Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by user_rm on Tue, 08 Sep 2015 13:07:05 GMT
View Forum Message <> Reply to Message

Thanks for your email. I suppose my confusion just boils down to 1 point - if each child in 1 cluster has a weight of "1.148330" as you explain, then when I am collapsing observations by district (such that I have 1 averaged datapoint for every district), do I apply the 1.148330 weight to the whole district?

e.g if there are 5 kids in 1 district, each with a weight of 1.148330. When I collapse the HAZ score, I get 1 average value for the district. Now when running regressions on average data, do I apply a weight for 1.148330 to that district or would the weight by 1.148330*5?

A.
The confusion arises in this context because I am trying to calculate the proportion of villages with Anganwadi centres in each district, weighted by the number of kids treated in that district. I am confused as to which command I should use in step 2:
1. gen weight = v005/1000000
2. collapse(mean) weight, by(District Village kids) or collapse(rawsum) weight, by(District Village kids)?
3. collapse(mean) kids[pweight=weight], by(District)

where kids treated = 1 if the village in which the kid lives has Anganwadi centre.
If I use the collapse (mean) command, then within a district, an average kid in 1 village and the treated/untreated kid get the same weight

Village kids District weight
25 0 AHMADNAGAR 2.060619
37 1 AHMADNAGAR 2.060619
40 1 AHMADNAGAR 2.060619
53 1 AHMADNAGAR 2.060619
56 1 AHMADNAGAR 2.060619
132 1 AHMADNAGAR 2.060619

If I use the collapse(rawsum) command, then the weight is different according to the number of kids treated in each village, which is kind of what I would like.

Village kids District weight
25 0 AHMADNAGAR 30.90929
37 1 AHMADNAGAR 26.78805
40 1 AHMADNAGAR 26.78805
53 1 AHMADNAGAR 35.03052
56 1 AHMADNAGAR 18.54557
132 1 AHMADNAGAR 10.3031

It's tricky because for other variables, I think a normal mean collapse command would work
e.g collapse(mean) HAZ WAZ [pweight=weight], by(District)

B. I then use the weighted treated proportion variable and HAZ WAZ, etc in a regression. Now then again, I would be using the survey weights? Is that right? I am very confused about this.
e.g gen surveyweight = v005/1000000
collapse(mean) surveyweight, by(District)
reg HAZ WeightedProp MothEducYrs, [aweight= surveyweight]

C. As a separate point, I wanted to include some summary statistics (ie mean and sd) tables for individual children related data (not averaged). I used the estpost tabstat command. But they don't let me use 'iweight' or 'pweight' I had to use 'aweight' Do you think that is okay?

Please clarify.

Many thanks
R