Subject: Weighting in pooled dataset, multiple countries and years
Posted by Marejoha on Mon, 05 Mar 2018 13:21:43 GMT
View Forum Message <> Reply to Message

Dear all,
I am working on a project in stata where I am pooling together data from 20 countries with between 2-5 surveys from each country. I am researching if women in national parliaments have an effect on women's health outcomes in developing countries. I will also combine DHS data with data from IPU and the World Bank and will perform OLS regressions. I am basing much of the data collection on a 2016 paper by Quamruzzaman and Lange.

Now this is quite a big project and I have not dealt with DHS data before. Currently I am a bit stuck on what to do about weighting. I have read several posts on this forum and found different opinions. From what I have gathered I need to first of all divide v005 by 1000000 in all the individual datasets, and then also tell stata to take into account the sampling design by svyset [pw=weight], psu(v021) strata(v022). However I found another post where it said that for regressions you do not actually need to divide the weights by 1000000, but just use pweight in stata (https://userforum.dhsprogram.com/index.php?t=msg&th=50&a mp;). What is the best to do?

Now after this I am a bit confused, I have found one post that said that with pooled data you either want to rescale the weights to add up to a fixed number (to give all surveys the same total weight), or that you can just use the weights that are in the data (which I have assumed to mean not doing anything more with the weights after the svyset command?). That was from this post:
https://userforum.dhsprogram.com/index.php?t=msg&goto=11
306&&srch=weights+multiple+countries+and+years#msg_1 1306

In another post I found the reference to the one pager by Ruilin Ren, explaining how to de-normalise weights for pooled data. Here it says you should use this:
V005*=V005×(total females age 15-49 in the country at the time of the survey)/(number of women age 15-49 interviewed in the survey)
This is done after having divided v005 by 1000000, but it does not mention the svyset command, which I have understood is very important to use if you are to produce standard errors and p values.

I have also read that you should modify the strata and cluster variables to be survey specific, adding multiples of 1000 to the stratum variable in each survey. Would this just be done by gen strata= v022+1000 ?

I am basically wondering which of these advices I should follow (rescaling/weights that are in data/Ruilin Ren example). I am sorry if this is a basic question, I am worried that I may have overlooked something, and just confused myself by reading lots of different suggestions that may not apply to my case.

I apologise for the long post! Hope someone can give me some tips!
Kind regards,
Maren