Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by user_rm on Tue, 18 Aug 2015 17:06:24 GMT
View Forum Message <> Reply to Message

Hi,

I am running regressions of a pooled cross section data for 2 years of NFHS (1992 and 1998). The data is

where i represents number of kids, j number of districts and t = 1992, 1998
Yit is the outcome variable e.g standard deviation of height for age for each kid (also called height for age z score) for 2 time periods
prop is the proportion of villages treated in each district j. I define treatment by the presence of an Anganwadi centre in village.
So in 1 district the proportion could be 1 (if all villages had the Anganwadi centre), 0 (if none had) or e.g 0.83 (if 5 out of 6 villages had the centre)

μj are the district dummies
controls would be mother's education, family size, etc
eit are the cluster standard errors

I want to run another version of this equation where I have average data by each district (e.g average height for weight z score). I am not sure how to do this in stata. I used the collapse command but I am having troubling in getting the specification of this command right. Can you please guide me in this regard?

1. I am confused as to how to get the proportion of villages treated in each district weighted by the number of kids treated in that district.
2. Will it be okay to weight the controls like wealth index, 0/1 variables in the same way?
3. The controls and the dependent variable will be weighted by the number of kids treated and non treated in the district? How can I do that in stata?
4. I should use cluster SE. So should the regression command have cluster(District) or vce(cluster District)? what is the difference between the two?

The collapse command I was trying to work is below:
gen ones=1
collapse(rawsum) ones(mean) propor [fw=ones], by (District treated)
where treated would be number of kids treated?

Please help in this regard.
R