Key Stata Commands for Constructing Variables

Hsueh-Sheng Wu

CFDR Workshop Series

July 11, 2022

BGSU

Center for Family and Demographic Research

Outline

- The importance of constructing variables
- Considerations on constructing variables
- Steps of constructing variables
- Key Stata commands for constructing variables
- Two Stata examples
  - Constructing variables using wide-format data
  - Constructing variables using long-format data
- Conclusions

The Importance of Constructing Variables

- The purpose of constructing variables is for researchers to eventually test theoretical relations of constructs among sampled respondents and generalize the findings to the target population.

- When submitting a manuscript, researchers need to state which variables were used, how variables were measured, what information these variables provided, how such information was aggregated to generate theoretical constructs, and how many respondents were in the final analytic sample.

- Researchers should be able to provide justifications on the way they chose to generate theoretical constructs in their research.

Considerations on Constructing Variables

- Are the data single-level or multilevel? The answer to this questeion will determine how to calcuate the number of valid respondents in the data.
- What item(s) can be used to construct variables? Researchers usually rely on theories and/or emprical studies to choose items for constructing variables.
- Are these items string or numerical variables? Different approaches are used to construct variables, depending on whether the items are string or numerical variables.
- Have the missing values of a variable been coded? The missing values of variables need to be correctly coded before variables can be used to construct other variables.
- How should the information of the items be aggregated? Different methods in aggregating the items have impacts on the number of respondents with valid values and the size of final analytic sample.
- Is there a way that can more accurately aggregate variables? It is difficult to consider the combination of three or more variables at the same time. Thus, it important to know how to aggregate several variables accurately.

Steps of Constructing Variables

- Check if the data are in wide format or long format
- Select the variables needed
- Check the attributes of these variables
- Check and code the missing values of these variables
- Decide what constructs to be created
- Decide what variables are to be used to construct the variables
- Decide how the information will be aggregated
- Check the number of respondents with valid values on the constructed variables
- Check the number of respondents without missing values on any variable used for the analysis
- Conduct the analysis and check the number of observations of the analysis

Key Stata Commands for Constructing Variables

Table 1. Stata commands

Stata Command		**Purpose**
clonevar		Clone existing variable
clonevar newvar = varname		generates newvar as an exact copy of an existing variable, varname, with the same storage type, values, and display format as varname.
Codebook		
codebook(varlist)		Display codebook for variables
des		Describe data in memory or in file
des		Describe data in memory or in file
des [varlist]		describe variables in memory

Key Stata Commands for Constructing Variables (Cont.)

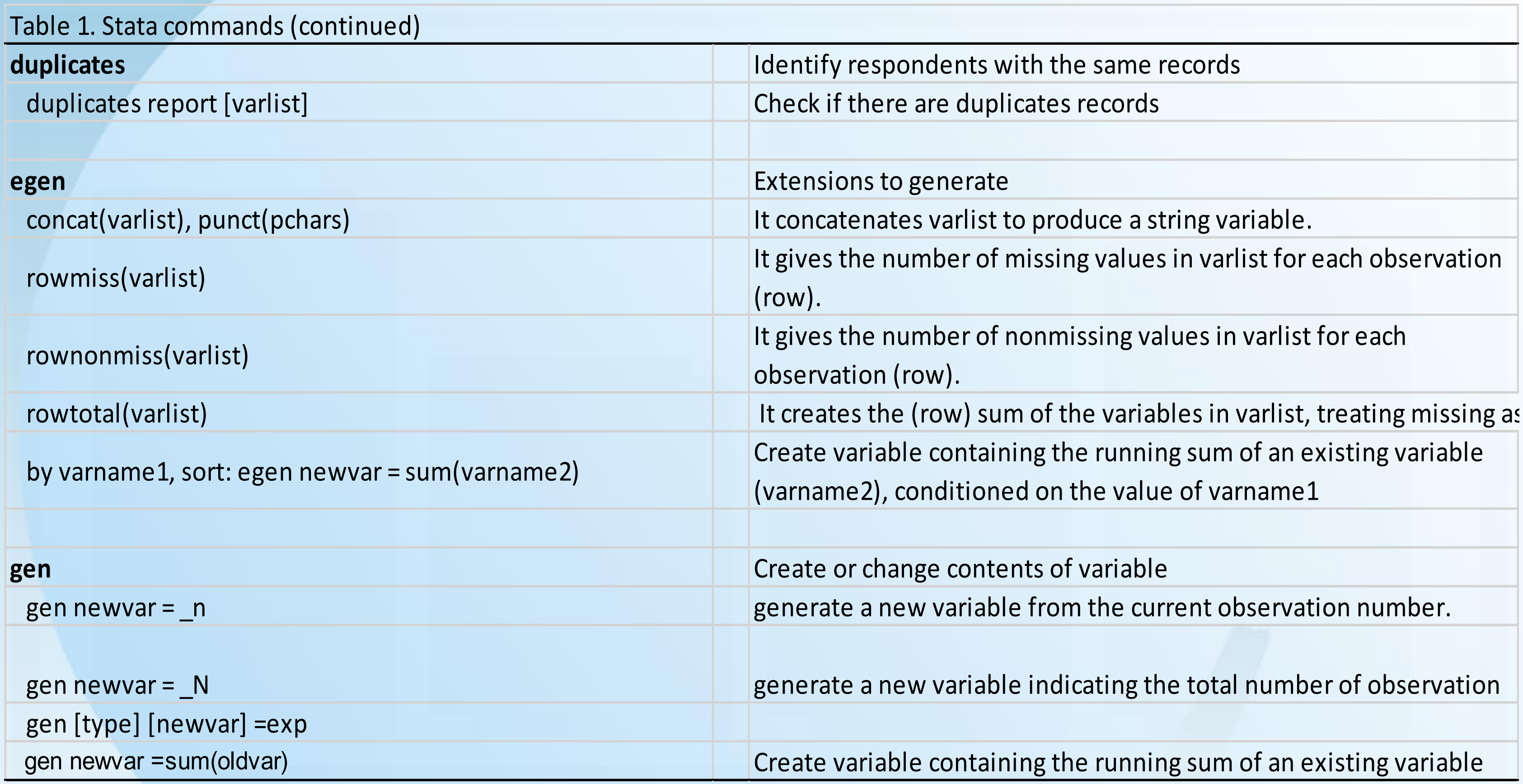

Table 1. Stata commands (continued)

duplicates	Identify respondents with the same records
duplicates report [varlist]	Check if there are duplicates records
egen	Extensions to generate
concat(varlist), punct(pchars)	It concatenates varlist to produce a string variable.
rowmiss(varlist)	It gives the number of missing values in varlist for each observation (row).
rownonmiss(varlist)	It gives the number of nonmissing values in varlist for each observation (row).
rowtotal(varlist)	It creates the (row) sum of the variables in varlist, treating missing as
by varname1, sort: egen newvar = sum(varname2)	Create variable containing the running sum of an existing variable (varname2), conditioned on the value of varname1
gen	Create or change contents of variable
gen newvar = _n	generate a new variable from the current observation number.
gen newvar = _N	generate a new variable indicating the total number of observation
gen [type] [newvar] =exp	
gen newvar =sum(oldvar)	Create variable containing the running sum of an existing variable

Key Stata Commands for Constructing Variables (Cont.)

Table 1. Stata commands (continued)	
If	The if expression at the end of the command applies the command to records meeting the specification of the if expression.
if varname == value	The condition is defined as a specific value of a variable
if inlist(variname, value1, value2,…)	The condition is defined as a list of values of a variable
if inrange(variname, value1, value2)	The condition is defined as a range of values of a variable
label	Manipulate labels
label variable [varname] ["label"]	Label variable
label define [lblname] # "label" [# "label" ...] [, add modify replace]	Define value label
label values varlist [lblname|.] [, nofix]	Assign value label to variables
label drop {lblname [lblname ...] | _all}	Drop value labels
list	List values of variables
list [varlist]	List values of variables
list [varlist], sepby (varlist2)	List values of variables and draw a separator line whenever varlist2 values change
list [varlist] [if]	List values of variables for records meeting the if condition
list [varlist], nol	List values of variables without the value label

BGSU

Center for Family and Demographic Research

Key Stata Commands for Constructing Variables (Cont.)

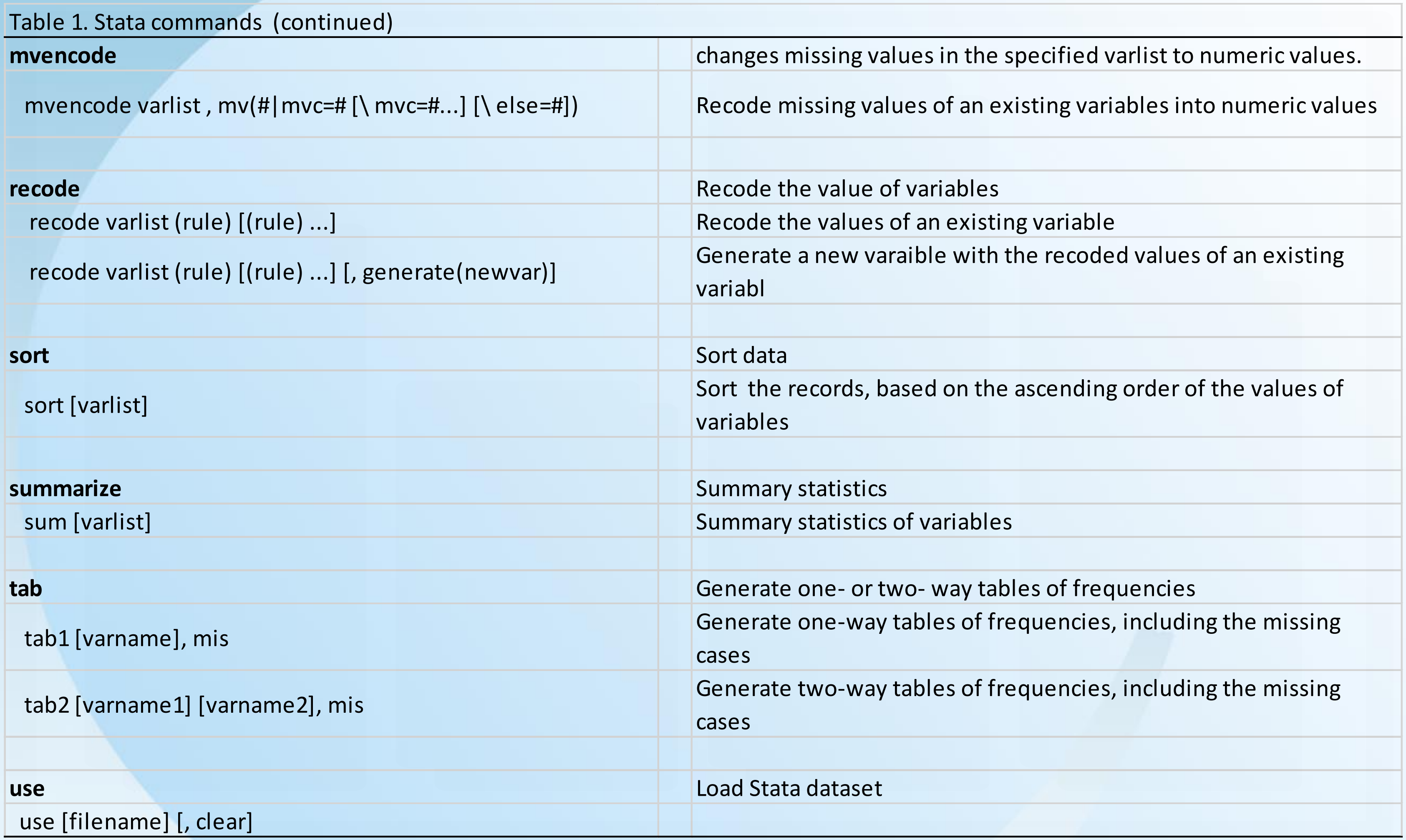

Table 1. Stata commands (continued)

mvencode		changes missing values in the specified varlist to numeric values.
mvencode varlist , mv(#|mvc=# [\ mvc=#...] [\ else=#])		Recode missing values of an existing variables into numeric values
recode		Recode the value of variables
recode varlist (rule) [(rule) ...]		Recode the values of an existing variable
recode varlist (rule) [(rule) ...] [, generate(newvar)]		Generate a new varaible with the recoded values of an existing variabl
sort		Sort data
sort [varlist]		Sort the records, based on the ascending order of the values of variables
summarize		Summary statistics
sum [varlist]		Summary statistics of variables
tab		Generate one- or two- way tables of frequencies
tab1 [varname], mis		Generate one-way tables of frequencies, including the missing cases
tab2 [varname1] [varname2], mis		Generate two-way tables of frequencies, including the missing cases
use		Load Stata dataset
use [filename] [, clear]		

Stata Example 1

Example 1: Constructing variables using data in wide format

Research question: how are race and experience of violence related to depressive symptoms?

Data: public data of Add Health at Wave III

The number of respondents in the original data: 6,504

The numbers of respondents with valid values of racial backgrounds, experience of violence, or depressive symptoms change with the way these variables were constructed.

The numbers of respondents in the analytic sample also vary, depending on how variables are constructed.

Stata Example 2

Example 2: Constructing variables using data in wide format

Research question: how are age and smoking related to the probability of live birth?

Data: Female pregnancy data (NSFG 2017-2019). The data set contains both respondent-level variables and pregnancy-level variables. Pregnancy records are nested within the respondent-level variables.

The data file contains 3,709 respondents and 10,215 pregnancy records.

- 3,662 respondents and 10,007 pregnancy records reported live births
- 3,709 respondents and 20,215 pregnancy records have valid information on the respondent's age
- 1,719 respondents or 2,293 pregnancy records have valid information on smoking during pregnancy

The final analytic sample consists of 1,719 respondents and 2,572 pregnancy records.

Stata Examples

See the Stata command and log files

Conclusions

- Constructing variables is a critical task because researchers usually need to construct some variables on their own before they can conduct analysis
- Constructing variables is more than summing up the values of individual variables. Researchers should have justifications for which variables should be used and how the variables should be aggregated.
- When more than two variables are used to construct variables, it is important to think through what strategies can aggregate these variables without risking committing any errors.
- During the process of constructing variables, researchers should keep tract of the number of respondents with (or without) valid values of these variables. It takes more effort to achieve this when the long-format data are used.
- Constructing variables usually does not require advanced coding skills, but does require patience and lots of data-checking. Therefore, researchers should allow themselves sufficient time to do it.
- If you have any questions about data management problems, feel free to contact Hsueh-Sheng Wu @372-3119 or wuh@bgsu.edu