

Introduction: Merging Datasets In STATA

Jen Cocohoba, Pharm.D., MAS
Health Sciences Associate Clinical Professor
UCSF School of Pharmacy

NOTE: tips tested for STATA v.12, though most info will be relevant for STATA 13

A little on merging datasets

- Understanding the language
 - Merge = add new variables from 2nd dataset to existing obs (across)
 - Append = add new obs to existing variables (under)
- Merging requires datasets to have a common variable (ID)
- Nomenclature for the datasets
 - One dataset is defined as the “master” (in memory) dataset
 - The other dataset is called the “using” dataset
- Many types – need to specify what kind for STATA

One to One/One to many

- merge 1:1

master			
id	zipcode	sex	age
1	94110	0	65
2	92154	1	46
3	94117	1	36
4	94143	0	47

using

id	wt	sbp	fbg	ldl
1	70	110	80	100
2	54	140	76	120
3	76	180	160	160
4	50	105	126	102

- merge 1:m

master			
id	zipcode	sex	age
1	94110	0	65
2	92154	1	46
3	94117	1	36
4	94143	0	47

using

id	clinicdate	wt	sbp
1	7/6/10	70	110
1	8/10/11	65	120
2	7/6/10	54	140
2	9/10/11	57	130
3	7/6/10	76	180
4	7/6/10	50	105
4	8/11/11	51	110

Many to one, Many to Many

- merge m:1

master			
id	clinicdate	wt	sbp
1	7/6/10	70	110
1	8/10/11	65	120
2	7/6/10	54	140
2	9/10/11	57	130
3	7/6/10	76	180
4	7/6/10	50	105
4	8/11/11	51	110

using

id	zipcode	sex	age
1	94110	0	65
2	92154	1	46
3	94117	1	36
4	94143	0	47

- merge m:m

master			
id	clinicdate	wt	sbp
1	7/6/10	70	110
1	8/10/11	65	120
2	7/6/10	54	140
2	9/10/11	57	130
3	7/6/10	76	180
4	7/6/10	50	105
4	8/11/11	51	110

using

id	clinicdate	fbg	ldl
1	8/10/10	80	100
1	11/10/11	120	70
2	8/6/10	76	120
2	10/8/11	100	130
3	7/7/10	160	160
4	8/6/10	98	100
4	9/11/11	126	102

Merging datasets

- Need to make sure data are sorted by the common variable AND saved
- Steps
 - Load the master dataset into memory
 - Sort (just to be safe) & save
 - Merge command
 - Check to make sure it makes sense
- See appearance of a “merge” variable which tells you where the observations came from (dataset 1, dataset 2, etc.)

Example: with two datasets called “wihsdrugs” (master) and “socdem” (using)

```
use “socdem.dta”
sort wihsid
save “socdem.dta”,
replace
clear

use “wihsdrugs.dta”
sort wihsid
merge 1:1 wihsid using
“socdem.dta”
browse
```

Introduction: Reshaping data

Jen Cocohoba, Pharm.D., MAS
Health Sciences Associate Clinical Professor
UCSF School of Pharmacy

NOTE: tips tested for STATA v.12, though most info will be relevant for STATA 13

Shape-shifting

- Conceptually difficult

– Example: chart with patients and average # cigarettes smoked per day over time

idnumber	1981	1982	1983
1	25	10	8
2	14	30	17
3	2	18	4

WIDE

- May want data to look different to manipulate

idnumber	year	cigs
1	1981	25
1	1982	10
1	1983	8
2	1981	14
2	1982	30
2	1983	17
3	1981	2
3	1982	18
3	1983	4

LONG

Reshaping wide to long

- Wide to long: make dataset with multiple records per patient

- Group variables need common “stub”

– In our example 1982, 1983, 1984

– STATA doesn't know to group these unless named similarly

rename 1982 **cigs1982**

rename 1983 **cigs1983**

rename 1984 **cigs1984**

- Nomenclature

– i = primary index variable (the patient identification number)

– j = secondary index variable (often generated from a “stub”)

reshape long *cigs*, i(*idnumber*) j(*year*)*

← Stub ← New variable

Reshaping long to wide

- Long to wide: one record per patient (our example)

reshape wide *datavariable*,
i(*indexvariable*) j(*2nd-indexvar*)

indexvariable	2ndindex1	2ndindex2	2ndindex3
1	cigs@subindex1	cigs@subindex2	cigs@subindex3
2	cigs@subindex1	cigs@subindex2	cigs@subindex3

reshape wide *cigs*, i(*idnumber*)
j(*year*)

← Stub: to be created

← Existing variable (going to be dropped)