Subject: Re: Weighting district-level data
Posted by Reduced-For(u)m on Thu, 17 Dec 2015 22:32:26 GMT
View Forum Message <> Reply to Message

This is (again) a question that has no definite answer, but there are better and worse ones.

First off, since you are using these types of estimators, should I assume that you are using 2 or 3 rounds of the DHS? In that case, you have to be fairly careful when you calculate your district-level means (which become the observations in your estimates). There are also some big issues with calculating standard errors (p-values) which relates to "clustering". Here is how I would do it, and then an alternative:

Weighting: you should calculate the survey-specific district means using the survey given weights. If you do this separately by survey round, you won't have to worry at all about re-noramlizing the weights because you'll be calculating a representative mean of the district level variables (what become your observations). Also note, if you wanted to make the final regression "population representative" you could weight each district-year-level observation by the population of the district - so larger districts would get more weight than smaller ones. This may or may not be reasonable, and is up to you and your specific interests/needs.

Clustering: clustering at PSU is not sufficient in this case, at least not usually. The rule of thumb here would be to cluster at the "district" level - the level at which you collapse observations/assign "policy intervention". The usual reference is Bertrand, Duflo and Mullainathan "How Much Should we Trust Difference-in-Difference Estimates". To do this in Stata, when you define PSU in your "svyset" command, you use the district identifier (that is common across survey rounds). You can include the strata here too, but I don't think it will make much of a difference (and if it does, it should make your p-values slightly smaller).

Note: If you have fewer than 30 or 40 districts, you should also see Cameron, Gelbach and Miller "Bootstrap Based Improvements for Inference with Few Clusters" - there is a new Stata package that makes doing those "wild-t bootstraps" very easy: see "cgmreg" group of .ado files you can download.

Now - there are also a couple more ways to do this. In particular, you could do this same analysis on individual-level data with group-level covariates. Everything is the same, but now instead of weighting when you collapse, you'd have to weight in the regression. In this case, you could either follow Gary Solon in "What are We Weighting For" and argue that, with causal effects, you don't need to weight, or you can follow standard DHS recommendation and re-normalize your survey weights and then apply those in the regression analysis.

Hope some of this helps. I can follow up with details on one of these methods if you decide you like one and still aren't sure what to do.