Documentation Tips

Based on *The Workflow of Data Analysis Using Stata* by Scott Long

Long's law of documentation: *It is always faster to document it today than tomorrow.*

It is important to make the tacit knowledge you accumulate as you are working on a project explicit for yourself later on and for others in the future.

When to document?

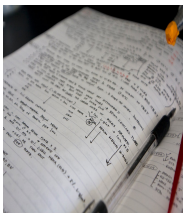
- Incorporate documentation into the procedures for each step of collection and analysis, rather than waiting until the end of the project.
- Schedule time for it each week, like a meeting on your calendar. Or use the last five minutes of each work period to document what you have done during that time.

What to document?

- Use the "hit-by-a-bus" test: If you were hit by a bus, would a colleague be able to pick up your work and continue it?
- Examples of what to document:
    - Data sources: Where did the data come from? How was it collected?
    - Data decisions: How were variables created and cases selected? What coding or scaling decisions were made and why? Also document decisions you made NOT to do something.
    - Statistical analysis: What steps were taken in what order? If an approach was explored but not taken in the final analysis, document this as well.
    - Software: Document the specific version and any packages. Things can change from version to version even in the same statistical package.
    - Storage: Where are the data stored and where are they backed up?
    - Ideas and plans: Write down ideas for future analyses, and tasks that need to be completed. Ideas that seem obvious now may not be obvious later.

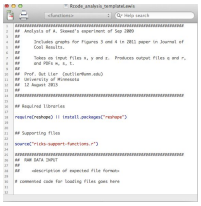
How to document?

The research log



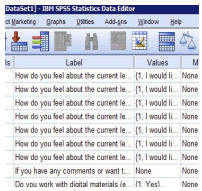
- Like a lab notebook in the sciences: Documents ideas, plans, the work that was done, who did it, and what the outcomes were.
- Also should indicate what other documentation is available and where it is located.
- Using a research log can also help you keep on track by specifying goals and plans ahead of time. Cuts the amount of time it takes to "catch up" each time you come back to the project.

Syntax files



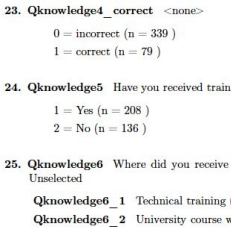
- Use a syntax file to record analyses:
  - In SPSS hit “paste” instead of “ok”
  - In R: Save your written scripts. Consider using markdown or sweave/knitR to keep output and scripts together.
- Make comments in your syntax:
  - In SPSS: use an asterisk (*); close with a period (.)
  - In R: use pound sign/hashtag (#)

Dataset documentation



- if you work with a lot of datasets, it is good to have a registry of datasets that will help you find the particular data you need.
- Add metadata to your dataset
  - Use short, descriptive variable names
  - Use Labels and Notes (stata), Variable and Value Labels (SPSS)

Codebooks



23. Qknowledge4_correct <none>

0 = incorrect (n = 339)
1 = correct (n = 79)

24. Qknowledge5 Have you received train

1 = Yes (n = 208)
2 = No (n = 136)

25. Qknowledge6 Where did you receive
Unselected

Qknowledge6_1 Technical training
Qknowledge6_2 University course w

- Quantitative Codebooks should include:
  - Variable names, values, labels/question text
  - How the variable was created (recoded, combined, etc)
  - Descriptive statistics such as Ns, percentages or means
  - If and how missing information is coded
- Qualitative Codebooks should include
  - Description of code
  - Guidelines for when to use the code and when to not use the code
  - examples of both included/excluded content