Subject: Merging KR file with PR file India DHS 2019-21 (NFHS-5)
Posted by preshit on Sat, 05 Aug 2023 13:40:27 GMT
View Forum Message <> Reply to Message

Hello,
I am attempting to merge the KR file with the PR file for India DHS 2019-21 (NFHS-5) dataset. I am following this post for NFHS-4, assuming it should work with NFHS-5 files as well.

Merging KR and BR data files [message #19180]:https://userforum.dhsprogram.com/index.php?t=msg&amp ;amp ;th=8882&start=0&

My codes are below:

****
*For KR file
****
use caseid v001 v002 b16 v024 v003 v004 using "IAKR7EFL.DTA", clear

keep if b16>0 & b16<. //this will remove children who are not listed in the HH list

rename v024 hv024
rename v001 hv001
rename v002 hv002
rename b16 hvidx
rename v003 hv003
rename v004 hv004

sort hv024 hv001 hv002 hv004 hvidx // hhid hv003

save "NFHS5_KRclean.dta", replace

****
*For PR file
****
clear all

use hhid hv001 hv002 hvidx hv003 hv004 hv024 using "IAPR7EFL.DTA", clear

sort hv024 hv001 hv002 hv004 hvidx //hhid hv003

merge m:m hv024 hv001 hv002 hv004 hvidx using "NFHS5_KRclean.dta"

Result   # of obs.

```
not matched                  2,621,943
    from master              2,621,942  (_merge==1)
    from using                       1  (_merge==2)

matched                        221,982  (_merge==3)
-----------------------------------------
```

As shown above, the merge did not match one observation from the KR file. I have further looked into the unmatched observation and have found that its b6/hvidx (line no.) from the KR file is not present in the PR file. Can DHS confirm this observation is additional in the KR file? Or is it a data processing error that can be rectified? I would also appreciate any feedback on my merging code if it is correct or if I need to modify it for correct merging.

I will appreciate any help with this.

Regards
Preshit

Subject: Re: Merging KR file with PR file India DHS 2019-21 (NFHS-5)
Posted by Bridgette-DHS on Mon, 14 Aug 2023 14:55:52 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

We apologize for the delay in this response, due to travel. I have slightly revised your code, as follows:


```
* specify workspace
cd e:\DHS\DHS_data\scratch

use "...IAPR7EFL.DTA", clear
keep hv001 hv002 hvidx hv024
gen cluster=hv001
gen hh=hv002
gen chline=hvidx
gen state=hv024
sort state cluster hh chline
save IAPR_temp.dta, replace


use "...IAKR7EFL.DTA", clear
keep v001 v002 b16 v024
keep if b16>0 & b16<.
gen cluster=v001
```

```
gen hh=v002
gen chline=b16
gen state=v024
sort state cluster hh chline

* At this point a merge will fail, because of a few duplicated values of b16 in the KR file.
* Identify the duplicates.
egen repeated_b16=seq(),by(cluster hh chline)
tab repeated_b16

* There are 7 households in which the same value of b16 appears twice
* Remove the second case in duplicates. Note: it is possible that in some households it is the
*   first, not the second, that should be removed. Values of b16 could be edited.
list if repeated_b16>1, table clean
drop if repeated_b16>1
drop repeated_b16

* sort again
sort state cluster hh chline
* Note that this is a 1:1 merge
merge 1:1 state cluster hh chline using IAPR_temp.dta
tab _merge

* There is one case with _merge=1. This is a child in the KR file with a valid value of b16 but
*  the value does not appear in the PR file. Should not happen; drop this case but could edit b16.
list if _merge==1, table clean
drop if _merge==1
```

There are 3 issues. First, the KR file includes 7 households in which the same value of b16 is repeated for different children. I simply drop the repeats. Ideally, someone could probably figure out a better solution for those 7 cases but there are nearly 222,000 children and the effect is negligible. Second, the KR file includes one child who has a valid value of b16 but that line number is not found in the household in the PR file. Again, I just drop the case but there are other options. Third, you do this as a m:m merge but it should be 1:1. The children with _merge=3 will be matched in the two files.