Subject: Weighting of pooled country and year data
Posted by Christiaan on Fri, 17 Jan 2025 15:15:02 GMT
View Forum Message <> Reply to Message

Dear DHS community

I have been performing analyses using DHS data pooled over various countries and years. I would like to weight this data accordingly. I have been reading up on how this should be done and I came across the following thread:

https://userforum.dhsprogram.com/index.php?t=msg&goto=99 82&S=Google

What interests me in particular is the first response by DHS Stata Specialist Shireen Assaf:

gen wt= v005/1000000

egen strata=group(v000 v025 ADM1_CODE) // strata also includes the survey (identified by v000) in the group command

egen v001r = group(v000 v001) // cluster also includes the survey in the group command

svyset v001r [pw=wt], strata(strata) singleunit(centered)
svy: tab ADM1_CODE

Now, I would like to display mean values of certain outcome variables by birth cohort. How would I weight this appropriately, drawing on the code provided above?

gen wt= v005/1000000

egen strata=group(v000 country v007 birth_cohort) // Where v007 is the year in which the survey was conducted and birth_cohort is the cohort in which the observation was born.

egen v001r = group(v000 country)

svyset v001r [pw=wt], strata(strata) singleunit(centered)
svy: sum outcome_var, by(birth_cohort)

Would this approach to weighting make sense? If not, how can I improve upon it? This is the first time that I am using weights and I would just like to make sure that I do it correctly.

Thank you so much for your kind assistance!

Subject: Re: Weighting of pooled country and year data
Posted by Bridgette-DHS on Tue, 21 Jan 2025 17:55:47 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

Unfortunately, Shireen no longer works at DHS.

Your question seems to be more about the specification of the svyset command than about weighting. There have been many related posts. The svyset command is determined by the sampling design. If you pool samples, you need to take into account that the clusters and strata are different in different surveys, even in the same command.

Construct a unique identifier for each survey, e.g. survey=1, 2, 3, etc. v000 does not actually serve this purpose because there can be two or even more successive surveys in the same country with the same value of v000. For each survey, define (say) cluster=v000 and stratum=v023 (or perhaps something else for older surveys). Then append the surveys and construct unique codes for cluster and stratum with

```
egen cluster_id=group(survey cluster)
egen stratum_id=group(survey stratum)
```

Then your overall svyset command will include cluster_id and stratum_id. There have been countless postings on the construction of weights in pooled files and I will not repeat any of them.

The birth year of respondents in the IR file is v010, and that is what I use to define birth cohort. I do not understand "birth_cohort is the cohort in which the observation was born." There must be a typo somewhere.

You do not need to adjust svyset for any covariates, birth cohort or anything else, because they are not part of the sampling process or design. You can analyze the data by age or cohort or year of data collection or region, etc., without changing svyset.