Subject: How would I use weights in this scenario?
Posted by RASimmons on Thu, 01 Dec 2016 15:17:40 GMT
View Forum Message <> Reply to Message

I apologize for the vague thread title, but the situation I am trying to use the DHS data in is rather complex, and I couldn't think of a quick way to summarize it in the title.

So, the basic idea is this: I am interested in looking at predictors of infant mortality across multiple countries in sub-Saharan Africa, AND how these predictors have changed over time.

Currently, I am using data from 8 countries, with multiple DHS surveys per country (using the birth recode data). Kenya (1998, 2003, 2008, 2014), Malawi (2010, 2013), Namibia (2006, 2013), Ghana (1998, 2003, 2008, 2013), Rwanda (2000, 2005, 2007, 2010), Senegal (2010, 2012), Tanzania (2004, 2010), and Uganda (2006, 2011).

However, one of the main research questions is the impact of access/quality of healthcare on infant mortality. To that end, the data from the DHS surveys has been linked at the country/region level to data from the SPA surveys. The way the matching operates is that any birth a given DHS survey round is linked to the closest SPA survey available for that country/region by date; e.g. if the birth was recorded as being in Kenya in 2011, it gets linked to data from the Kenya 2010 SPA, while if the birth was in 2006 it gets linked to the 2004 Kenya SPA, etc.). We then subset the data so that the maximum distance in time between any given birth and the nearest available SPA is 4 years (we then apply a couple of extra dataset restrictions; e.g. excluding records where the maternal age at birth is recorded as being less than 15, etc.). So, this gives us a smaller available subset of these DHS surveys (some DHS surveys are eliminated completely, because there isn't an SPA available within the given time frame, while others are only partially excluded based on the child's birth date). To be clear, this means that all of the births within the same region-country-year that are successfully matched to an SPA all have the same values for whatever those SPA variables are (e.g. number of facilities per 1000 population in that region, etc.).

Now, how would I use weights in this scenario? I've read a bunch of the threads on here and understand the basic concepts of re-normalizing, etc. However, I also understand that the fact that I am using not only multiple countries, but multiple rounds PER country makes the question of how to re-normalize a bit more complicated, and it gets even MORE complicated since I am dealing with a subset of the data due to our various dataset restrictions. Even ignoring the fact that the inclusion of the SPA data (which is another complexity), it isn't clear to me precisely how I would go about weighting this set of DHS data. Further, I think it is well understood in the survey analysis community that using the WRONG weights can actually induce more bias than not weighting at all ... so, is this a situation where on a philosophical level I might be more justified in just not weighting and accepting the fact that the oversampled surveys will contribute more to the results of the model than the undersampled surveys (so, in this case, the Kenya 2014 DHS will drive the parameter estimates)? Or do you think it is possible to construct a sensible weighting schema given the complexities of the dataset?

Very curious to hear your thoughts!

Subject: Re: How would I use weights in this scenario?
Posted by Bridgette-DHS on Thu, 01 Dec 2016 16:56:07 GMT
View Forum Message <> Reply to Message

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

Quote:It's very important that you use the weights. Otherwise the within-survey distributions will over-represent the strata that were over-sampled, etc. The only question is whether, in the pooled file, you want to re-scale the weights. The alternatives are re-scale them to add up to a fixed number in each survey or to add up to a number that is proportional to the estimated population sizes at the times of the surveys. Between the two, I personally prefer giving every survey the same total weight. In a sense, THIS is the option that corresponds to not using weights at that level (but using them within the surveys). If you do not adjust for the sample size, then you are allowing a completely arbitrary characteristic of the survey to influence the importance of that survey. If you make the total weight for a survey proportional to the population size, then large countries will completely swamp small countries. Within the series of surveys, in many countries, the recent surveys are much larger than the early surveys.

But for many purposes you can just use the weights that are in the data. The issue of how to adjust the total weight for each survey is really only relevant if you plan to produce pooled estimates, such as "the mean contraceptive prevalence rate (CPR) in Ghana from 1990 to 2010" or "the mean CPR in West Africa in 2000". I don't think those are meaningful parameters to try estimate, and I don't think DHS surveys cover enough years and countries. If you are interested in estimating changes and differences, which I think is more appropriate, then you can just leave the weights as they are.

Subject: Re: How would I use weights in this scenario?
Posted by RASimmons on Mon, 05 Dec 2016 14:51:51 GMT
View Forum Message <> Reply to Message

Hi Bridgette (and Tom),

Thanks for the reply!

What about the issue of linking the data from the SPA surveys to the DHS surveys? Should the SPA data be weighted differently, to take into the account that the data was sampled differently? Or for the purposes of the analysis I described, is it sufficient to make the DHS surveys representative in the way you describe?

Subject: Re: How would I use weights in this scenario?
Posted by Bridgette-DHS on Mon, 05 Dec 2016 18:36:16 GMT
View Forum Message <> Reply to Message

Following is a response from DHS Senior Research Associate, Wenjuan Wang:

Quote:The SPA data should be also weighted using the weight variables included in SPA data files. There are facility weight, provider weight, and client weight, depending on your unit of analysis. For example, if you link DHS data to the regional level SPA indicator: proportion of health facilities that provide Cesarean delivery services, you would use the facility weight.

Hope this helps!

Subject: Re: How would I use weights in this scenario?
Posted by RASimmons on Tue, 06 Dec 2016 15:11:08 GMT
View Forum Message <> Reply to Message

Thanks for the reply!

Do I need to follow the same sort of process for rescaling the weights as I do for the DHS weights? I would assume I do.

As for actual implementation of the weights, to my knowledge standard statistical software won't let me use two different sets of weights in one model (usually you just specify the weight variable in the PWEIGHT command, for example), so how would I actually perform this type of weighting? Can software apply different weights to different sets of variables (e.g. the DHS weights to the variables from the DHS, the SPA weights to the variables from the SPA)? If not, then how would I actually do this? Would I have to apply the weights manually to the data (so multiply each observation by the reciprocal of the weight value), and then use manually adjust the standard errors (or use robust standard errors, or bootstrap, etc.)?

Subject: Re: How would I use weights in this scenario?
Posted by Bridgette-DHS on Tue, 13 Dec 2016 20:29:58 GMT
View Forum Message <> Reply to Message

Following is a response from DHS Senior Research Associate, Wenjuan Wang:

Quote:From your previous description, I understood that you want to examine how access/quality of healthcare is associated with infant mortality and you plan to link DHS and SPA data at the regional level, right? If this is the case, I see the following major steps of the analysis and the weight variable you would use in each step.

You first need to generate health service variables of interest at the regional level based on SPA datasets. For example, you want to look at the availability of C-section services in the region, you may generate this indicator: proportion of health facilities in the region that offer C-section. When generating this indicator, you will use SPA facility weight. Sometimes you also want to use provider data, for example, % of facilities with at least half of the providers who provide delivery services trained in C-section. You will need to use both provider weight and facility weight to

create this indicator.

After you prepared these SPA indicators, you then merge them into the DHS dataset, for example, BR file using the region variable. You do need to make sure the region categorization is the same between DHS and SPA for the same country. They are not always the same.

After merging, let's say you use survival analysis to test the association between healthcare access variables and risk of death before one. You will apply the DHS weight.

Hope this helps!
