Subject: Merging and appending Kenya DHS
Posted by laura on Fri, 04 Aug 2017 17:51:49 GMT
View Forum Message <> Reply to Message

Hello,

I am using Kenya DHS for my analysis. I first want to merge the PR, IR and MR data sets for each year (2003, 2008, 2014). I will then append the merged files for all the 3 years as I want to do a panel data analysis.

I would like some clarification on using weights.

Let's start with when merging the files:
I have gone through posts here that have given me an insight on how to de-normalize the weights (attached document has been particularly useful). However, I would like to know after de-normalizing the weights in the PR, MR and IR and then merging, should I create a variable by adding the de-normalized weights of all the three files? This is because it is tricky to choose which of the 3 de-normalized weights to use once I have merged and I want to declare svy in stata.

Weights when appending files:
When I am appending the files for the 3 DHS (2003, 2008,2014), do I still need to alter the weights?
If yes, how do I go about de-normalizing weights if I want to append?
(I will also appreciate a rationale of why we need to de-normalize weights when appending files if it is feasible to do so)

Lastly,
Which variables can I use to declare xtset once I append the files?
I tried using the region variable and year variable but got an error message.

File Attachments
1) Note+on+de-normalization+of+DHS+standard+weight (1).pdf, downloaded 637 times

Subject: Re: Merging and appending Kenya DHS
Posted by Bridgette-DHS on Mon, 07 Aug 2017 18:06:36 GMT
View Forum Message <> Reply to Message

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

I prefer "renormalize" to "denormalize"....

When using a cross-sectional analysis (e.g. for 2014) you do not need to renormalize, but you do

need to choose between weights. If any variables in a regression or tabulation, etc., come from the IR file I would use v005 rather than hv005. v005 is equal to hv005, except for a slight adjustment for loss of a few women respondents. Similarly, if you are using variables from the MR file, then use mv005 rather than hv005. If you were using the couples file, that includes both v005 and mv005 on the same record, then preference is given to mv005 because male nonresponse is higher than female nonresponse.

I went ahead and did the merge for 2014 with the following Stata lines:

```
use e:\DHS\DHS_data\MR_files\KEMR70FL.dta, clear
keep mv001 mv002 mv003 mv005
rename mv001 hv001
rename mv002 hv002
rename mv003 hvidx
gen sex=1
save e:\DHS\DHS_data\scratch\KE_temp.dta, replace

use e:\DHS\DHS_data\IR_files\KEIR70FL.dta, clear
keep v001 v002 v003 v005
rename v001 hv001
rename v002 hv002
rename v003 hvidx
gen sex=2

append using e:\DHS\DHS_data\scratch\KE_temp.dta
sort hv001 hv002 hvidx
save e:\DHS\DHS_data\scratch\KE_temp.dta, replace

use e:\DHS\DHS_data\PR_files\KEPR70FL.dta, clear
keep hv001 hv002 hvidx hv005
sort hv001 hv002 hvidx
merge hv001 hv002 hvidx using e:\DHS\DHS_data\scratch\KE_temp.dta
tab _merge
keep if _merge==3
drop _merge

summarize *v005
pwcorr *v005
```

Here are the results from the last two lines:

I see that this survey only had a subsample of men. That may be an issue for the kind of analysis you want to do.

I would not say that you have a panel study; you have repeated cross-sections. If you are looking

at changes from one survey to the next, you do not need to alter the weights. Because you have completely different men and women in each cross-section, and you are not (I hope!) trying to combine successive surveys in an additive way, then you do not need to renormalize. You definitely do not need to add up the weights.

If you still have doubts, please let me know (with an example) the sort of thing you plan to do with the combined file....

File Attachments
1) v005.jpg, downloaded 1842 times

Subject: Re: Merging and appending Kenya DHS
Posted by laura on Mon, 07 Aug 2017 19:04:39 GMT
View Forum Message <> Reply to Message

Thank you so much for the response.

I want to analyze how various characteristics (household's, mother's and father's) affect the children's schooling. I wanted to do a panel analysis, but from your response, it seems this is impossible.
"I see that this survey only had a subsample of men. That may be an issue for the kind of analysis you want to do." Do you think this will be an issue in my analysis?

I would appreciate clarification on when to renormalize weights.
I assumed that since I will be merging various files, then I would need to renormalize.

What I have understood is that it is better to use the weight from the file likely to have the highest nonresponse.
I will be using variables from all the 3 files (PR, MR and IR)and below is my svyset code after merging the files:

gen weight = mv005/1000000
svyset hv021 [pweight=weight], strata(hv023) singleunit(scaled)

Ps. I will later on merge the gps file to the already merged file. I hope I do not have to worry about weights when merging the gps file to the merged file.

Subject: Re: Merging and appending Kenya DHS
Posted by Bridgette-DHS on Tue, 07 Nov 2017 12:56:14 GMT
View Forum Message <> Reply to Message

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

I think there are still a couple of questions hanging from your August posts about using the PR, IR, and MR files from the Kenya 2014 DHS. Sorry for the delay. I will try not to repeat what I said earlier. First, on "when to renormalize the weights", this is mainly an issue when pooling surveys from different countries or several surveys from the same country. This amounts to finding some survey-specific number for survey i, call in ki, to re-scale each survey up or down. ki could be the population size (at the time of the survey) Ni divided by the sample size, ni, in which case the weighted number of cases can be interpreted as population estimates. That is, ki=NI/ni. Then, in the pooling, the relative weight of each survey will be proportional to the population size. This sounds good but there is a down side--the pooled results are hardly affected at all by the smaller countries. The alternative is to weight each survey equally. For example if n is the total sample size in a pooling of 20 surveys, and ni is the sample size for survey i, then you ki will be ki=(n/20)/ni = n/(20*ni). This would be my preference. (To be very specific, I am saying that you have a command such as "gen hv005_rev=hv005*ki".)

In Stata, pweights are always rescaled so that they have a mean of 1. Thus [pweight=mv005] will give you exactly the same result as [pweight=weight] where weight=mv005/1000000. Try it both ways and you will see.

Second, I said, "I see that this survey only had a subsample of men". This would be a problem if, say, you merged the IR and MR files with the PR file and then wanted to analyze, say, men and women age 15-49. The PR file includes 32,172 women who are age 15-49 and de facto residents (hv103=1); all of them were eligible for the interview of women (hv117=1). The PR file includes 29,514 men who are age 15-49 and de facto residents. Of them, 13,337 lived in households selected for the male interview, i.e. were eligible for the male interview. If you want to pool the men and women, using variables that are in both the IR and MR file, to get an estimate for men and women combined, you will have to weight up the men, basically with a factor 29514/13337, but actually the factor should be the ratio of the sums of the weights for the 29,514 and the 13,337 cases.

You only need to make these adjustments to the weights if you want to produce pooled estimates. If you just want to compare surveys or compare men and women, it is better to leave the weights alone.