Understanding Society User Support - Support #1298

Matching youth data to parent data in Understanding Society

01/11/2020 08:29 AM - Paul Downward

Status:	Resolved	**Start date:**	01/11/2020
Priority:	Urgent	**% Done:**	100%
Assignee:	Paul Downward		
Category:			

Description

Dear colleague,
I wonder if you can help me with the above. I have used the USS before and merged and matched waves and, following your online course, matched adults in a household. I am now experimenting with matching individuals from the youth file to, say, their mothers and whist I can match the files I end up with very small matched samples and wonder if I am doing something silly.

To illustrate based on some reduced files - I have also been saving replacing the files as I go to check each step - as I have learned syntax as I address specific projects

If I create a 'mum' file with a few variables

```
use "C:\ukhls_w8\h_indresp.dta"
keep if h_sex==2
save "C:\ukhls_w8\mum data.dta",replace
use "C:\ukhls_w8\mum data.dta"
keep pidp h_sex h_hidp h_scsf1 h_mnspid h_pno h_childpno h_intdaty_dv h_dvage
drop if h_mnspid==-8
rename (h_sex h_hidp h_scsf1 h_mnspid h_pno h_childpno h_intdaty_dv h_dvage) (msex mhidp mscsf1 mnspid mpno mchildpno mintdaty_dv dvage)
save "C:\ukhls_w8\mum data.dta",replace
```

I then created a youth file with a couple of variables in

```
use "C:\ukhls_w8\h_youth.dta"
keep pidp h_mnspid h_ypsrhlth h_hidp h_dvage
drop if h_mnspid==-8
rename (h_mnspid h_ypsrhlth h_hidp h_dvage) (mnspid yypsrhlth yhidp ydvage)
save "C:\ukhls_w8\youth data.dta",replace
```

I have then tried an m:1 merge

```
Use C:\ukhls_w8\youth data.dta"
merge m:1 mnspid using "C:\ukhls_w8\mum data.dta"
save "C:\ukhls_w8\Total W8.dta",replace
```

I get the message

```
variable mnspid does not uniquely identify observations in the using data
```

So, I checked the duplicates in the mum file and I get

```
duplicates report, mnspid

Duplicates in terms of all variables

--------------------------------------
copies | observations       surplus
----------+---------------------------
1 |        2738           0
--------------------------------------
```

and in the youth file I get

```
Duplicates in terms of all variables

--------------------------------------
copies | observations       surplus
----------+---------------------------
```

```
1 |        3174           0
--------------------------------------
```

So there doesn't seem to be an issue with duplicates but, if I repeat the steps above and this time remove the duplicates by force

e.g. in the mum file
. duplicates drop mnspid, force

Duplicates in terms of mnspid

(413 observations deleted)

and I leave any duplicates in the youth file as I assume it makes sense that there are duplicates of mnspid as a parent can be shared.

If I now follow the m:1 merge above I get

```
Result                           # of obs.
    -----------------------------------------
    not matched                         4,458
        from master                     2,598  (_merge==1)
        from using                      1,860  (_merge==2)

matched                               576  (_merge==3)
    -----------------------------------------
```

This seems to be a very small subset of cases. Am I doing the right thing here? Any help would be greatly appreciated. My plan was to do this for each wave and then append the waves.

Thank you.

Paul.

History

#1 - 01/13/2020 05:29 PM - Stephanie Auty

- Status changed from New to Feedback

- Assignee set to Paul Downward

- % Done changed from 0 to 60

- Private changed from Yes to No

Dear Paul,

In the w_indresp files, w_mnspid is the identifier for that individual's mother, while pidp is the identifier for that individual. In w_youth, w_mnspid is the identifier for that young person's mother, while pidp is the identifier for that young person. To match young people to their mothers, you need to match the young person's mother's ID with their mother's own ID, that is, match w_mnspid in the w_youth file to pidp in the w_indresp file. To do this you need to rename one of these variables so they have the same name and then merge on that variable name. Since you are renaming all of the mother variables, you could drop mnspid from indresp, then rename pidp to mnspid, then merge on mnspid.

As you are planning to append files from each wave and have removed the wave prefixes, remember to create a wave variable otherwise you will not know which wave each observation has come from.

When you use duplicates report, do not use a comma before the variable name. Stata thinks you are trying to look for duplicates in terms of all variables as anything after the comma is an option rather than a parameter.

Best wishes,
Stephanie

#2 - 01/13/2020 06:18 PM - Paul Downward

Hi Stefanie,

This is very helpful indeed. I noticed from the reply to Understanding Society User Support issue#252 that this strategy was recommended. I have experimented with this and it is working.

Thank you very much. A great help.

Paul.

#3 - 01/16/2020 12:34 PM - Stephanie Auty

- Status changed from Feedback to Resolved

- % Done changed from 60 to 100