Data Management with Stata

Hsueh-Sheng Wu

CFDR Workshop Series

May 22, 2023

BGSU

Center for Family and Demographic Research

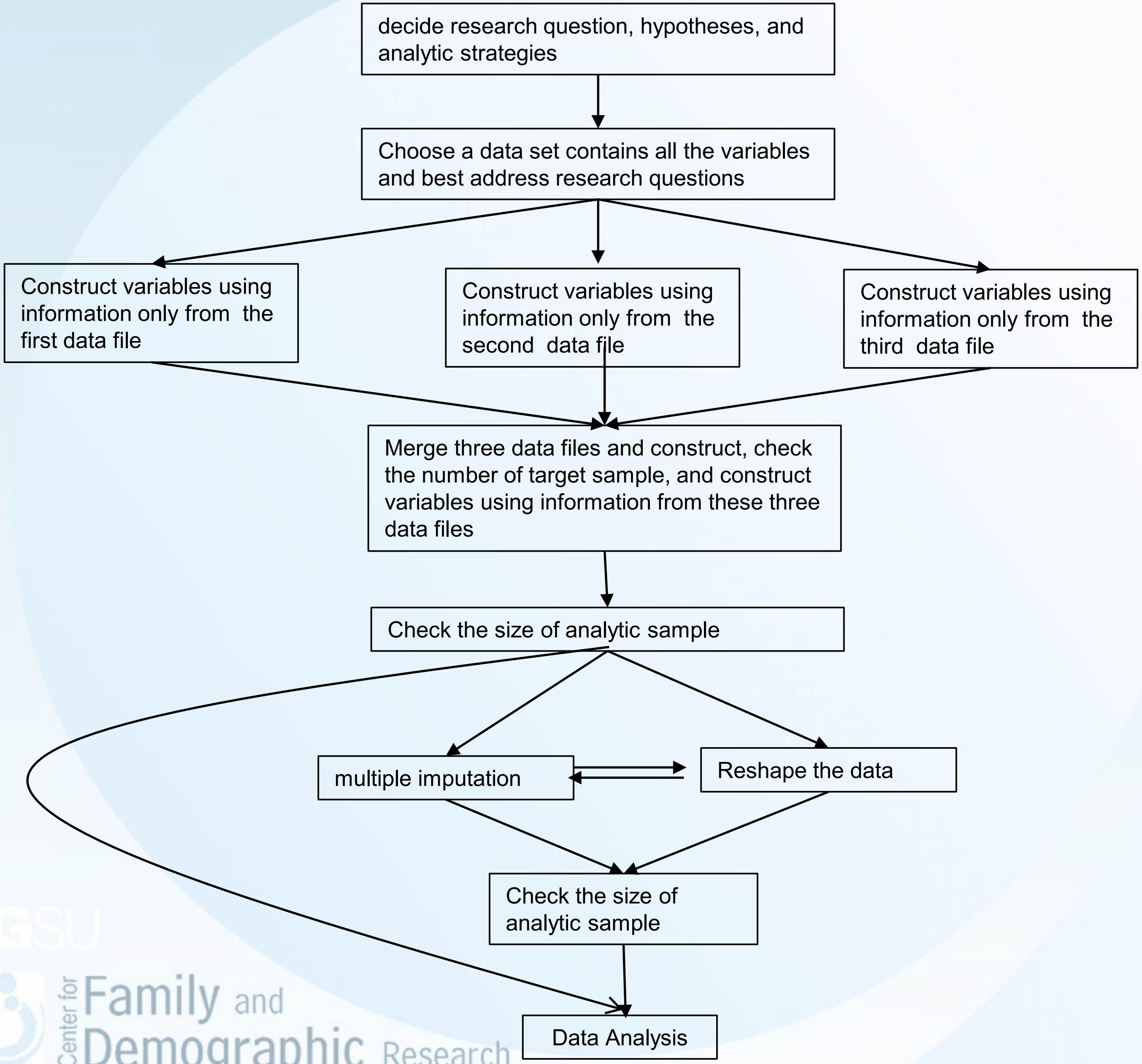
Outline

- Importance of data management in research
- A flowchart of data management
- Check list at each step of data management
  - Preparation works before data management
  - Decide on the data file
  - Check the accuracy of each data file
  - Construct variables
  - Merge all the data file
- An example of data management
- Conclusion

Importance of Data Management

- Data Management plays a critical role in social science research, but received little attention
- The major tasks of Data Managements are
  - Data preparation
  - Data Cleaning
  - Variable Transformation
  - Subset Creation
  - Data Documentation and reproduction
- Social Science research often involves lots of decisions on how to analyze imperfect data. Thus, good data management skills help social sciences researchers keep tract of why and how they code and construct variables and make it easy to modify if necessary, variables.
- Reviewers of manuscripts often raises many methodological questions. Good data management skills help researchers justify their decisions on these methodological questions.

A Flowchart of Data Management



decide research question, hypotheses, and analytic strategies

Choose a data set contains all the variables and best address research questions

Construct variables using information only from the first data file

Construct variables using information only from the second data file

Construct variables using information only from the third data file

Merge three data files and construct, check the number of target sample, and construct variables using information from these three data files

Check the size of analytic sample

multiple imputation

Reshape the data

Check the size of analytic sample

Data Analysis

Check list at each step of data management

Preparation works before data management

- Decide research questions
  - 1. What are research questions?
  - 2. What are research hypotheses for each research question?
  - 3. What variables are needed
  - 4. How these variables have been conceptualize and operationalize these variables?
  - 5. What is the population of research questions?

- Analytic Strategies
  - 3. what analyses are need for respective hypotheses?
  - 4. What is the data format for testing each hypothesis?

Check list at each step of data management (continued)

Decide on the data file

- Locate data files contain variables you need. For constructing complex variables, you need to make sure you have all the variables needed.
- Check out wordings and response categories of questions in codebooks to determine if these questions adequately measure each variable of interest.
- Examine the sampling frame and decide if the data file covers the target population
- Look at the sampling methods and decide if the analysis result should be weighted.
- Depending on research questions, other data attributes like data collection years, the number of waves may also need to be considered.
- Decide the data file or files to use
- Document what file each variable is from and identify possible inconsistencies across variables

Check list at each step of data management (continued)

Check the accuracy of each data file

- Read in data files without any error message
- if the number of observations in the data file match that of the codebook
- locate the ID variable or variables and check how many observations are associated with each ID value

Check list at each step of data management (continued)

Construct variables

- Check the attributes and/or frequencies of the ID variable, independent, dependent, control, and weight variables. (i.e., string or numerical; value labels, variable labels
- Generate new variables from these variables
- Add value and variable labels to generated variables
- Recode or replace the values of new variables if need
- Check if new variables are created correctly in each data file
- Construct variables and note the number of valid observation in the data file
- Save variables that you will use In a new data file

Check list at each step of data management (continued)

Merge all the data files into a new data file

- Merge data files together
- Identify the sample of target population
- Check if variables that should not vary across data files do not vary across data files.
- Construct variables using different sources of data
- Identify the size of analytic sample
- If necessary, use multiple imputation to reduce the problem of missing data
- If necessary, reshape the data for analysis using data in long format
- If multiple imputation or reshape command are used, recheck the size of analytic sample.

An example of data management

- Research questions:

  Whether race and gender have different impact on the trajectory of depressive mood between ages 18 and 45.
- Data: Add Health public data at Waves 3,4, and
- Variables needed: age, gender, race, and depressed mood
- Statistical technique: hierarchical linear model
- Final data format: the long format.

An example of data management (continued)

Table 1. The variables and Data files to be used

		W3		W4		W5	
		original data	new data	original data	new data	original data	new data
	data files	21600-0008-Data.dta	w3.dta	21600-0022-Data	w4.dta	21600-0032-Data	w5.dta
		original variable	new variable	original variable	new variable	original variable	new variable
	ID varialbe	aid		aid		aid	
variables to be constructed	available variables						
Age	Inteview Year	iyear3	Interview_y3	iyear4	Interview_y4	iyear5	Interview_y5
	Interview Month	imonth3	Interview_m3	imonth4	Interview_m4	imonth5	Interview_m5
	Birth Year	h3od1y	birth_y3	h4od1y	birth_y4	h5od1y	birth_y5
	Birth Month	h3od1m	birth_m3	h4od1m	birth_m4	h5od1m	birth_m5
Gender	Gender	bio_sex3	female_w3	bio_sex4	female_w4	h5od2a	female_w5
Race	White	h3od4a	white_w3			h5od4a	white_w5
	Black	h3od4b	black_w3			h5od4b	black_w5
	Indians	h3od4c	indian_w3			h5od4f	indian_w5
	Asians/islander	h3od4d	asian_islander_w3				
	asian only					h5od4d	asian_w5
	islander					h5od4e	islander_w5
	Best Category for multiple racial background	h 3od 6	best_category_w3			h5od8	best_category_w5
	Hispanic	h3od2	hispanic_w3			h5od4c	hispanic_w5
	Others					h5od4g	other_w5
Depressed Mood	Depressed	h3sp9	depressed_w3	h4mh22	depressed_w4	h5ss0b	depressed_w5
	Sad	h3sp12	sad_w3	h4mh26	sad_w4	h5ss0d	sad_w5
	Cannot shake of the blues	h3sp6	blues_w3	h4mh19	blues_w4	h5ss0a	blues_w5

An example of data management (continued)

- See the accompanying command and log files

Conclusions

- Data management skills are extremely important to social sciences researchers
- Research questions and hypotheses provides the roadmap for how variables and data sets should be constructed. Thus, it is critical to think them through before working on data.
- Data Management skills focus on Data preparation, Data Cleaning., Variable Transformation, Subset Creation, and Data Documentation and reproduction. Each of these skills requires different but simple Stata commands
- The use of command and log files is very important in achieving these data management tasks
- When working on data, always keep tract of which respondents are in the sample and which respondents have missing values on what variables. Such information identifies the size of final analytic sample.
- If you have any questions about data management, please send an email to me (wuh@bgsu.edu) or drop by my office. I am glad to help.