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!