Data Management Workshop

Hsueh-Sheng Wu

CFDR Workshop Series

September 12, 2016

BGSU

Center for Family and Demographic Research

Outline

- What are data management problems?
- How do data management problems occur?
- Organize your files
- Data construction and analysis
- Sample commands for creating log and data files
- Create a folder system
- Conclusions

What Are Data Management Problems?

There are many types of data management problems. You may be experiencing these problems if you:

- work on a very long command file
- always start your analysis with the original data file
- have many data files in your folder
- have many similar command files in your folder
- cannot quickly locate variables or files that you need
- cannot remember whether and why you create, recode, modify certain variables

What Are Data Management Problems? (Cont.)

Consequences of data management problems?

- They impede your research progress
- They impede other people’s research because when the CFDR network drive does not have enough storage space, other users cannot run analyses and open or save files on the server.

Why Do Data Management Problems Occur?

The completion of a research project requires you finish many tasks

- Planning ahead
  - What is your research question?
  - What are research hypotheses?
  - What variables will be needed?
  - What is the target population?
  - What data set will be used?
  - What analyses will be conducted ?
  - What software will be used?

Why Do Data Management Problems Occur? (Cont.)

- Data construction
  - Obtain the data files
  - If necessary, link different data files
  - Select the variables for the project
  - Correct errors and inconsistencies in the variables
  - Construct the independent, dependent, and control variables
  - Select the target population
  - Save a copy of the data (Do not overwrite the original data!)

Why Do Data Management Problems Occur? (Cont.)

- Data analysis
  - Conduct the analysis

- Reporting the results
  - Create tables and graphs
  - Write up the paper

BGSU

Center for Family and Demographic Research

Why Do Data Management Problems Occur? (Cont.)

Files accumulated at each stage

Planning Stage:

- Journal articles
- Book chapters
- Web documents on data and methods
- Drafts of the research proposal

Variable construction stage

- Stata or SAS command files
- Log files for Stat users and log and outcome files for SAS users

Data analysis stage

- Stata or SAS command files
- Log files for Stat users and log and outcome files for SAS users

Reporting results stage

- Words files
- Excel files
- PowerPoint files

Organize Your Files

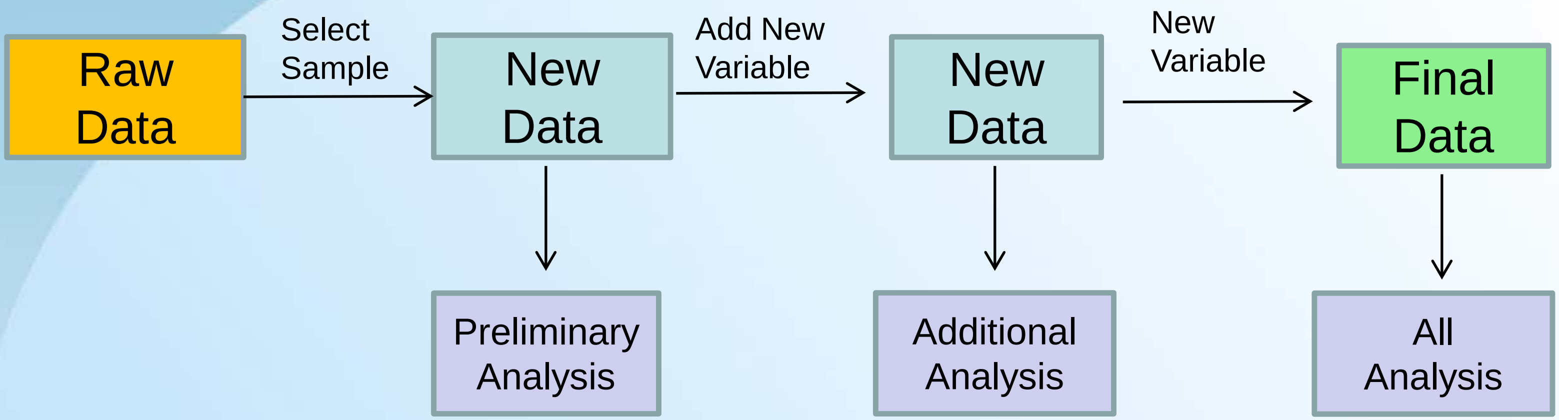
- Create a folder system to help you organize files for the same project

- Create a spreadsheet to keep track of your files
  - Name the command, log, result, and data files wisely, so you can easily find the correct files
  - Remember to include the version number or the date in the file name
  - Remember to merge and purge command files for data construction

Organize Your Files (Cont.)

- Always write separate command files for constructing variables and for data analysis
- Command files should have their corresponding log/result files
- Because data construction has its continuity, you usually just need one command file linking the original data to data file that you are currently working. Don't generate a new data file each time you modify a command file.
- Because results of data analysis are determined by the data used, you do not need to merge the command files for data analysis if they use different data sets.
- Keep only the variables and observations that you need in your data files
- Document the purpose, data file, log file, result file, problems, decisions, thoughts, doubts in the command files
- Document how new variables are created from the original variables
- When constructing variables, you should not overwrite the original variables
- You should have "permanent file" and "working file" when working on your project

Data Construction and Analysis



- You should have separate command files for constructing data and analyzing data. Don't mix these files.

- This diagram will have three data construction steps and three data analyses

- How many files do you expect to have in the end?
  - An excel file documents the purposes and contents of all command, log, result, and data files
  - A command file for data construction
  - A log file for data construction
  - A final data file
  - Three command files for data analysis
  - Three log/result files for data analysis

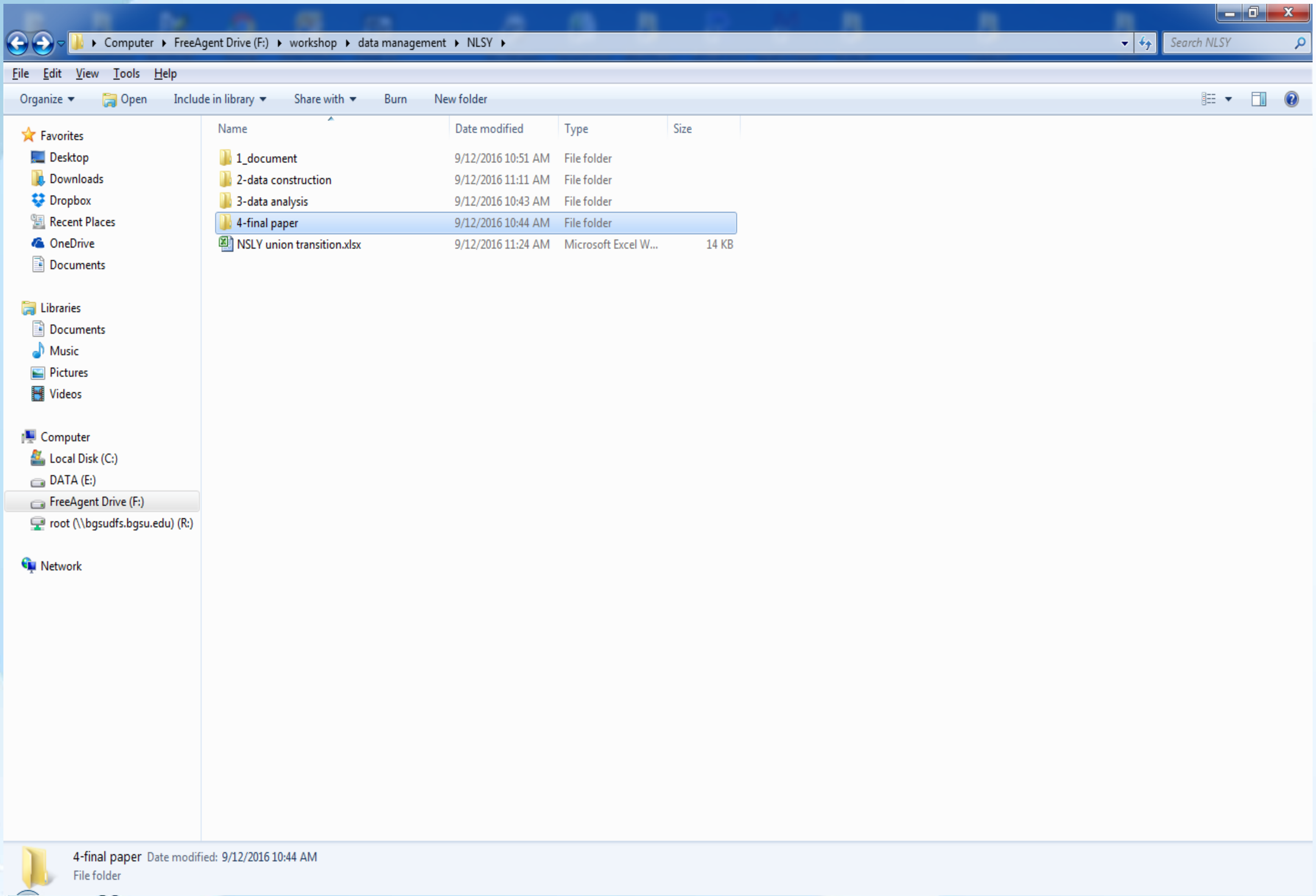
Stata for Creating Log and Data Files

```
***********************************************************************************************************
* Stata command file: auto_stata_9_12_2016
* Log file: auto_stata_9_12_2016.log
* This command file lists the frequencies of two variables: make and price
* The original data, auto.dta, was used for the analysis
* No new data file was created
************************************************************************************************************

*create the log file
log using “e:\temp\ auto_stata_9_12_2016.log”, replace

*open the original data
use “e:\temp\auto.dta”, clear

* create the frequencies of make and price
list make price

* Save the log files
log close
```

SAS for Creating Log and Result Files

```
* SAS command file: auto_sas_9_12_2016;
* Log file: auto_sas_9_12_2016.log;
* Result file: auto_sas_9_12_2016.lst;
* This command file lists the frequencies of two variables: make and price;
* The original data,. auto.sas7bdat, was used for the analysis;
* No new data file was created;

PROC PRINTTO LOG=‘e:\temp\auto_sas_9_12_2016.log’ NEW; /* create the log file*/
RUN;

PROC PRINTTO PRINT='e:\temp\auto_sas_09_12_2016.lst' NEW; /* create the result file*/
RUN;

PROC PRINT DATA='e:\temp\auto'; /* create the frequencies of make and price*/
VAR make price ;
RUN;

* Save the result and log file;
PROC PRINTTO PRINT=PRINT;
RUN;

PROC PRINTTO LOG=LOG;
RUN;
```

Create A Folder System



Name	Date modified	Type	Size
1_document	9/12/2016 10:51 AM	File folder	
2-data construction	9/12/2016 11:11 AM	File folder	
3-data analysis	9/12/2016 10:43 AM	File folder	
4-final paper	9/12/2016 10:44 AM	File folder	
NSLY union transition.xlsx	9/12/2016 11:24 AM	Microsoft Excel W...	14 KB

4-final paper Date modified: 9/12/2016 10:44 AM
File folder

Conclusions

- Data management problems influence not only your research but also other people’s.
- The completion of a research project involves the creation of many files. Therefore, you need to consider how to organize them to help your research. Create a folder system is very important.
- Create an Excel file to keep track of all the files
- Create separate command files for data construction and data analyses. Document everything in your command files
- Each command file should have its corresponding log file.
- Use consistent names for command, log, result, and data files.
- After you complete a task, merge the command file with the previous ones.
- If you have any questions about data management problems, feel free to contact Hsueh-Sheng Wu @372-3119 or wuh@bgsu.edu