Understanding Society User Support - Support #1949

Merging two datasets

08/03/2023 11:08 AM - Taiba Chaudhry

Status:	Resolved	**Start date:**	08/03/2023
Priority:	Normal	**% Done:**	100%
Assignee:			
Category:	Data management		

Description

Hi, I am currently trying to merge g.indresp and g.youth together. I have tried using the household identifier g.hidp, but an error message appears stating that the id is not uniquely identified. I would appreciate any help in terms of merging these two datasets. Thank you.

History

#1 - 08/03/2023 11:14 AM - Taiba Chaudhry

I am trying to look at parental income and children health. By merging the two datasets I can see the children within each household alongside their parents. I have tried merging but hasn't worked. I have been looking into reshaping the data and then merging but still not sure.

#2 - 08/03/2023 12:02 PM - Understanding Society User Support Team

- Private changed from Yes to No

Hi Taiba,

Which statistical package do you use?

Best wishes,
UKHLS User Support

#3 - 08/03/2023 01:16 PM - Taiba Chaudhry

I am using Stata

#4 - 08/03/2023 01:53 PM - Understanding Society User Support Team

- Category changed from Data analysis to Data management

- Status changed from New to Feedback

- Assignee deleted (Alita Nandi)

- % Done changed from 0 to 80

This happens because these are individual level files, a lot of individuals live together in the same household therefore, in both datasets, a given g_hidp number can appear more than once. To merge these files without an error message you would need to use merge m:m:

use g_indresp, clear
merge m:m l_hidp using g_youth

However, for the research problem you described it is perhaps better to merge these datasets using parents identifiers available in the youth datafile, g_fnspid (father) and g_mnspid. These variables tells you the pidp of, respectively, father and mother of the child. You can use the syntax file available on our website:
https://www.understandingsociety.ac.uk/sites/default/files/downloads/documentation/mainstage/syntax/stata/stata-parents-children-matching.do

Best wishes,
Piotr,
UKHLS User Support

#5 - 08/03/2023 02:10 PM - Taiba Chaudhry

Thank you for getting back to me. I have done a few steps so far.

I merged the individual and household data files, that is indresp and hhresp to get household data such as gross household income. The command I did was
use individual_data
merge m:1 g_hidp using household_data.dta

Then I merged this newly individual-household data with the child data using the following command.
merge 1:1 g_hidp g_pno using child_data.dta

However, as you said it may be better to merge based on the parents ID. I could create a separate sample for mothers and fathers and see how that affects the child health. Is my thinking correct?

But I was also wondering would I have to create household level variables because there maybe multiple children within the household. For example creating the mothers qualification at the household level. Or would this not be a problem to worry about.

#6 - 08/03/2023 05:07 PM - Understanding Society User Support Team

Hi Taiba,

The step 1 is fine, you correctly linked the hhresp to indresp. However, as you are using indresp which includes only respondents 16 years old and older, in your second step

> Then I merged this newly individual-household data with the child data using the following command.
> merge 1:1 g_hidp g_pno using child_data.dta

you will have 0 matches.

This because in the master dataset you have only adults and in the linked child data only children (regardless whether it is child or youth file, both are at the child level=each row is one child and the pno refers to the child).

My advice would be to:
1) make a plan of your analysis, write down what exactly you need to get,
2) check where this information is stored, what the level of these files is (e.g., child or adult), what identifiers you need and how you can use them. Everytime you are merging two files you need to carefully check what each identifier means in a given context, e.g. if you mechanically tried to merge indresp with youth using mnspid, it would work, but you would be merging siblings. So, sometimes you need to change the name of some identifiers in some of the files to get the result you're looking for.

If you are planning merging parents information to children, then I'd recommend checking this stata syntax:
https://www.understandingsociety.ac.uk/sites/default/files/downloads/documentation/mainstage/syntax/stata/stata-parents-children-matching.do It does all of the work for you so it will save a lot of time.

Best wishes,
Piotr Marzec,
UKHLS User Support

#7 - 11/30/2023 01:23 PM - Understanding Society User Support Team

- Status changed from Feedback to Resolved

- % Done changed from 80 to 100