Understanding Society User Support - Support #792

Merging datasets

06/07/2017 03:21 PM - Emily Lowthian

Status:	Closed	**Start date:**	06/07/2017
Priority:	High	**% Done:**	100%
Assignee:	Emily Lowthian		
Category:	Data analysis		

Description

Hi there,

My name is Emily - I am a MSc student, currently working on a dissertation looking at how socioeconomic status moderates the impacts of trauma on adolescent well-being.

I essentially need to merge 2 datasets in wave 4 - d_indsamp and d_indresp. I am using stata V14.2 for this.
However, despite trying to use the pidp and d_pno as the unique identifiers Stata keeps flagging up an error that the ID variables do not uniquely identify observations in the master dataset (indsamp).
I'm not entirely sure why this is happening - please could you help?

Additionally - I would like to merge on the wave 4 hhresp dataset on to this (I will just duplicate the hh responses on to each indivdiual). Then wave 6 youth dataset will be appended on based using hh ID so the children sit in the parental household and it can be submerged later on.

Any help will be greatly appreciated as I have hit a wall here!

Emily

History

#1 - 06/07/2017 03:22 PM - Emily Lowthian

Stata code:

use d_indresp, clear
sort pidp
describe
save "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 4\d_indresp", replace

```
use d_indsamp, clear
    sort pidp
    describe
    save "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 4\d_indsamp", replace

use d_indsamp, clear

merge 1:1 pidp d_pno using d_indall
```

#2 - 06/17/2017 01:27 PM - Alita Nandi

- Status changed from New to In Progress

- Assignee changed from Alita Nandi to Emily Lowthian

- % Done changed from 0 to 70

- Private changed from Yes to No

Hello Emily,

Each row in INDSAMP is uniquely identified by PIDP and D_FINLOC
Each row in INDALL is uniquely identified by PIDP

So if you want to match these files you will have to do a many to one matching.

use d_indsamp, clear
merge m:1 pidp using d_indall

However, before doing that please consider what D_INDSAMP is useful for and what D_FINLOC means. For example, if you are only interested in

the final location of the person then you should first select d_finloc=1 cases and then do a 1:1 matching on PIDP.

use d_indsamp, clear
keep if d_finloc==1
merge 1:1 pidp using d_indall

Best wishes,
Alita

#3 - 06/19/2017 01:54 PM - Emily Lowthian

Hi Alita,

Thank you for getting back to me.

My apologies but I realised that I did not need to merge these in the end because I noticed that Indresp and indsamp merged together allowed me to get all the data I needed without merging lots of little datasets - still learning!

I have so far merged indresp and indsamp (both wave 4) on stata like this:
use d_indresp, clear
sort d_hidp pidp
describe
save "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 4\d_indresp_sort"

```
use d_indsamp, clear
        sort d_hidp pidp
        describe
        save "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 4\d_indsamp_sort"

use d_indresp_sort, clear //master file//
            merge 1:1 d_hidp pidp using d_indsamp_sort //using file//
            sum _merge
            save "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 4\d_indsamp+indresp"
```

I then have merged indresp and indsamp with hhresp (all wave 4) like this - I followed the guidance on the moodle:
use d_indsamp+indresp, clear // master file //
drop _merge
merge m:1 d_hidp using d_hhresp // using file //

save "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 4\d_indsamp+indresp+hhresp" //original saved on 14/06/2017//

I've checked around 1,000 cases and 8 variables comparing the raw to the merge and they look to have merged correctly for the wave 4 dataset. I am now at the stage where I want to merge Wave 6 youth cases on. However, I am unsure whether to merge the variables from Wave 6 on to the Wave 4 dataset I have created first, or to append the Wave 6 cases on first. Any support regarding this would be appreciated! So far I have written this, I am yet to run it as I am tied up with other things!

- Generating Wave Number and removing prefix d_ in W4 adult** EL 15/6/17

use "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 4\d_indsamp+indresp+hhresp", clear // master file //
drop *merge*
gen wave = 4
rename d **
save "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 4\d_indsamp+indresp+hhresp-d_"

- Generating Wave Number and removing prefix f_ in W6 youth EL 15/6/17
  use "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 6\f_youth", clear //using file//
  sort pidp
  gen wave = 6
  rename f_ * *
  save "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 6\f_youth-f"

- Merging on youth data ** EL 15/6/17
  use "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 4\d_indsamp+indresp+hhresp-d_", clear //master file//

merge 1:1 hidp pidp using "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 6\f_youth-f" //using file//
sum _merge

save "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 6\W4indsamp+W4indresp+W4hhresp+W6youth"

- appending on W6 cases ** EL 15/6/17

use "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 6\W4indsamp+W4indresp+W4hhresp+W6youth", clear
append using "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 6\f_youth-f"
save "H:\My Documents\SSRM\Dissertation\Data\Workable files\Wave 6\W4indsamp+W4indresp+W4hhresp+W6youth_FINAL"

Many (super) thanks for this,

Emily

#4 - 06/27/2017 12:58 PM - Alita Nandi

Hello Emily

Our remit at the User Forum is to answer queries related to Understanding Society data and provide general advice about how to manage the data. Given the number of users we have I'm afraid we cannot advise on individual users' analysis syntax specifically. We provide online and in person training to use the data set and for setting up datasets for different kinds of analysis. As you are following that then that should be fine.

My understanding is that you are interested in analysing Wave 6 youth respondents and want to use information collected about them or their family members in Wave 4. If that is the case, then you should merge the Wave 6 youth file with Wave 4 datasets. But you are suggesting merging these two files using hidp pidp - this is not possible as hidp is not unique across waves. W_HIDP is unique within a wave. Note that we provide identifiers for father, mother, grandmother, grandfather, spouse and partner. You can use these to directly link to parents and grandparents information. Also, the Wave 4 merged file you are producing is a multilevel file - please be careful how you use the information in that file.

My suggestion (based on my understanding about your research qs) is: First take the Wave 6 youth file. Then merge father and mother IDs (F_FNSPID F_MNSPID) from F_INDALL. Separately create a Wave 4 individual level file by merging D_INDRESP with D_HHRESP using D_HIDP. Then take the Wave 6 file, rename F_FNSPID to PIDP and merge with the Wave 4 file you just created and save as Father_File after renaming the variables you need (e.g, D_DVAGE Father_DVAGE). Do the same for mother's information. Then you will have a Wave 6 youth file with father and mother information from Wave 4 attached. As the household level information is also attached to the Wave 4 individual level file you can attach those via the father or mother. NOTE: I am only providing the general principles for this type of data management. You should check every step carefully and write the code that best suits your purpose.

Best wishes,
Alita

#5 - 06/27/2017 02:39 PM - Emily Lowthian

Hi Alita,

Thank you for your detailed reply. I completely understand and I am sorry to have burdened you with all of the code! I have signed up to some training but I am slightly late to the party and I am currently on a waiting list.

Thank you for providing that information about hidp, pidp etc - that is all very useful! I did plan on merging them by hidp, so I am glad I did not try that. I have done a similar thing to what you have suggested, so I made the wave 4 dataset which includes indsamp, indresp and hhresp and then I appended the youth data on. This is less elaborate then what you have suggested but I think for my current skil set what I have done should be suitable; I also have checked every step very carefully. I can use the identifers to link youth's to their mother/father which should help me a lot - thank you for that information that is key here.

Could I check that FNSPID and MNSPID are the same across waves? And these match on to the variable PIDP or PID? that is the key information so I can match and submerge information.

Thank you so much - it is super appreciated

Em

#6 - 06/27/2017 02:58 PM - Alita Nandi

- % Done changed from 70 to 90

FNSPID and MNSPID are the PIDP of the father and mother not PID (even though the variable name as PID instead of PIDP!) These variables have a wave prefix which reflects that these are the PIDP of the parents who are living with them in the same household in that wave. In most cases, there will be no change. The only change you will see is that the child is no longer living with the parent (say parents separate and one parent moves into a different household). In that case this variable will have a value of -8. But if the actual identity of the parent changes (say a parent dies, the other parent remarries and this new spouse adopts the child) then the value of this variable will change across waves.

When you say you merged INDSAMP, INDRESP and HHRESP - have you made sure that you have linked the INDSAMP with INDRESP only after deleting FINLOC=0 cases from INDSAMP?

Best wishes,
Alita

#7 - 06/27/2017 06:13 PM - Emily Lowthian

Hi Alita,

Thanks for getting back to me so quickly.
Ok that's perfect I can definitely use that then - I will just have to accept where things change or cannot be followed up! I have found that in research you cannot account for every single thing!!

A short answer to your question is no I did not remove finloc= 0 when merging indsamp.. Is it possible to remove the finloc=0 cases now with the final merged dataset? As you mentioned they were duplicates from moved houses which I would not like to include in the analysis for obvious reasons. Or

is it a case of having to remerge?

Thanks so much!

Em :-)

#8 - 07/18/2017 03:39 PM - Alita Nandi

Hello Emily,

Sorry for the delay in replying.

You should remove the finloc=0 cases at the beginning, that is first remove these cases and then do the merging.

Best wishes,
Alita

#9 - 08/08/2017 03:53 PM - Stephanie Auty

- Status changed from In Progress to Resolved

- % Done changed from 90 to 100

#10 - 08/30/2017 09:50 AM - Stephanie Auty

- Status changed from Resolved to Closed