Subject: Revisiting the topic of weighting data
Posted by acseng on Tue, 21 Jan 2025 09:09:48 GMT
View Forum Message <> Reply to Message

Dear colleague,

I would like your advice and opinion. I accept my question has been discussed many times in this forum, and I have already read most of the replies, including the file pool_and_reweight_surveys_do_22Ot2024.txt written by Tom Pullum.

My point to revisit is on how to append different DHS databases and weights: let's say I want to use data from children under 5 years using KR databases from the same country and from 4 different survey years. My intention is to create a database that after adjustment for potential confounders would represent a hypothetical average population. My questions:
1) in the Stata routine in the txt text, what is the target population to generate the var factor? a)the total population, b) women 15-49 yo, or c) children <5yo?
2)would it make sense to use weightr as wtr=hv005r/1000000/4 to create a kind of "average" population of the 4 years?

Thank you in advance for your help!

Subject: Re: Revisiting the topic of weighting data
Posted by Bridgette-DHS on Tue, 21 Jan 2025 19:14:11 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

I am not enthusiastic about pooling surveys to get some kind of average over a long period of time. However, if you do that, a major issue is that the different surveys will have different sample sizes, and if you don't adjust for that, your results will be most influenced by the largest survey and therefore the conditions at the time of that survey.

Say that n1, n2, n3, n4 are the four sample sizes and the total is N. Say that the weight variables in the samples are w1, w2, w3, w4. You can construct w1'=w1*N/(4n1), w2'=w2*N/(4n2), etc. Then, the sum of the weights should be the same in each survey.

You can define the population of interest however you want but it sounds like you want the children under 5 to be the cases, and you would pool the KR files.

Most analysis uses pweights, and they are the weights in svyset. Pweights are automatically normalized in Stata to have a mean of 1 in the separate files and in the pooled file, so it doesn't really matter if you have a factor of 1000000 or something else.