

Using Stata to Work with Long-Format Data

Hsueh-Sheng Wu
CFDR Workshop Series
November 6, 2023

BGSU



Center for Family and
Demographic Research

Outline

- Long format versus wide format data
- Why do we need data in long format?
- Variable description
- Common Stata commands for working with data in long format
- Examples:
 - Cross-sectional family relation data
 - What is the size of the family?
 - Do householders have kids living at home?
 - What is the gender of the spouse of the household head?
 - Finding change in an variable and create discrete-time event history data
 - Using the information from the previous time point to substitute the missing value
 - Determine whether the respondents entered marriage
 - Determine how many times respondents transited into marriage
 - Determine the time when respondents entered the first marriage
 - Remove data that occurred after respondents entered the first marriage
- Conclusions

BGSU

Long Format versus Wide Format Data

Sample data in Long Format			Sample Data in Wide Format								
Family ID	pernum	sex	Family ID	sex1	sex2	sex3	sex4	sex5	sex6	sex7	sex8
9	1	2	9	2	2
9	2	2	8212	1	2	1	1	2	2	1	2
8212	1	1									
8212	2	2									
8212	3	1									
8212	4	1									
8212	5	2									
8212	6	2									
8212	7	1									
8212	8	2									

- Long format has 10 rows of data; wide format has two rows.
- Both formats have the case ID variable: Family ID
- Long format has an index variable: pernum
- Long format has one sex variable, but wide format has eight because the information of the index variable is incorporated into the sex variables
- Both formats provide same information for individuals
- The index variable can indicate the ID within family or for age/time
- The ID variable indicates a family or individual.

Why Do We Need Data in Long Format?

- Data in long format preserve the original data structure, so it is easier to check the data and construct variables.
- Certain analyses (e.g., discrete-time survival analysis) require data in long format.
- Stata commands of constructing variables are different for data in long format than for data in wide format
- For types of variable and data constructions:
 - Reshape data between wide format and long format
 - Merge data in long format with data in wide format
 - Generate summary statistics
 - Copy information from one row to another (e.g., the gender of the spouse for the household head)
 - Create a time interval capturing the change of a status (e.g., the duration for individuals to change from being single to being married)

Variable Description

Variable Descriptio of the Sample Data			
Definition	Name	Value	Value label
Survey Year	year	2016	
Survey month	month	1-12	January - December
Household serial number	fam_id	1- 94097	
Person number in sample unit	pernum	1-16	
Person ID	id	20141000000601 - 2016120746002	
Relationship to household head	relate		
		101	Head/headholder
		201	Spouse
		301	Child
		501	Parent
		701	Sibling
		901	Grandchild
		1001	Other relatives, n.s.
		1114	Unmarried partner
		1115	Housemate/roommate
		1241	Roomer/boarder/lodger
		1242	Foster children
		1260	Other nonrelatives
Age	age		
		0-85	0-85 years old
Gender	sex		
		1	Male
		2	Female
Marital Status	marst		
		1	Married, spouse present
		2	Married, spouse absent
		3	Separated
		4	Divorced
		5	Widowed
		6	Never married/single
		9	Not in a union

Common Stata Commands for Working with Data in Long Format

Command commands for Working with Data in Long Format	
<code>reshape wide year month marst sex age t_time, i(id) j(time)</code> <code>reshape long year month marst sex age t_time, i(id) j(time)</code>	Change data between the wide format and the long format
<code>duplicates report fam_id pernum</code>	Check whether each record is a unique one
<code>merge 1:1 id using "the name of a data set"</code> <code>merge 1:m id using "the name of a data set"</code> <code>merge m:1 id using "the name of a data set"</code>	merge data in wide format to data in wide format merge data in wide format to data in long format merge data in long format to data in wide format
<code>sort id time</code>	Arrange data in a special order
<code>by id: gen n=_n</code>	Create an indicator variable for each record of the person or family
<code>by id: gen N=_N</code>	Calculate the total number of records of the person or family
<code>gen marst_r = marst</code> <code>replace marst_r = marst[_n-1] if marst_r ==. & marst_r[_n-1] ~=.</code>	Handling the missing value
<code>by id: replace c_mar = 1 if marst_r[_n-1] >=4 & (marst_r ==1 marst_r ==2)</code>	Code the transition into marriage
<code>by id: gen i_c_mar = sum(c_mar)</code>	Create an indicator for the time of entering marriage
<code>by id: egen s_c_mar = sum(c_mar)</code>	Create an indicator for number of times of entering marriage
<code>by id: gen time_mar1 = time if c_mar ==1 & i_c_mar ==1</code>	Extract the time when the first marriage take place
<code>by id: egen m_time_mar1 = max(time_mar1)</code>	Expand the time of the first transition in marital status to all records of the individual
<code>by id: drop if time > m_time_mar1</code>	Remove records that occurred after the first transition in marital status
<code>by id: replace s_sex = if pernum ~=1</code>	Replace the values in irrelevant records

Examples

- See accompanying Stata commands and log files

BGSU



Center for **Family and**
Demographic Research

Conclusions

- With data in long format, researchers usually create summary statistics, copy information from one row to another, or create a time interval capturing the change of a status.
- Stata commands greatly reduce the difficulty of accomplishing these tasks.
- When working with data in long format, be sure that there are no duplicate records per family or individual and that data are sorted correctly.
- You can use the `–reshape-` command to switch between long format and wide format to speed up the variable construction process.
- If you have any questions about long format data, please come see me at 5D, Williams Hall or send me an email (wuh@bgsu.edu).