Data Management Tips

Advice from an Old Analyst Scold

Jason T. Newsom

Department of Psychology
Portland State University

Data Management

- With “data management” I am generally referring to data preparation after a study has been conducted and the data have been entered.

- Anything in from the data entry stage to the analysis stage

Data Collection

- Make your future life easier by very carefully, even obsessively, planning how the data are recorded
- My favorite ***bad*** example

Next I have some questions about your health. Would you say your health is excellent, very good, good, fair, or poor?

---

1. Excellent
2. Very good
3. Good
4. Fair
5. Poor

Data Collection

- Think ahead to analyses
  - Example of sufficiently continuous variables

Data Entry and Labeling

- Double entry and/or careful checking on accuracy of input
- Spend time brainstorming possible things that can go wrong and think ahead to analyses and data uses.
- Codebooks valuable but good data labeling really necessary—variable labels and value labels
- Larger the data set and the more people using it, the more formal and elaborate these steps need to be

Data Entry and Labeling

- Variable names
  - Short
  - Meaningful
    - e.g., `anxious`, `restless` is better than `q1`, `q2`
  - Longitudinal – consistency across waves, making it easier for quick coding and global changes
    - e.g., `anxt1`, `anxt2` or `w1anx`, `w2anx`

Data Preparation/Management

- Back up data
- Syntax, syntax, syntax (or in R, Script, script, script)
  - Record keeping of transformations critical
  - Notes and documentation for others who use the data and for yourself
  - You just can’t keep track of these things with menus in SPSS or using the console command line piecemeal in R

Data Preparation/Management

- My process is to create one large syntax file for each paper
- Run this file every analysis, leaving the data file unchanged (don't save when exiting)
- Saving over the data file runs the risk of rerecoding or recomputing if you use the same variable name
- Use file location at the top (in SPSS using `get file="c:\path location…"`, so you know the name of the data file and where it was stored.
- Code can be reused for other papers (e.g., BMI computation, alcohol drinks per week)

Data Preparation/Management

- Clear output file after every run so you can find the output you need more quickly
- In SPSS add `output close *.` at the top of the syntax file.
- In R, add `cat("\014")` at the top of the script file.

Data Preparation/Management

- For some larger studies, we have devoted an extensive syntax file just to recoding and computing scales, and then saved a prepped version of the data file
  - Drawback is that if you decide you need to change or modify something (e.g., dropping an item), you need to update or go back to that original data file
  - Creating multiple data files can be confusing about contents, locations etc.

Data Preparation/Management

A few more specific remarks

- Missing values need to be identified

```
missing values w1emo1 w1emo2 w1emo3 (-99).
```

- Carefully consider how missing data are treated when computing scales
  - See more detail at http://web.pdx.edu/~newsomj/uvclass/ho_missing.pdf
  - `Sum` function treats missing items for a case as a 0
  - `Mean` function treats missing items for a case as the average value
- Can require minimal number of items present, e.g.,

```
COMPUTE w1neg=MEAN.10(w1unw1,w1unw2,w1unw3,w1dwn1,w1dwn2,w1dwn3,
  w1out1,w1out2,w1out3,w1fai1,w1fai2,w1fai3).
```

Data Preparation/Management

A few more specific remarks

- Temporary command in SPSS is handy instead of creating multiple filter variables

```
temporary.
select if w1neg lt .42.
desc vars = w1health w2health w3health w4health w5health w1neg.
```

Thank you!

Please contact Jason Newsom, newsomj@pdx.edu, with comments or questions.