Subject: Re: Merging three different files
Posted by Liz-DHS on Wed, 01 Jul 2015 19:00:23 GMT
View Forum Message <> Reply to Message

Dear User,
Here is a response from one of our technical experts, Dr. Tom Pullum:
Quote:I'll begin with some friendly suggestions, based on the Stata lines that you sent. First, I strongly advise against having more working files than you need. Second, as much as possible, use file names that are short and close to the original file names. (There are actually two child files, the KR and BR, and two household files, the HR and PR, and two surveys were done in Uganda in 2011, the DHS with code 60 and the AIS with code 6A. I am assuming that you are talking about the KR and PR and 60 files) Third, use the old version of merge, at least for as long as Stata will allow us to use it.

In general, when doing a merge, I start with the larger file and use the smaller file as the "using" file. I always think of how I would do the task manually, and that's how I would do it. It may not make any difference. However, I have had instances with merge and append when the operation would fail if I started with the smaller file.

In your merge, I would include hvidx, the line number in the PR file, and b16, the matching line number of the child that is found in the KR and BR files. Virtually all of the relevant household data is on every line in the PR file, so if you do a 1:1 merge you should get everything you need. (But with the old merge command you don't have to specify 1:1 or 1:m or m:1.)

I have a "scratch" folder where I put temporary files that I need only during file construction, such as a sorted file. So long as I save the program, it is not necessary to refer to the files again, so I put "temp" in the file name and they can be over-written.

Keep in mind that there are some children in the PR file who are not in the KR file. Those are children for whom the mother is not also in the household. There are some children in the KR file who are not in the PR file. These are children whose mother is in the household but the child is not. You need to decide whether you just want the children who are in both files.

To be sure that you have the cases you want, you could use _merge (as in tab _merge) but I will suggest another approach. Here is how I would do the merge, using the folders where I keep these files:

```
use c:\DHS\DHS_data\KR_files\UGKR60FL.dta, clear
gen hv001=v001
gen hv002=v002
gen hvidx=b16
gen in_KR=1
sort hv001 hv002 hvidx
save c:\DHS\DHS_Data\scratch\temp.dta, replace

use c:\DHS\DHS_data\PR_files\UGPR60FL.dta, clear
gen in_PR=1
sort hv001 hv002 hvidx
```

merge hv001 hv002 hvidx using c:\DHS\DHS_Data\scratch\temp.dta
tab _merge
tab in_KR in_PR,m

gen KR_PR_merge_result=.
replace KR_PR_merge_result=1 if in_KR==1 & in_PR==1
replace KR_PR_merge_result=2 if in_KR==. & in_PR==1
replace KR_PR_merge_result=3 if in_KR==1 & in_PR==.

label define KR_PR_merge_result 1 "In both PR and KR" 2 "In PR only" 3 "In KR only"
label values KR_PR_merge_result KR_PR_merge_result
tab KR_PR_merge_result _merge,m
drop _merge

* next save the file, all or just what you need

Quote: Here is the tab at the end:

KR_PR_merge_resul t	_merge 1	2	3	Total
In both PR and KR	0	0	6,898	6,898
In PR only	38,079	0	0	38,079
In KR only	0	980	0	980
Total	38,079	980	6,898	45,957

Quote:Most of the 38,079 cases that are in the PR file only are not children. You probably only want the 6,898 children who are in both files, but you may want the 980 who are only in the KR file.

This was a longer answer than necessary. Some steps could be done differently. Please re-post if you had something different in mind.