Subject: Question re-weighting combined survey data
Posted by DHS user on Wed, 20 Feb 2013 17:15:27 GMT
View Forum Message <> Reply to Message

Is there any recommended strategy to compute a new weighting variable when combining data from multiple countries to identify pooled estimates? If no, which weight variable should I use if I want to combine data from multiple countries?

Subject: Re: Question re-weighting combined survey data
Posted by Bridgette-DHS on Wed, 20 Feb 2013 17:17:48 GMT
View Forum Message <> Reply to Message

Here is a response from one of our DHS Sampling experts Ruilin Ren, that should answer your question.

Attached is a one page note on weight variable rescaling when pooling data from different surveys. The DHS recode data file presents a normalized weight which is a relative weight. The attached note explains how to de-normalize weights for pooled data or for estimating totals.

I hope this helps.

Bridgette-DHS

File Attachments
1) Note on de-normalization of DHS standard weight.pdf, downloaded 4231 times

Subject: Re: Question re-weighting combined survey data
Posted by Reduced-For(u)m on Sat, 30 Mar 2013 04:34:28 GMT
View Forum Message <> Reply to Message

Bridgette,

Wow. This is actually kind of scary to me. From that one-pager:

For example, to de-normalize the household standard weight HV005, one should divide the household standard weight by the household survey sampling fraction, that is, the ratio of total number of households interviewed in the survey over the total number of residential households in the country at the time of the survey....The second piece of information is usually obtained from population projections for a period close to the time of the survey fieldwork, based on the latest population census. The de- normalized weight is very sensitive to the second piece of information, so one should guarantee that the source of information is reliable; otherwise, it can lead to erroneous statistical conclusions.

That is actually sort of scary, right? Has anyone ever tried to figure out whether anything at all is gained from trying to re-normalize these? Supposing we appended 10 country survey rounds together and we didn't do any weighting at all...then we would have a mean that was biased toward the over-sampled populations and ignored the population-size differences between the countries, but which is interpretable as the mean of that particular sample. But if we grab some estimate of the total number of households (or people) in a country and that estimate is, say, 5-10% off (which strikes me as a conservative guess at how well we know how many households are in a country), then are we really getting anything approaching the population mean, or are we possibly getting a worse estimate that is totally uninterpretable?

I've generally sided with weighting over not-weighting, but I might be tempted to re-think that in situations where we are using pooled data. Alternatively, anyone have an interpretation of what we are estimating if we just use the regular weights after appending 10 countries together? Is that like weighting within country but ignoring population differences across countries (the implicit population weight being the sample size)?

One last thing: If we append together multiple rounds of the same survey, do we still need to re-normalize, and what are we estimating if we don't relative to if we do - meaning what are we implicitly assuming about the sample sizes and population growth over time? Sorry if this is asking too much, but if anyone has any insight on this, I'd love to hear it. Weighting in these ways is kind of hard to think about.

Subject: Re: Question re-weighting combined survey data
Posted by bsayer on Wed, 03 Jul 2013 22:17:50 GMT
View Forum Message <> Reply to Message

I am doing some somewhat similar things, so I'm going to take a stab at this. However I am not sure I am understanding exactly what is meant by "de-normalization". I take it to mean reversing the process of normalizing the weights to the sample size. So my comments are in that line.

I understand the trepidation of DHS-User. With regard to the statement about needing the second piece of information, I think that may vary by survey. For example, I have been working with Uganda AIS surveys (the standard DHS uses the same sampling scheme) and the documentation indicates that the number of households in the PSU (enumeration area) is found by a canvas of the PSU before the interviews start. So only the housing count of the PSU is needed, not the population - at least in that survey. That is, if one one feels that de-normalization is necessary. It would be nice if DHS posted the spreadsheet of counts for each stage of weighting.

Now on to the issue of "is it necessary". I don't usually work with normalized weights - DHS is the only data that I use that has them. I am also assuming that survey design software is being used to analyze the data. So I'm always wondering what the impact of normalizing is. For one activity I was doing (creating weights for the couple file) I did a test to make sure that we got the same results using the normalized household weight as we do with an original, not normalized weight. We know that E[c*X] equals c*E[X] so I expected that it did not matter, and it doesn't. But this is not a case of combining surveys, this is just one survey.

What is the purpose of combining surveys and what is the resulting estimate? That is probably the

more important question. We regularly combine multiple years of U.S. surveys (NHIS, NHANES, MEPS, etc) and we do not de-normalize the weights. In fact, the weights are much more complex than the DHS weights. What we do is adjust the weight to reflect the number of surveys we are combining. This usually takes the form of dividing the weight by the number of surveys - nothing more than that. The reason we do this is for totals, not means. The result is that we are creating an estimate for the mid-point (in time) of the survey. Of course, these surveys are designed for this, and that may have an impact.

So what is the purpose of combining DHS surveys? Is it to get sufficient sample to evaluate a small population? Is it to compare countries? I can't tell this from the question posted, and the answer may impact what should be done with the weights. What I will point out is that the stratum variable needs to be modified to be survey specific. It is important that the strata and PSU information cannot be combined across surveys, unless that is planned for in the survey design, and to my knowledge none of the DHS surveys do this. This is easy to do, just add some multiple of 1000 to the stratum variable for each survey, i.e. 1000 for country A, 2000 for country B, 3000 for country C, etc. (Make sure the order of magnitude of the additive is larger than the order of magnitude of the stratum variable in every survey).

Back to the question of the weights. If the only reason for combining surveys is to make it easier to estimate survey specific results, for sure the weight is fine. Just fix the stratum variable.

If the purpose is some form of pooled estimate then it may also depend on the time of the survey. For example, pooling multiple countries that are all surveyed at approximately the same time then I don't see a reason to alter the weight variable. Now if multiple surveys from different time points for the same country are being combined, this might be more complicated. But even here I am not certain it is necessary to de-normalize. I think this specific case may require more investigation.

One final thought. Perhaps de-normalizing is necessary in the context of using software that does not have survey capabilities. In which case the strata and clustering is being ignored anyway, so we know the variances are not design correct. The solution here is simple - use the correct software. But I generally don't see a reason why the scale of the weights matters when using the correct software.

So, my take is this: 1) Use survey design capable software; 2) Modify the stratum variable to be survey specific; and 3) altering the weights is not necessary, but it is necessary to use weights.

Subject: Re: Question re-weighting combined survey data
Posted by Reduced-For(u)m on Thu, 12 Sep 2013 00:27:19 GMT
View Forum Message <> Reply to Message

bsayer,

This was a great post that summarizes really well a lot what has been talked about on these weighting threads. In particular, I think the whole "modify the stratum variable to be survey specific" is something we should put in a sticky thread at the top of the weighting thread, along with the code for svyset (and its SAS/SPSS equivalent code) and a list of which weights to use for

which dataset/recode.

Now just between us - since we've gone a little back-forth on this in the past - let me give one instance in which I think you might want to re-scale the weights in some way.

Suppose you want to know the correlation between sanitation access and child mortality in all of sub-Saharan Africa. So you pool together the last DHS from each of a bunch of countries into one dataset, and regress mortality on sanitation access. Now, you could do this separately by each country, but maybe you just want to know the average impact across the region.

If you use the original DHS weights, which, within country sum to the sample size, are you not implicitly weighting the pooled regression by the sample sizes in each country (so each country's total regression weight comes out to (Nsurvey/Nallcountries)- the fraction of the total observations that came from that country)? Wouldn't a better weighting scheme be to have each country's weights sum up to that countries population? Then, when STATA re-normalizes all the weights (summed for all observations in all countries) to 1, people in small countries will have (by design) less weight than people from large countries. And isn't that what we'd want? Wouldn't we want the larger countries to get more weight in this case - assuming we are looking for some "population average" in our regression coefficient?

I know we've been over this a bit before, and I have conceded that I might be missing something, but I still think that in some cases you would want to re-scale the DHS weights so that they could simultaneously act as population weights. Or, if you want each country to have equal weight, then re-normalize so that each country's weights sum to 1/Ncountries.

If anyone thinks it is worth it, I'll try to do some regressions of this sort weighted in various ways and post the results here, but I can't do it at the moment and wanted to respond before I forgot to.