful; date your comments!)
  - Check results or output at regular intervals
  - Integrate command files with writing (e.g., MarkDoc in Stata or knitr/Sweave in R, with LaTeX, markdown, or HTML)

Best Practices - Stata

See template provide (.do file)

- Integrative tool: MarkDoc
- Text processing tools: LaTeX, markdown, or HTML
- Command files
  - File path for figures directory may need to be close to root directory
  - Open and close logs
  - Call a .do file from within another do file:
    - “do” and “run”
  - Comment characters:
    - *, /* */, or ///
  - Advanced topics to avoid mistakes and increase efficiency during repeated tasks
    - Macros
      - Local and global
      - Consider tradeoffs
        - Local macro will not carry over from one code block to another
        - Yet, global macro may remain in memory if not cleared
    - Loops
    - .ado files

Best Practices – R and RStudio

See templates provide (.Rnw and .Rmd files)

- Integrative tools: Sweave (.Rnw file) or knitr (.Rmd file)
- Text processing tools: LaTeX, markdown, or HTML

*** Easiest to use within RStudio, which is already an integrated development environment (IDE) ***

- Command files
  - Calling R scripts from within another R script
    - source()
  - Comment character:
    - # in R scripts or Rstudio
    - ``` in Rstudio to comment out large blocks (careful with this as close to knitr format for code chunks)
  - Advanced topics to avoid mistakes and increase efficiency with repeated tasks
    - Loops
    - Define own functions
    - Generate own packages
  - Advanced integration of R and Stata
    - Call Stata from R using knitr package (easiest to implement in RStudio)

Sources/Resources

Video: Bill Jacoby and Arthur (Skip) Lupia at ICPSR 2017: https://www.youtube.com/watch?v=wleSZ8PnCD0

Book:

J. Scott Long, 2009, *The Workflow of Data Analysis Using Stata*

More info here: https://www.stata.com/bookstore/workflow-data-analysis-stata/

Slides: Scott Long's work flow slides to accompany 2009 book: https://www.ihrp.uic.edu/files/Workflow%20Slides%20JSLong%20110410.pdf

Shawna Smith's teaching materials: http://shawnasmith.net/icpsrcda/