Subject: Duplicates in IR file when merging
Posted by nholla on Wed, 25 Nov 2015 01:25:32 GMT
View Forum Message <> Reply to Message

Hi DHS,

I'm trying to merge the individual to the birth recode files for several countries using the following code in stata, where the birth recode is the "master" file:

merge m:1 v001 v002 v003 using "filename", gen(ir)

However, I'm getting a "does not uniquely identify observations in the using dataset" error. I observed duplicates in terms of v001 v002 v003 for the following files:

Mali 2001
Niger 2000
Niger 1992
Senegal 1998
Senegal 1986
Nigeria 1990

Can multiple women have the same "respondent" in the IR file? If so, how could this occur? For a few of them, I was able to find an additional identifier (such as sconces or snumber). For those surveys I can't find an additional identifier, should I be merging on caseid?

Thanks!

Subject: Re: Duplicates in IR file when merging
Posted by Bridgette-DHS on Wed, 02 Dec 2015 13:14:51 GMT
View Forum Message <> Reply to Message

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

For these surveys, unfortunately, another variable is required to identify households within households. If you look at hhid in most surveys, you will see that it is constructed by combining hv001 and hv002. In the surveys you listed, however, hhid includes a third variable. In the HR file for the Mali 2001 survey (MLHR41FL.dta), for example, there are four households with hv001=1 and hv002=3. For these sub-households, there is another variable that takes the values 8, 11, 27, and 50 that has been incorporated into hhid. In the Mali 2001 survey, this is a survey-specific household variable (prefix sh) called shconces.

If you look at the household questionnaire at the back of the main report on this survey, the top of the first page, you will see "numero de grappe", which is French for "cluster number", and right under that "numero de concession". So--for this survey you need to include shconces every time you sort and merge. In the other surveys you will have to hunt for that variable. There is probably a list somewhere of the name of this extra id variable--I don't think it is always called shconces.

I believe this code is not included in surveys more recent than the ones you listed. For example, looking at the Mali 2006 survey, I see that "numero de concession" is included on the household questionnaire, but hhid is constructed solely from hv001 and hv002. The easiest way to check whether this is an issue is to open the HR file and then enter the following:

gen n=1
collapse (sum) n, by(hv001 hv002)
tab n

You have a household id problem if "tab n" produces more values of n than n=1. If you do not have an HR file, you can use the PR file, enter "keep if hvidx==1", and then enter those three lines.

Using the full id for these surveys will be important if you are merging. It could be relevant if you are using the relation to head code (hv101). I will list some variables for the 13 cases in the Mali 2001 file with hv001=1 and hv002=3. For many kinds of analysis it is irrelevant. Good luck.

File Attachments
1) duplicates.jpg, downloaded 1665 times

Subject: Re: Duplicates in IR file when merging
Posted by nholla on Thu, 17 Dec 2015 16:19:23 GMT
View Forum Message <> Reply to Message

Thank you for your detailed explanation this. Given that the shconces variable is only on the household file, how could I merge this additional identifer onto the individual and birth recode files? I can't merge on hv001 hv002 since this doesn't uniquely identify households (I guess I could do an m:m merge but I'm hesitant to do this).

My whole purpose in doing this is to just get the mother contraception variable onto the birth recode file. In order to simplify things, am I ok in just using the caseid in these cases to merge the two files together?

Also, suppose I want to output descriptive statistics on the characteristics only of the mothers of the children recorded in the birth recode file. (I'm describing the characteristics of a subsample of children in this file, and subsequently am describing the characteristics of their mothers). Am I safe in deleting duplicates in terms of the caseid of the mom to narrow my file down to 1 observation per mother? Here I'm assuming that observations for every "mom" variable in the birth recode (such as education, marital status, etc) are the same given the same caseid. I was going to do a similar thing for households, but given the extra variable problem above I'm unsure whether I just delete duplicates on v001 v002 given the issue with shconces.

Thanks!

Subject: Re: Duplicates in IR file when merging
Posted by Bridgette-DHS on Wed, 13 Jan 2016 13:18:12 GMT
View Forum Message <> Reply to Message

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

If you list out caseid and v003 for the first few cases in the IR file, and list out hhid for the first few cases in the HR file, you will see that caseid is simply a combination of hhid and v003.

Both caseid and hhid are character strings (str15 and str12, respectively); v003 is numeric.

Before you do the merge, when you open the IR file, use this line: gen hhid=substr(caseid,1,12). Then merge with the IR file using hhid. I think this will work for what you want to do, but let me know if it does not.