Subject: Merging issue
Posted by chopotin on Fri, 12 Jun 2015 22:42:58 GMT
View Forum Message <> Reply to Message

Hello!

I have Stata 12.1. For some reason I am having trouble merging the HIV dataset with the Individual dataset with Ethiopia 2011. I have merged these two types of datasets with other countries with no issues. However, after doing it with Ethiopia I get the following error message: "variable id does not uniquely identify observations in the master data".

Here is my process for merging these two datasets:
1) First I open the HIV dataset and put the code: gen long id=((1000+hivclust)*10000)+(hivnumb*100)+hivline

2) Then I save it and open the Individual dataset and put the code: gen long id=((1000+v001)*10000)+(v002*100)+v003

3) I leave the individual dataset open and put the code to merge: merge 1:1 id using "C:/Users/EthiopaHIV2011.dta"

Then I get the error message as shown above.

Why is unique about the Ethiopia Individual and/or HIV dataset compared to other country datasets where this code does not work?

Thanks for any help you can provide ahead of time!
Leo

Subject: Re: Merging issue
Posted by Bridgette-DHS on Wed, 17 Jun 2015 18:22:16 GMT
View Forum Message <> Reply to Message

Following is a response from DHS Senior Stata Specialist, Tom Pullum:

As I understand it, you want to merge the AR data with the IR data. Here are Stata lines to do this:

use c:\DHS\DHS_data\AR_files\ETar61FL.dta, clear
rename hivclust v001
rename hivnumb v002
rename hivline v003

```
sort v001 v002 v003
save c:\DHS\DHS_data\scratch\temp.dta, replace
use c:\DHS\DHS_data\IR_files\ETIR61FL.dta, clear
sort v001 v002 v003
merge v001 v002 v003 using c:\DHS\DHS_data\scratch\temp.dta
tab _merge
keep if _merge==3
drop _merge
```

You will need to change the paths, of course. I am using the old version of the merge command, but the version you used would work equally well. Your formula for an id code did not produce unique values. Look at the following results for the IR file:

```
gen long id=((1000+v001)*10000)+(v002*100)+v003

. gen n=1

. collapse (sum) n, by(id)

. tab n
```

There are 601 id codes that appear 2 times, 28 that appear 3 times, and 1 that appears 4 times. It is safer to use a hierarchical sort command, such as " sort v001 v002 v003". Also easier. To get a truly unique id you could use this: "egen id=group(v001 v002 v003)". Your formula with powers of 10 will not always work.

File Attachments

1) tab.jpg, downloaded 1121 times