Constructing Variables with Stata

Hsueh-Sheng Wu

CFDR Workshop Series

January 31, 2022

BGSU
Center for Family and Demographic Research

Outline

- Importance of learning how to construct variables
- Steps of data analysis
- Key information about data
- Key information about cleaning data and constructing variables
- A demonstration of how to clean and construct marriage history data
- Conclusions

Importance of Learning How to Construct Variables

- Many theoretical constructs used in social sciences are measured either by combining several indicators measured at the same time (e.g., depressive symptoms) or by combining measures across different time points (e.g,. martial history).
- Many students mistakenly think that to construct variables, they just need to aggregate the information of different indicators into a theoretical construct. Subsequenlty, they underestimate the time and effort they need to spend in understanding and cleaning up the data before they can start constructing variables.
- This workshop will demonstrate how to use Stata to look at the data, clean up errors in the data, and then construct marriage history for individual respondents.

Steps of Data Analysis

- Steps
  - ❖ Step 0: Decide research questions and hypotheses
  - ❖ Step 1: Obtain original data
  - ❖ Step 2: Understand data by reading questionnaires, codebooks, and user's guides and browse through data files
  - ❖ Step 3: Clean data and construct variables
  - ❖ Step 4: Run statistical analyses
  - ❖ Step 5: Obtain final results
  - ❖ Step 6: Complete the project
- Timeline

- Cleaning data and constructing variables are usually very time consuming,and thus analysts should plan ahead and allow themselves to have sufficient time to complete these two tasks.

Key Information about Data

Before start using the data, users should know the following information about the data:

- What is the number of observations in the data?
- What is the total number of variables?
- Are variable and value labels associated with the variables?
- Are the variables string or numerical?
- Which one(s) are the ID variable(s)?
- Are the data in long or wide format?

Key Information about Cleaning data and Constructing Variables

- Social data may contain errors because of entry errors, design effects, or respondents' erroneous reports. Thus, users need to clean the data before constructing variables.
- When cleaning data, users usually look at two issues:
  - ❖ Coding errors are impossible values of variables, which is very easy to identify.
  - ❖ Inconsistences are conflicting information across indicators, which is usually difficult to identify.
- Cleaning data is time-consuming because users need to replace the values of the variables for individual respondents.
- To clean data and construct variables, users need to know what variables will be used to construct new variables, what errors may exist for these variables, and what other variables to be used to identify possible problems in these variables.
- When cleaning data, users may need to make some judgement calls.

A Demonstration of How to Clean Marriage History Data

Sample data used in this workshop:

- The information about marriage history was collected from 14 respondents between December 2020 and January 2021.
- The respondents reported a maximum of three prior marriages.
- The goal is to construct variables to describe the stability of marital status between the start of first marriage and the interview date.

A Demonstration of How to Clean Marriage History Data (Cont.)

- Variables in the Sample Data

Table 1: Description of Variables in the Sample Data

variable	Variable Label	Range of Valid Values
id	Id	1-14
inter_y	Interview Year	2020-2021
inter_m	Interview Month	1-12
mar_sy1	Start of First Marriage, Year	1970-2017
mar_sm1	Start of First Marriage, Month	1-12
mar_ey1	End of First Marriage, Year	1969-2021
mar_em1	End of First Marriage, Month	2-11
mar_sy2	Start of Second Marriage, Year	1990-2018
mar_sm2	Start of Second Marriage, Month	1-11
mar_ey2	End of Second Marriage, Year	2001-2018
mar_em2	End of Second Marriage, Month	3-12
mar_sy3	Start of Third Marriage, Year	2006-2019
mar_sm3	Start of Third Marriage, Month	1-10
mar_ey3	End of Third Marriage, Year	2016-2019
mar_em3	End of Third Marriage, Month	5-11
marital	Current Marital Status	0-1
mar_num	Total Number of Marriages	1-3

A Demonstration of How to Clean Marriage History Data (Cont.)

Table 2. Sample Data

ID	Interview Year and Month		First Marriage				Second Marriage				Third Marriage				Currently Married	Total Marriage
	Year	Month	Start Year	Start Month	End Year	End Month	Start Year	Start Month	End Year	End Month	Start Year	Start Month	End Year	End Month		
id	inter_y	inter_m	mar_sy1	mar_sm1	mar_ey1	mar_em1	mar_sy2	mar_sm2	mar_ey2	mar_em2	mar_sy3	mar_sm3	mar_ey3	mar_em3	marital	mar_num
1	2020	12	1974	3	1980	7									0	1
2	2020	12	1976	4											1	1
3	2020	12	1984	8	2000	3	2004	4	2008	7					0	2
4	2020	12	1986	9	2004	3	2008	1							1	2
5	2021	1	2002	5	2005	3	2006	10	2016	11	2017	4	2019	5	0	3
6	2021	1	2004	6	2006	2	2007	11	2018	12	2019	1			1	3
7	2012	1	2000	4	2001	4	2005	9	2014	8	2016	3	2018	6	0	3
8	2021	14	1994	1	1996	5	2000	6							1	2
9	2020	12	1970	1	1969	11									0	1
10	2020	12	1972	2	2021	9									0	1
11	2021	1	1992	12	2016	2	2018	2							1	2
12	2021	1	2016	3	2019	3	1990	11							0	2
13	2021	1	1986	9	2004	3	2000	5	2001	5					0	2
14	2021	1	2017	4	2019	5	2002	5	2005	3	2006	10	2016	11	0	3

BGSU Center for Family and Demographic Research

A Demonstration of How to Clean Marriage History Data (Cont.)

Table 3. Final Data after errors have been corrected

ID	Interview Year and Month		First Marriage				Second Marriage				Third Marriage				Currently Married	Total Marriage
	Year	Month	Start Year	tart Month	End Year	End Month	Start Year	tart Month	End Year	End Month	Start Year	tart Month	End Year	End Month		
id	inter_y	inter_m	mar_sy1	mar_sm1	mar_ey1	mar_em1	mar_sy2	mar_sm2	mar_ey2	mar_em2	mar_sy3	mar_sm3	mar_ey3	mar_em3	marital	mar_num
1	2020	12	1974	3	1980	7	.	.	.	.	.	.	.	.	0	1
2	2020	12	1976	4	2020	12	.	.	.	.	.	.	.	.	1	1
3	2020	12	1984	8	2000	3	2004	4	2008	7	.	.	.	.	0	2
4	2020	12	1986	9	2004	3	2008	1	2020	12	.	.	.	.	1	2
5	2021	1	2002	5	2005	3	2006	10	2016	11	2017	4	2019	5	0	3
6	2021	1	2004	6	2006	2	2007	11	2018	12	2019	1	2021	1	1	3
7	2021	1	2000	4	2001	4	2005	9	2014	8	2016	3	2018	6	0	3
8	2021	1	1994	1	1996	5	2000	6	2021	1	.	.	.	.	1	2
9	2020	12	1970	1	.	.	.	.	.	.	.	.	.	.	0	1
10	2020	12	1972	2	.	.	.	.	.	.	.	.	.	.	0	1
11	2021	1	1992	12	2016	2	2018	2	2021	1	.	.	.	.	1	2
12	2021	1	1990	11	.	.	2016	3	2019	3	.	.	.	.	0	2
13	2021	1	1986	9	2004	3	.	.	.	.	.	.	.	.	0	2
14	2021	1	2002	5	2005	3	2006	10	2016	11	2017	4	2019	5	0	3

BGSU Center for Family and Demographic Research

A Demonstration of How to Clean Marriage History Data (Cont.)

Table 2. Stata commands used

Stata Command	Purpose
destring [varlist] , generate(newvarlist)	Convert string variables to numeric variables
generate newvar = oldvariable	Create new variable
label variable varname ["label"]	Label variable
count	Count observations
des	Describe data in memory
duplicates report [varlist]	Report duplicate observations
keep if exp	Keep observations that satisfy specified condition
label values varlist [lblname|.]	Assign value label to variables
log close	Close a log file
log using filename	Open a log file
replace oldvar =exp [if]	Change contents of variable
reshape	Convert data from wide to long form and vice versa
sort	Sort data
sum	Provide Summary statistics
tab1 varlist	Create One-way table for each variable
tab2 varlist	Create Two-way table of frequencies

Conclusions

- Constructing variables is an important step of data analysis. When there are errors in variables, it is impossible to accurately constuct new variables. Thus, it is important to consider cleaning data and constucting variables together.
- To clean data and construct variables, users need to know what variables will be used to construct new variables, what errors may exist for these variables, and what other variables are used to identify possible problems in these variables.
- Codebooks and user's guides are very useful for identifying coding errors in variables, but not so much for inconsistencies between variables. Thus, when correcting inconsistencies between variables, researchers may need to make some judgement calls.
- Only simple Stata codes are needed for cleaning data and constructing variables. However, users may need to switch between the wide format and the long format to quickly clean the data and construct variables.
- Correcting the errors of variables usually needs to be done for individual records and will be time-consuming. Therefore, researchers should allow themselves sufficient time to do it.