Workflow of Data Analysis:
General Observations and Examples for Stata

Mark Fossett
Sociology, Texas A&M University
m-fossett@tamu.edu

Sociology Quantitative Methods Series, Summer 2021

Overview of the Presentation

The goal of this session is to provide a brief introduction to the topic of "Workflow of Data Analysis".

- **Note what is involved in best professional habits and practices of managing workflow of data analysis in sociological research.**
- **Comment on why these habits and practices are worth the trouble.**
- **Note some useful skills and tools**

Anticipating Conclusions

Having sound habits and practices for managing workflow of data analysis is an important attribute for professional researchers.

- **It pays dividends in saved time, effort, and emotional distress.**
- **It is a part of professional ethics – supporting the principle of replication and transparency.**
- **It promotes greater accuracy and trustworthiness of results and findings.**
- **It enhances one's professional reputation.**
- **It is a sign one takes the enterprise of research seriously.**
- **It can help protect against legal liability.**
- **It can help deal with mandatory requirements.**
- **It is a valuable job skill for research positions in both academic and applied settings.**

Therefore, it is desirable to adopt sound habits and practices as early as possible in one's professional career.

Assumptions

This presentation presumes familiarity with:

Basic research methods – both quantitative and qualitative – at the level of first graduate courses

A full introduction to review of workflow would require taking a course, and/or attending a series of extended workshops, and/or extended review of the emerging literature.

That is not the goal here.

The goals here are:

To raise awareness about the topic focusing primarily on quantitative methods.

To highlight some of the habits and practices that are easy to adopt for starters to provide a base to build on.

For a more in-depth review, see Long (2009) and/or consider the seminar in Urban Planning for spring 2022 (see example syllabus posted to the series website)

Workflow in Broadest View

In broadest view, workflow encompasses every aspect of the research and data analysis enterprise starting with study design and on to data collection, analysis, and reporting results.

Professional practices can assure

- **Accuracy, integrity, and robustness of results and findings**
- **Reproducibility**
- **Meeting highest ethical and scientific standards**

Professional practices can help one AVOID

- **"Black box", "trust me" (wink, wink) results and findings**
- **Irreproducible results**
- **Subpar professional reputation**
- **Legal liability**

Cautionary Tales from the Inter-Galactic Truth Squad

Sadly, replication has historically not been valued sufficiently in social science research

- **Replication studies are surprisingly (shockingly!) hard because information needed for replication often is not available**
- **Results of replication studies are hard to publish**
- **Replication studies are less "sexy" for research funding**

This is changing

- **Some funding agencies are now mandating that proposals must establish that research results will be replicable**
- **Some journals are now making submission of data and analysis programs prerequisites for article review and publication**

Attention to reproducibility and ACTUAL replication are more common in some disciplines

- **It fosters better and more cumulative science and also higher credibility and prestige**

Assume these Trends are Going to Continue

Social science research is assigned less credibility and prestige today than in the halcyon days of the 1960's.

- **Inattention to rigor in documentation and a culture of replication and transparency undermines credibility**
- **"Trust me, I'm a scientist" doesn't fly anymore.**

Greater access to data, survey tools, and easy-to-use analysis software has created a proliferation of "junk science" analysis

- **Junk science has all the superficial trappings of high-quality science. The difference can be hard to distinguish (especially for lay audiences)**

So, more than ever, trust and credibility rest on having high standards in professional ethics and research rigor

- **Attention to documentation and a culture of replication can help separate high-quality research from junk science and build trust and credibility for the discipline**

There is no good argument for neglecting these practices.

Have the "Right Stuff" and Do the Right Thing

Put on your lab coat. You're a scientist dagnabit!

Take your work seriously; act like it is important

- **We're not putting on a neighborhood play.**
- **Not every analysis leads to policy changes or paradigm shifts.**
- **But, if the work is worth doing, adopt high standards**

For motivation …

- **First, assume reviewers adopt a skeptical view toward your work until you convince them otherwise.**
- **Next, if your work challenges received wisdom, assume resistance beyond "standard" skepticism.**
  - **One's analysis must be squeaky clean to overcome gatekeepers & naysayers**
- **Additionally, work on funded contracts could be reviewed by a disinterested auditor with possible implications for payment disputes or even potential legal liability**

Some Selected Topics

The Basics

- **Clarity and detail of documentation of research**
- **Provenance of data**
  - **Could a disinterested party reproduce the analysis data?**
- **Provenance of results and findings**
  - **Could a disinterested party reproduce the findings?**
- **Backup, Archiving, and Sharing**

Some Details

- **Working in groups**
- **Confidentiality and ethics**

Clarity and Detail of Documentation

Research projects involve hundreds of tasks and decisions, some large and some small, in the journey from research idea to findings.

While it is impractical to imagine that every single step in a large-scale project is documented in full detail, ask the following basic question.

> **Could a disinterested party (with basic professional training) understand and, at least in principle, implement the research plan?**
>
> **If the answer is no, there is a basis for concern.**

Sources of Documentation

- **Proposals can document basic design: data, sample, methods**
- **Addendums can document changes in basic research design**
- **Project guides/notebooks, programs, working papers, and supplementary materials can provide additional documentation not found in articles and books**

Provenance of Original Source Data – I

Primary data should be documented with special rigor and detail.

- **Ask, "Would a disinterested reviewer accept your data claims?"**
- **In principle, primary data should be available for replication and review (possibly with restrictions to protect proprietary interest).**

Documentation and sources of secondary data.

- **Obtain data and detailed documentation from original source (e.g., NORC GSS, federal agency, etc.).**
- **Alternatively, obtain data from an established and highly respected data distributor (e.g., ICPSR, IPUMS, NHGIS).**
- **READ THE DOCUMENTATION & TECH REPORTS; you own it!**

AVOID using data sets from colleagues and informal repositories

- **Short-term convenience undermines provenance claims**
- **Despite good faith and best intentions, unknown (and unknowable) problems can creep in**

Provenance of Original Source Data – II

Maintain a detailed version history of PRIMARY data sets.

- **Archive initial raw data entry files**
- **Document data problems and data cleaning/editing operations in detailed project notebooks.**
  - **Where feasible, automate this work in programs.**
  - **Use comments extensively.**
- **Maintain archives of input files and output files leading to final "cleaned" version of the data set**
  - **original raw data file, leading to**
  - **intermediate cleaning files (multiple), leading to**
  - **final cleaned data set**

Prepare detailed documentation of the cleaned data set with description of the process of data preparation.

- **Ask, "Would a disinterested party believe the data claims?"**
- **Ask, "Is it adequate for replication and reproducibility by a disinterested party?"**

Provenance of Original Source Data – III

Document chain of possession and modification for SECONDARY data sets (website links, download dates, etc.)

Maintain archival copies of the original data

Perform a series of appropriate checks of secondary data

- **Use programs to reproduce sample counts, tabulations, and other information reported in codebooks and technical documentation**

Provenance of Analysis Data Sets – I

Use programs to produce analysis data sets. Program code should document

- **Case selection**
- **Creation of new variables**
- **Data recodes/ correction/adjustment**
- **Data cleaning should be implemented in code where possible**

AVOID

- **Manual editing and direct user modification**
- **Gaps in chain of possession**

Maintain archival copies of the original data and also of intermediate data sets leading up to the analysis data set

Separate Production Work and Exploration

Production work – Work producing stable products such as:

- **Products used by other programs (e.g., analysis data sets)**
- **New variables and measures**
- **Work documenting decisions and protocols**
- **Results used in papers and presentations**

Exploratory Analysis

- **Work of an ephemeral nature.**
- **Of course, this can be very useful. But, if it does lead to something useful, it will become production work**

Production work needs to meet high standards for rigor and documentation

One can be more "casual" and informal with exploration, BUT ONLY UNTIL the point that it has consequences for production work

Separate Personal and Professional Computing

Production work should be conducted in a dedicated computing environment separate from personal computing.

- **No games, music player, video player, or other non-essential software**
- **Minimum personal-use programs and materials**
- **Rationale: All personal-use programs pose unnecessary risks to your professional work.**

YES, THIS IS CONVENIENT

- **Sad, but true.**
- **But it takes only a moment's reflection to see the risks in mixing professional and personal computing**

Know the Relevant Features of Your Digital Data

Digital storage of information has many quirks hidden surprises

- **bytes, integers, long (examples of whole numbers)**
- **floats, doubles (examples of fractional/floating point numbers)**
- **characters, words, strings (examples of alphanumeric types)**
- **Whole numbers and alphanumeric types can be represented with exact accuracy in digital storage. Fractional numbers can only be approximated.**
- **The differences can matter, especially on equality tests.**

Know how missing data are handled and represented

Be aware display formats and internal representations of variables may not be the same.

Know Relevant Features of Your Digital Files

Digital data sets can have hidden surprises

Be aware of types of data files

- **Text files – simple and low tech, but reliable and robust**
  - **fixed field vs. free field (relevant for usage & file size)**
- **CSV files – a popular version of text-based data files**
- **System files – potentially superior and efficient, but with costs**
  - **May be hard to share across platforms and software**
  - **May have embedded variable labels and value labels with not code to document assignments**
  - **May vary in format across software versions**

"Hierarchical Files" vs. "Flat Files"

- **Hierarchical structure can be efficient**
- **Flat files are less efficient, but are easier to use**

Merges, Joins, & Concatenations– Oh My!

Merges (joins) are some of the most important, but also most treacherous data operations. When possible …

- **Merge on strings or high-precision integers**
- **Avoid merging on floating point numbers**

ALWAYS !!! Perform checks on merged data

- **Perform the checks with programs that generate log files to document the success of the merge**

Concatenation of data sets creates opportunities for problems

- **duplicating cases (inflating sample size)**
- **over-writing of cases (altering variable values)**

ALWAYS !!! Perform checks after concatenations

- **Perform the checks with programs that generate log files to document the success of the concatenation**

Case Selection

Case selection operations are common in data analysis

- Subsetting cases for analysis data sets
- Subsetting cases for application of recodes, etc.
- Subsetting cases for analysis procedure

Case selection involves logical tests

- Know your software functions and relevant digital issues
- When possible select on integers and strings
- AVOID equalities involving floating point variables
- Consider creating sample selection variables
  - Maintains code for complex case selection criteria in a single location in the program
  - Simplifies the terms needed for case section in analysis (avoids complex selection statements)

Variable Names and Value Labels

Variable names are often cryptic.

- **Avoid this where possible, but it is not always possible.**
- **Careful codebooks are usually necessary**
- **Thoughtful variable labels are an intermediate step**
- **Think carefully and try to develop conventions and protocols for naming variables to convey their content and purpose**

Take care with value labels.

- **Implement with program code that can be reviewed and corrected if necessary.**
- **Try to isolate the code in a routine that is called as needed (to avoid applying value labels across programs via cut-and-paste).**

Use Data Base Software

Database software and statistical software with data base features may have a steep learning curve. But the investment to learn the software is worth it.

- Strong "typing" of variables reduces errors and unexpected results
- Easier to automate tasks via programming code
- Greater robustness

"Death to Excel" – An example of useful but dangerous software

- Inconsistent numerical operations (e.g., display precision and calculation precision may not be as expected)
- Very little documentation of workflow
- No data typing
  - alphanumeric, floating point, and integer values can be intermingled and conflated
- Very limited options for automation
- GREAT for exploration; Problematic for production

Develop a Sequence of Limited Task Programs

Automate as many tasks as possible using programs.

AVOID OMINIBUS PROGRAMS.

- **Constant revising creates risks for accidents**

Instead, try to develop single-task programs.

- **Single-task programs do one or a few things.**
- **Once developed, they can remain stable**
- **When needed, they can be re-run and/or called from other programs**
- **Thus, editing and revisions can be limited in scope to just the programs needing attention**

In the ideal, strive for a situation where a "master program" would call other programs to reproduce the analysis from start to finish.

- **Individual routines can be run on a contingent (as needed basis)**

Organize Data and Products across Folders and Media

When possible, keep original data, analysis files, and research products, in a common environment

- **When it is necessary to spread materials across drives, computers, and environments, create a protocol and follow it.**

Use thoughtful folder structures to organize work by tasks, stage of analysis, versions, and other relevant concerns

Respect your folders!

- **Take the time needed to place materials in the correct locations so you can find them later**

Give thought to file sharing and backup needs.

- **For example, separate original data from analysis files for easier backup**

Don’t Be Penny-Wise and Pound-Foolish

- **Inconvenience in the short run will save time, energy, and drama in the long-run**

Backup Strategies

Backup is essential. It not *will* you need a backup. It is *when*.

Some Options

- **Main analysis computer**
- **External hard drives**
- **Cloud storage**
- **Secondary computers**

Adopt and follow protocols for frequency of backup

Adopt and follow protocols for version control and synchronizing

Use encryption and compression wisely

- **Sometimes these are necessary**
- **But they create a layer of complication and risk**

Be wise about person-time efficiency and computing efficiency (buying an external hard drive can be cheaper than reusing drives)

Create a Workflow Document for Special Tasks

When automation is not feasible, create workflow documents to review tasks, protocols, and important concerns

Use Program Comments Liberally (But Accurately)

Use comments extensively in programs.

It never hurts (unless it is incorrect or misleading, which of course it won't be).

Adopt the following attitude

- What seems obvious and unnecessary to comment today will be obscure and impenetrable tomorrow. So, comment well.
- What is obvious to you may not be obvious to a collaborator or to someone seeking to replicate your work. So, comment well.

Automate Data Management and Data Analysis

Nowadays most aspects of data management and analysis can be automated

AVOID doing any production work manually, or by "point-and-click" operations

- **This is inherently hard to reproduce**
- **It often does not leave a "paper trail"**
- **Provenance of results breaks down**

Use relative file and folder names so programs are not closely tied to specific computers, drives, etc.

Avoid Hand Calculations

Avoid hand calculations of any values that are going to be "production" values or used to create production values.

Learn to use scalars, macros, and similar tools to store and manipulate constant values used in any aspect of production work.

- **Greater accuracy**
- **Easier to replicate**
- **Reduces opportunities for human error**
- **Enhances provenance of results**

Automate Graphs

Even graphs and figures can be automated.

Use software that permit automation of graphs via a command language

- **See earlier comment on "Death to Excel"; it can be convenient, but it cannot be automated.**
- **Stata, R, etc. can automate graphs**
  - **The amount of detail and control over specifics is high**
  - **It is easy to modify "styles" and sizing to create consistent results**
- **Point and click operations can be used to generate command syntax and code for producing complicated graphs**
- **Once created, graph code can be recycled**
  - **At this point, graphing becomes easier, more efficient, and more consistent, and more robust**

Automate Table Production

Table production is tedious, time-consuming, and prone to error.

It is increasingly easy to automate production of even complex tables

Use software that permits preparation of tables via a command language

A Few Relevant Stata Procedures

- **The "table" command (especially as enhanced in v17).**
- **The "estimates , table" command**
- **The "estimates , stat" command**
- **Third party routines such as the "tabout"**

Increasingly, procedures can produce Word documents and PDFs with near-publication quality features

- **This is amazingly efficient for working on papers and reports where results must be updated**

Working with Confidential Data

Data confidentiality concerns pose problems for replication

- **Plain and simple, this is a major problem for science**
- **There is no way to put lipstick on this pig. It may be unavoidable for ethical and other reasons. But, it nevertheless places limits on best scientific practice.**

Research Data Centers Practices

- **Research Data Centers are developing protocols to permit replication work using confidential data**
- **This shows it is possible to protect confidentiality while still meeting the needs and norms of scientific research**

Working in Teams

Collaboration multiplies problems of coordination and provenance of data sets and results

Develop protocols for the division of labor for production and sharing of work

- **See Long (2009) for extend discussion**

AVOID the WILD WEST

- **Absence of structure, a guiding vision, and protocols is very likely to lead to inefficiencies and problems, possibly fatal to the project.**
- **Backing up and "untangling" work that has not been guided by protocols can be an enormous time suck**

Conclusions – Summing Up

Having sound habits and practices for managing workflow of data analysis is an important attribute for professional researchers.

- **It pays dividends in saved time, effort, and emotional distress.**
- **It is a part of professional ethics – supporting the principle of replication and transparency.**
- **It supports greater accuracy and trustworthiness of results and findings.**
- **It protects one's professional reputation.**
- **It signals one takes the enterprise of research seriously.**
- **It can protect against legal liability.**
- **It may be mandatory!**
- **It is a valuable job skill for research positions in both academic and applied settings.**

It is desirable to adopt sound habits and practices as early as possible in one's professional career.

Parting Advice

Adopt professional habits of workflow of data management and analysis

Reasons

- **Makes sense if you are ethical and rigorous**
- **Makes sense if you want a good reputation**
- **Its easier than you may think**
- **Makes sense if you are lazy!**
  - **In the end it saves time and effort.**
- **Its fun! (okay, that's not really true)**

End of Slides

Thank you for your attention.