Subject: Re: Weighting when combining datasets - AIS and non-weighted datasets
Posted by Bridgette-DHS on Thu, 04 Feb 2016 14:40:45 GMT
View Forum Message <> Reply to Message

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

As I interpret the situation, you want to append another file to a DHS file and analyze them as one file. I strongly recommend that you not do this. It will be much better to analyze the two files separately and then present the results side by side.

The weights, clusters, and strata are determined by the design of the data collection. If you do not have information about the design of the non-DHS survey, then you cannot safely make any assumptions about those things. It is very unlikely that, as you say, the second sample does not need weights. It's much more likely that the information required for calculating weights was not saved. Almost no samples are genuinely self-weighting.

The weights in a DHS survey reflect relative under-sampling or over-sampling (and response rates) of the clusters. They ignore a factor which is basically the ratio of the national household population to the sampled household population, i.e. the number of cases in the PR file. If you want a pooled DHS sample and non-DHS sample to be nationally representative, you must adjust for the difference in the overall sampling fractions. If n1 is the size of the DHS sample and n2 is the size of the non-DHS sample, then as a minimum you would assign a pseudo-weight n1/n2 to the cases in the non-DHS sample. This would weight the non-DHS sample up if it is smaller than the DHS sample, and down if it is larger than the DHS sample. If you are working with Stata, pweights are always adjusted so that the total weight in the full sample is equal to the unweighted total number of cases. Therefore you can do this kind of adjustment to either sample, without affecting the robust standard errors, etc.

To repeat, I would not go down that path. It would be much better to do parallel analyses of the two data sets and then compare the results. I don't think any journal reviewers, for example, would accept an analysis that was based on the kind of pooling that you suggest.