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 1669 times