STATA 3

Sara Nadel

sara_nadel@hksphd.harvard.edu

Getting Started

- Copy and paste the *MPAID_STATA_Session3* file from the shared folder to your personal folder
- Open STATA on your computer
  - Start → All Programs → STATA
- Open a *do-file*
  - Window → Do File Editor → New Do-File
  - (*also: CTRL-8)*
- Check the auto-save on do/run box
  - On do-file page: edit→ preferences → Auto-Save on Do/Run
- Open data
  - Type the following into do file:

```
clear
cd “m:\MPAID_STATA_Session3”
set mem 100M
use expchina.dta, clear
```

Do the do-file by clicking CTRL-D

The **egen** command

- Unlike **gen, egen** usually works within several data points of a single variable
- Go to **he egen** to see other uses of **egen**
- Examples, comparing gen v. egen:
  - gen tot_exp1 = sum(value)
  - egen tot_exp2 = sum(value)
  - bys importer: gen tot_exp1_byimp = sum(value)
  - bys importer: egen tot_exp2_byimp = sum(value)
- Other egen examples:
  - egen avg_value = mean(value)
  - bys importer year: egen max_value = max(value)

The **collapse** Command

- Shrinks your dataset horizontally by combining rows
- Go to **he collapse** for information on what the command looks like.
- Things to remember:
  - It only keeps the variables you specify
  - The first variables have to be numeric
  - Cannot be undone: must re-load original data to uncollapse
- **by** indicates the variable (or variable combination) that will be a unique indicator
- Examples:
  - collapse (sum) value, by(year importer)
  - collapse (sum) value, by(year sitc4)

The **merge** Command

- Joins two datasets that share at least one variable
- Used to combine data from multiple datasets that relate to the same topic.
- Go to **help merge**
  - Scroll down to description
  - Scroll down to examples at the bottom

Typical configuration

Master Database

var1	var2	var3	var4
1	S	1	A
1	S	3	B
1	S	4	C
1	s	5	D
.	.	.	.
.	.	.	.
.	.	.	.

Using Database

var4	var5	var6	var7
.	.	.	.
B	F	2	4
C	F	2	2
D	F	2	4
E	F	2	2
F	F	2	4
G	f	2	2

Typical configuration

Master Database

Using Database

var1	var2	var3	var4	var5	var6	var7
1	S	1	A	.	.	.
1	S	3	B	F	2	4
1	S	4	C	F	2	2
1	s	5	D	F	2	4
.	.	.	E	F	2	2
.	.	.	F	F	2	4
.	.	.	G	f	2	2

Typical configuration 2

Master Database

var1	var2	var3	var4
1	S	1	A
1	S	3	B
1	S	4	C
1	s	5	D
1	S	6	e
.	.	.	.

Using Database

var4	var5	var6	var7
.	.	.	.
B	F	2	4
B	F	2	2
D	F	2	4
E	F	2	2
F	F	2	4

Typical configuration 2

Master Database

Using Database

var1	var2	var3	var4	var5	var6	var7
1	S	1	A	.	.	.
1	S	3	B	F	2	4
1	S	3	B	F	2	2
1	S	4	C	.	.	.
1	s	5	D	F	2	4
1	S	6	e	.	.	.
.	.	.	E	F	2	2
.	.	.	F	F	2	4

Merging: the basics

- “Master” dataset: the one that is open
- “Using” dataset: the one you want to join with the master dataset
- Requirements:
  - There has to be at least one variable with the same name in both to be able to merge
  - Both master and using datasets must be sorted by the “link” variables
- COMMAND: merge *varlist* using *filename*

Merging

- What if the merging variable occurs multiple times in one of the databases?
- What if there is a second variable with the same name in both datasets and we don't put it in the command?
- How can you tell how many variables merged?
  - **tab _merge**

Exercise: Create a dataset with the following characteristics…

- Merges the education and household information of the two datasets *education.dta*, and *household.dta*
- Generates a variable for the number of individuals in the household
  - Hints:
    - Assume that each individual is represented in the education dataset
    - Generate a variable called one which equals 1 all the way down
    - Use bys: egen to sum the number of individuals in the household
- Determine how many meals are consumed on average by family size using…
  - tabstat
  - egen
  - collapse

Congratulations!

You are now all STATA Masters.