Understanding Society User Support - Support #21

Merging household and individual data set-wave1, 2009-2010

01/20/2012 10:25 AM - Anita Staneva

Status:	Closed	**Start date:**	01/20/2012
Priority:	High	**% Done:**	50%
Assignee:	Redmine Admin		
Category:	Data analysis		

Description

I want to merge the data set into one file. I started first with merging all household files. Let's say I want to merge a_hhsamp with the a_hhresp. I suppose to keep (1 3) of the resulting merge, however I had only merge= 2 3, or I have exactly the number of the first household file in 2, which means the two files were not merged.
I try with distributing household level information to the individual level, where I am using the a_hidp as identifier and follow the example you gave in the documents. Now my merge is fine, but by keeping merge 1 and 3 my sample size increase dramatically and I had duplicate observations.
Next I continue with the I individual files, where I am using a_hidp and a_pno as unique identifier in order to match correctly individual files, however again the resulting merge is not fine.
Could you advise me please how to deal with matching the files? Do you have some users do files which would help us to combine all the data sets from the wave 1, 2009-2010?
Many thanks
Anita

History

#1 - 01/20/2012 01:29 PM - Redmine Admin

- Category set to Data analysis

- Status changed from New to In Progress

- Assignee set to Redmine Admin

- % Done changed from 0 to 50

Anita,

I have tried to reconstruct your example here:

```
use a_hidp using a_hhresp,clear
merge 1:1 a_hidp using a_hhsamp,keepus(a_ivfho_dv)
table a_ivfho_dv _m,row col
```

household response outcome	_merge using only (2)	_merge matched (3)	Total
f2f - all eligible hh intv		21,694	21,694
f2f - interviews + proxies		2,630	2,630
f2f - interviews + refusal		5,708	5,708
hh comp + ques only		137	137
lost capi interview	21		21
demolished/derelict	605		605
building not complete	133		133
institution, not private hh	198		198
no hh member contact	2,240		2,240
unable to locate address	201		201
contact made but not with correct people	526		526
unknown eligibility	483		483
other non-contact	3,121		3,121
refus to rsrch cntre	976		976
refusal to intviewer	17,183		17,183
language problems	531		531
other ineligible	38,921		38,921
Total	65,139	30,169	95,308

The master data set is a_hhresp.dta, the using data set is a_hhsamp.
The households that match (_m==3) are those with a productive interview outcome, while the unmatched households are those with unproductive

outcomes (_m==2).
This fits with the description of a_hhsamp as the data file with data on all enumerated households and a_hhresp for all responding households.
If we had chosen to open a_hhsamp first and then merged it to a_hhresp, the results would have been the same except for the _merge variable would have had the values 1 and 3 instead.

> Next I continue with the I individual files, where I am using a_hidp and a_pno as unique identifier in order to match correctly individual files, however again the resulting merge is not fine.

You can use pidp as the personal identifier on all individual level data files.
Do you have a specific example here?

Some more general advice...
The data are released in a set of data files that allows users to construct working data sets for a multitude of purposes. Due to the relative complex data structure, we recommend that you study the questionnaires and online data documentation and select the variables you need for a given study purpose. In that way, the working data sets remain of a manageable size and there should also be less scope for confusing variables with similar names but different meaning on from different files.
See also free course materials from some of our training courses or news of forthcoming training courses

Hth

Jakob Petersen

#2 - 02/23/2012 01:38 PM - Redmine Admin

- Status changed from In Progress to Closed

#3 - 11/10/2015 03:01 PM - Gundi Knies

- Target version set to M1