Subject: Re: Using weights in regression analysis
Posted by Reduced-For(u)m on Fri, 14 Jun 2013 21:24:32 GMT
View Forum Message <> Reply to Message

Hey there. These are all really good questions. I'll go through them best I can. But first off, just to be clear, I'm not a DHS employee, and have no special insight other than what I've gleaned in my working with the DHS data, discussions with other users, and my general econometrics training. So nothing I say should be taken as the voice of the DHS speaking, or even the advice of some super-expert, just another practitioner trying to figure these things out. With that out of the way...

Weighting, clustering and stratification for regression

(1)...just about everything you say here is new and interesting to me. I had never used the "svy" command before using the DHS - I always weighted and specified standard error calculations manually. What I know about using "svy" for the DHS mostly comes from the DHS FAQ and this paper: http://eprints.soton.ac.uk/8142/

Basically, I have no real insight on the proper use of "svy" to deal with complicated survey designs.

(2) This is one of those times I wish I had said less (it happens), just because leaving something fuzzy like that is probably not helpful. First things first, when it comes to standard errors/inference, we aren't really talking about weighting, we are talking about stratification and clustering. The weighting problem is really just a question of what population you want your results to be representative of (the survey population, or the national population, or the regional population or whatever). Weighting doesn't require the use of "svy".

As for statistical inference (standard errors), to me, one way of thinking of the DHS standard error assumptions is that IF the DHS had used a simple random sampling, then we could just use OLS standard errors (and weight manually with [pweight=weight]). However, in many applications, like difference-in-difference estimation or a cohort fixed-effects-type regression, even with simple random sampling, this is probably an un-conservative technique. Coming from the Labor econ world, I think two really good introductions to the problem are "How much should we trust difference-in-difference estimations" (Bertrand, Duflo, Mullainathan) and "Robust inference with clustered data" (Cameron and Miller). These papers focus on situations where there is likely to be auto-correlation and/or heteroskedasticity in error terms within "clusters" like states or counties (not to be confused with sampling clusters, but some larger grouping of people). As another example, and closer to home, when estimating cohort determinants of HAZ (like, say, effect of month of birth, or the effect of some shock in the birth cohort) I find that the "svy" technique leads to rejection rates on a placebo treatment over 25% (when it should be 5%) and up to like 70% in some cases.

One important caveat though is that all of the things I mention above are uses of the DHS for which it was probably not originally designed. I think that when it comes to things like the effect of maternal age on HAZ, then the DHS method will produce better sized standard errors. I haven't run any placebo tests on that to check implied rejection rates, but you could probably do it fairly easily.

Even though I am using the sample weight, my tabulations differ from those in the country tables

Hopefully some "-DHS" will respond to this, as I have only two (maybe, maybe not) helpful comments and one useless one.

1 - I use "pweight" instead of "iweight". My guess is that it will not make a difference, but these are probability weights (best as I can tell) and since using pweight automatically scales everything to sum to 1, it might make a difference.

2 - Still on weights...when appending multiple rounds, my understanding is that this induces a new weighting problem, as DHS weights within a survey sum to the sample size. So, by just using the given weights, you are not weighting each survey the same, you are implicitly weighting it by the sample size. I'm not sure if that is what you want or not. An alternative would be to re-scale each survey's total weight to sum to one manually, preserving probability of sampling within survey but making each survey have the same total weight (assuming population size is constant, each survey is actually "representing" the same number of women).

```
egen surveytotalweight = total(weight), by(survey)
gen new_weight = weight/surveytotalweight
```

I so far have done that AFTER I dropped all observations that wouldn't go in the regression I use or the statistics I'm tabulating.

3 - I know nothing about replicating DHS tables, so I'll go back to hoping someone "-DHS" responds.

****

I hope this has been in some way helpful. I've been struggling with the weighting thing myself, and how to deal with multiple survey rounds. Truth is, I don't think there are really "perfect" answers out there, and a lot of us are trying to figure things out on our own and doing things in different ways depending on our backgrounds. So my perspective is one that comes from dealing with problems in the Labor Econ world, and Epidemiologists or Nutritionists would have different opinions and different modelling concerns. For example, my field has basically stopped using any random effects models and switched to using an "arbitrary" or "cluster-robust" variance/covariance matrix estimation - I haven't been able to confirm, but I think that the "svy" command uses some weird random-effects-type specification of the V/C matrix. So my biases and "insights" (such as they are) come from that world, and may not be totally appropriate here. These are just my thoughts. I'd love to learn more if someone thinks I'm missing something obvious or important or just fundamentally not understanding something.