Subject: Merging Individual and HH Member Recode Files
Posted by amw289 on Tue, 26 Jul 2016 19:37:15 GMT
View Forum Message <> Reply to Message

Hello,

I am working with the continuous Peruvian DHS (2004-2012). I have created two data sets. The first is an appended file of all individual data files from 2004-2012. The second is an appended file of all household member data files from 2004-2012. For brevity, I refer to these as the "individual" and "household member" files below. Here are the steps I have taken:

1. Renamed hvidx to v003 in the household member file (to match the variable name in the individual file)
2. Attempted to merge from the individual file to the household member as follows:
   1:m v000 v001 v002 v003 v007

   *** When I did this, stata informed me that v000 v001 v002 v003 and v007 do not uniquely identify observations in the individual file. Upon further investigation, I discovered
   288 duplicate observations in terms of v000 v001 v002 v003 v007. All duplicates are from the year 2012. Is this an accident? If not, how can I identify unique individuals and
   create identifiers for them so that they can be merged with the household member dataset? ***

3. I then forcibly dropped the duplicate observations from the individual data set (just to see if I could successfully merge without them anyway).
4. Again I attempted to merge the individual file to the household member as follows:
   1:m v000 v001 v002 v003 v007
   The result of the attempted merge is as follows:

Result	# of obs.	
not matched	460,200	
from master	29	(_merge==1) *these are observations from the individual file
from using	460,171	(_merge==2) *these are observations from the household member file
matched	135,622	(_merge==3)

   ***Of the 460,171 observations not matched from the household member file, 78% have eligible women in the household. What am I doing wrong? Why don't these households match to
   any women in the individual file?***
   Why are there 29 women who don't match to any household member? Should all individual women have a corresponding household?

Thank you so much in advance for your help. This has been perplexing me for quite some time!

Subject: Re: Merging Individual and HH Member Recode Files
Posted by Bridgette-DHS on Wed, 27 Jul 2016 21:12:06 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Stata Specialist, Tom Pullum:

You are probably working with successive rounds of the Peru CS, and the cluster numbers, in particular, are being re-used for different rounds. That is, cluster 1, say, does not refer to the same cluster for all rounds. That would be a major problem. The households are numbered within clusters and the line numbers are within household, so the problem is mainly with the cluster id codes, which are being re-cycled. Each round has different cases in it. The Peru CS is not longitudinal; it is successive cross-sections.

The safest thing to do would be to do these merges within each round. Add a variable which is "round" or something like that. After you have done all the merges within rounds, append them to make one long file. Then "round" will be part of the id information. If you use svyset you will have to construct unique identifiers for the strata, with something like "egen strata=group(round v023)". You may want to re-normalize the weights, as described in several other posts. But I recommend that you only use the pooled file for calculating differences and changes across rounds. If you try to calculate a mean, say, for the full file, it will not have a clear reference date and will be confusing.

Subject: Re: Merging Individual and HH Member Recode Files
Posted by amw289 on Sun, 31 Jul 2016 16:02:45 GMT
View Forum Message <> Reply to Message

Thank you for your response. I understand that cluster numbers are repeated across waves, which is why I added v007 to the end of the list of merging variables. v007 captures the year the survey data were collected (and years were not repeated across waves, as far as I know). I have tried your suggestion of merging individual and household member files separately for each wave, but I still run into the same two problems.

1. In 2012, there are 288 duplicate observations in terms of v001 v002 v003 in the individual file (PEIR6IFL). Yet there are no duplicates in terms of caseid. Isn't the caseid is based on v002 v002 and v003? If so, shouldn't there be no duplicates in terms of v001 v002 and v003? Please advise.

2. When I try merging the individual file to the household member file, separately by wave, I run into the same problem I described earlier, where there are many observations from the household member file that remain unmatched. For example, just using the data files from 2011, when I attempt a 1:m merge for (individual to household member), I get the following results in stata:

Result	# of obs.	
not matched	76,145	
from master	0	(_merge==1)
from using	76,145	(_merge==2) (these are all from the household member file)
matched	22,517	(_merge==3)

When I go to check if there are eligible women among the unmatched observations from the household members file (_merge==2) this is what I get:

Number of
eligible
women in HH Freq. Percent Cum.

0 19,335 25.39 25.39
1 41,003 53.85 79.24
2 12,080 15.86 95.11
3 2,926 3.84 98.95
4 699 0.92 99.87
5 85 0.11 99.98
6 9 0.01 99.99
7 8 0.01 100.00

Total 76,145 100.00

In other words, of the 76,145 unmatched household members from 2011, 79% reside in a household with at least one woman who was eligible to participate. Is this just telling me about non-response rates? Do I not need to worry about these households not being merged to observations in the individual file?

Thank you again for your time and assistance!

Subject: Re: Merging Individual and HH Member Recode Files
Posted by Bridgette-DHS on Mon, 01 Aug 2016 16:04:38 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Specialist, Shea Rutstein:

Since 2009, the sampling design of the Peru surveys consists of two nationally representative halves ("semesters"). For one semester, the same clusters are used from one of the previous year's semester, either the first or last, and a new set of clusters are selected for the other semester, which will be reused in the following year. A set of clusters is only used twice. Within both the reused and newly selected clusters, a complete household listing is done for each year and a new selection of households is done with a new numbering. Thus there is no guarantee that the same households are selected in neighboring years, and most likely they are not reselected. Therefore, only half of the clusters match but none of the households nor members.

For a single year, the household and individual data files can be merged by using the cluster and the household numbers. The semester is given in SHSEMES and SSEMES but I don't think it is necessary to use them for merging. I am not sure if I would follow a panel of clusters from one survey to the next. To do so, one would have to id each semester as to which clusters were in the

previous year and also denormalize the weights. The reuse of half the clusters was done to reduce the variance for trends.

Be careful about the sampling weights when pooling the years. The weights have been normalized to the sum of total households or individual interviews for each year.

Subject: Re: Merging Individual and HH Member Recode Files
Posted by amw289 on Tue, 02 Aug 2016 16:05:30 GMT
View Forum Message <> Reply to Message

Thank you for your response. It is very important to me that I understand exactly what is going on in the data so that I can make informed decisions about how to analyze the sample, so that I can accurately describe all the pertinent aspects of the data in a publication, and so that my results are reproducible. I continue to be confused about:

1. Why there are duplicate observations in terms of v001 v002 v003 when there are no duplicate observations of caseid in the individual data file in 2012. I would like to know exactly how caseid is constructed if not solely based on v001 v002 v003.

2. Why there are households with eligible women who did not participate in the household member file. Why did these women not participate? What should I make of these households?

Thank you so much again for your time and patience.

Subject: Re: Merging Individual and HH Member Recode Files
Posted by Bridgette-DHS on Thu, 11 Aug 2016 20:07:49 GMT
View Forum Message <> Reply to Message

Another response from Senior DHS Specialist, Shea Rutstein:

In 2012 and perhaps other years, there is an additional variable, hv002a, which identifies a separate household within the same dwelling (found during interviewing, such as another structure or the same structure in back with a separate entrance). This would seem to be the reason for duplicates.

From the individual survey data, hhid should match the hhid in the household data, then the specific household members record hvidx should match v003 from the individual file (not hv003 in the HH). Another way would be to create a variable by concatenating hvidx to the end of hhid in the household data file. This would match caseid in the individual file. Of course there are many more members of the household than interviewed women (only women 15-49 who slept the night before and were successfully interviewed are in the individual file).

I would NOT perform the matching on a pooled file. I would do it year by year and then pool.

Subject: Re: Merging Individual and HH Member Recode Files
Posted by amw289 on Tue, 16 Aug 2016 14:02:59 GMT
View Forum Message <> Reply to Message

Thank you again for your response and clarification. I continue to have problems identifying unique households withing 2012. In the Household Recode file:

"duplicates report hv000 hv001 hv002 hv002a" yields:

copies observations surplus

1 26481 0
2 570 285
3 117 78
4 40 30
5 10 8

This appears to be because all of the observations in the dataset are missing for hv002a. Please advise.

Thank you again!

Subject: Re: Merging Individual and HH Member Recode Files
Posted by Bridgette-DHS on Tue, 16 Aug 2016 16:13:30 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Stata Specialist, Tom Pullum:

I'm sorry that the previous responses did not resolve your problem. The Peru Continuous Survey is now operating almost completely independently of DHS.

It is clear that hv002a is not what you need. It is an empty variable, and why it is even included in the data is a mystery. I still have no idea what is the name of the variable that identifies sub-households.

As a kind of last resort, I have checked the household id variable, hhid. This is a character string that is usually made up from hv001 and hv002. In this survey, however, there are two additional characters at the end. Those last two characters appear to be the missing variable. It is possible to extract them as a separate variable.

If you do the following lines, you will extract that variable, which I call "final2":

* use PEPR6IFL.dta, clear
list hhid hv001 hv002 hvidx if _n<=50, table clean
* the last two characters of hhid must be extracted

```
gen str2 final2=substr(hhid,8,2)
destring final2, replace
tab final2
```

This gives you the variable you need in the HR or PR file.

To get it in the IR or KR file, you need to extract from caseid, as follows:

```
* use PEPR6IFL.dta, clear
list caseid v001 v002 v003 if _n<=50, table clean
* the last two characters of hhid must be extracted
gen str2 final2=substr(caseid,11,2)
destring final2, replace
tab final2
```

The substring command goes to columns 8-9 for hhid and 11-12 for caseid. You can identify those positions by trial and error.

After you have constructed "final2" or whatever you want to call it in every file, you would include it in the sort and merge and I think everything will be fine.

There are other ways to do this, even simpler, but possibly confusing, so I will just recommend this strategy.

Subject: Re: Merging Individual and HH Member Recode Files
Posted by amw289 on Tue, 16 Aug 2016 16:36:32 GMT
View Forum Message <> Reply to Message

Great, thank you so much! This worked and was very helpful.

Subject: Re: Merging Individual and HH Member Recode Files
Posted by rayhangog@gmail.com on Thu, 25 Jan 2018 17:55:19 GMT
View Forum Message <> Reply to Message

Dear,
I am trying to merge IR file with KR file, and PR file with KR+IR files from Peru's latest DHS data. But, in the first step, KR & IR does not match. I am following the below command in Stata;

```
*Merge child+women file
sort v001 v002 v003 merge
m:1 v001 v002 v003 final2 using "E:DataDHSPeruPEIR6IFL.DTA"
```

"variables v001 v002 v003 do not uniquely identify observations in the using data r(459);"

While the files from other countries have been merged with the same command. Please help me with this and provide me proper command

* Command for merging child+women and household member file (PR)
*rename in child file (KR+IR) . rename b16 mergeid . sort v001 v002 mergeid
*rename in household member file (PR)
rename hv001 v001
rename hv002 v002
rename hvidx mergeid
sort v001 v002 mergeid

Thank you

Subject: Re: Merging Individual and HH Member Recode Files
Posted by Bridgette-DHS on Sun, 28 Jan 2018 19:49:14 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Stata Specialist, Tom Pullum:

I don't know why you would want to merge the KR and IR files. Almost all of the information about the mother is on the child's record in the KR file. All the information about the children is on the mother's record in the IR file. No matter how you think of this merge, you will get a lot of duplicated information. It makes more sense to start with the PR file, merge the IR file with it, and then merge the KR file with the PR+IR file. You can then decide what cases to retain. The following Stata lines will do this.

* Prepare KR for merge
use e:\DHS\DHS_data\KR_files\PEKR6IFL.dta, clear
gen line=b16
sort v001 v002 line
save e:\DHS\DHS_data\scratch\KRtemp.dta, replace

* Prepare IR for merge
use e:\DHS\DHS_data\IR_files\PEIR6IFL.dta, clear
gen line=v003
sort v001 v002 line
save e:\DHS\DHS_data\scratch\IRtemp.dta, replace

* Prepare PR for merge
use e:\DHS\DHS_data\PR_files\PEPR6IFL.dta, clear
gen v001=hv001
gen v002=hv002
gen line=hvidx
sort v001 v002 line

```
* Merge IR and KR with PR
quietly  merge v001 v002 line using e:\DHS\DHS_data\scratch\IRtemp.dta
rename _merge _merge_IR
sort v001 v002 line
quietly merge v001 v002 line using e:\DHS\DHS_data\scratch\KRtemp.dta
rename _merge _merge_KR

* Decide which records to retain based on the combinations of the two _merge codes
tab _merge*,m
```

Subject: Re: Merging Individual and HH Member Recode Files
Posted by rayhangog@gmail.com on Sun, 28 Jan 2018 20:10:31 GMT
View Forum Message <> Reply to Message

Dear Pullum,
Thank you very much.
