Subject: Weighting variables in DHS India data (1992 and 1998)
Posted by user_rm on Thu, 06 Aug 2015 17:20:51 GMT
View Forum Message <> Reply to Message

I am using the 1992 and 1998 DHS datasets for India. I am unable to get my head around how to re-adjust weights in both datasets due to the following reasons.

For my analysis, I have to make both datasets consistent with each other because I want to use them in the same regression. Therefore, I had to drop certain observations specified below:

1) The data for kids in 1998 dataset is for kids under the age of 35 months. So I had to drop data for kids greater than 35 months in the 1992 file.
2) Merging village files with the household and kids files in both datasets results in unmatched observations for urban towns. So I had to drop these and keep observations for only rural areas.
3)There are around 370 districts in both datasets. However, the 1998 dataset increased the coverage to include additional 50 districts. But I had to drop these observations to make the 2 data-sets consistent.

Can you please suggest how I should re-adjust weights in both the datasets? I am working in stata.

Many thanks
R

Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by Bridgette-DHS on Fri, 07 Aug 2015 13:33:44 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Stata Specialist, Tom Pullum:

I do not recommend that you adjust the weights at all. According to your first and second points, you are just studying children under 3 years of age (that is, age 0-35 months) in rural areas. So long as you say that, explicitly, in your analysis, the weights would not be affected. Your third point, about dropping the 50 extra clusters, is more of an issue. It is possible that when they are dropped, the weights should be re-calculated. However, this would require information about the sample design that is no longer available. You should mention explicitly that those 50 districts have been dropped from the 1998 data in order to have better comparability with the 1992 data. That's a good decision.

It's important for other people to be able to replicate whatever you do, and by being explicit about 1) - 3) you will make that possible. If you start tinkering with the weights, then people will NOT be able to match what you have done, even though I'm sure the differences would be very small, probably negligible.

Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by user_rm on Fri, 07 Aug 2015 15:00:43 GMT
View Forum Message <> Reply to Message

Many thanks for your reply. I have 2 follow up questions:

1. I am not clear on why the weights wouldn't change if I use data for kids below the age of 35 months? I say this because the 1992 dataset has kids data upto 4 years while 1998 dataset has kids data upto 3 years. To both datasets consistent, I had to drop the kids of age 4 years from the 1992 dataset. Wouldn't that mean that the weights, in principle, change?

2. Even if I don't re-adjust weights, I will still have to weight the data using the weights provided in the data? How do I do that?
Or is the NFHS data already weighted?

Please clarify.

Regards,
R

Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by user_rm on Wed, 12 Aug 2015 23:01:08 GMT
View Forum Message <> Reply to Message

In the 2 datasets (1992-93 and 1998-99), I wanted to weight the data to make it nationally representative. I am using kids data (as described earlier) and I have also merged the village files to get information on the Anganwadi, schools, etc.
I am unsure as to which weights I should use to weight the data in both years?

There is a variables V005, which is the relevant sample weight to be used here I think?
But do I also have to use the village weights for all the village variables?
I am unclear as to which set of variables I have to apply the sample weights and village weights?

Can you please clarify?

Many thanks
R

Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by Bridgette-DHS on Thu, 13 Aug 2015 15:26:47 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Stata Specialist, Tom Pullum:

The weights are used to inflate subpopulations that have been under-sampled or to deflate subpopulations that have been over-sampled. If you have dropped a subpopulation entirely, such

as kids in some age range, then that age group is gone. You cannot replace them by weighting up some other subpopulation, unless the subpopulation that remains is the same, in every way, as the subpopulation that you dropped. In your example, you cannot assume that.

The representativeness of the remaining sample is unaffected by dropping some other part of the sample, so long as you make it clear what population the remaining sample comes from. You do not need to change the weights for the cases that remain to be representative.

The weight variable is v005, as you said. If individual cases are your units of analysis, then you do not need any other weight. These weights apply to the case, and are not different for different variables, such as variables that describe the child, the mother, the household, the village, etc. To repeat, the weight is specific to the case and does not depend on what variables you are using.

I suggest that you try to match some basic numbers in the relevant survey reports. If you are using the weights in the same way that DHS uses them, then you should get a match in the weighted number of cases and the weighted means, proportions, etc.

Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by user_rm on Fri, 14 Aug 2015 18:07:20 GMT
View Forum Message <> Reply to Message

Many thanks for your reply.

I am using the individual level data (ie for kids) for 1992 and 1998 for rural households. In stata, I have the 1998 data appended below the 1992 data for the same set of variables (ie it's a pooled cross-section analysis).

I wanted to specify the survey design in stata to make the data nationally representative. The NFHS manual states that the survey is a two stage stratified sample design for rural areas. The first level of stratification consists of identifying districts within each state. Districts were further sub-divided into regions and households for villages were identified for each region. In terms of specifying the sampling design in stata, do I need to specify the cluster number (v001) and household number (v002)? The weights used would be sWeight = v005/100000. Is the code below correct for this purpose? Or am I missing something here?

svyset v001 [aweight=sWeight]|| v002 [aweight=sWeight]

Please clarify.

Many thanks
R

Subject: Re: Weighting variables in DHS India data (1992 and 1998)

Posted by user_rm on Sun, 16 Aug 2015 16:14:15 GMT
View Forum Message <> Reply to Message

Is it necessary to use the svyset command for weighting?

Can one not just use pweights = v005 in the regression equation itself without using the svyset command?
e.g regress y x [pweight=v005] ?

Do you have any thoughts in this regard?

Many thanks
R

Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by Reduced-For(u)m on Sun, 16 Aug 2015 21:11:25 GMT
View Forum Message <> Reply to Message

The problem with doing the weighting manually (instead of via the "svy" prefix) is not in the weighting, but in the stratification and clustering. That is, your command will weight the observations just fine, but won't account for other aspects of the survey design.

The clustering at PSU could be done manually ", cluster(psu)", but I am not sure there is a good, easy way to manually account for the stratification (in the sampling procedure).

Your version (just the manually used weights) will produce standard errors that are too small (your p-values will be too small, and you'll reject a true null hypothesis too often). Adding in clustering at the PSU level will produce almost right standard errors (or p-values, or CIs) but they may be a little too big since they don't account for the stratification. Using the "svy" prefix guarantees not just that you get good population level estimates of means/differences (weighting) but also more appropriate statistical inference properties (rejection rates, CI coverage rates, however you want to think of that).

Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by user_rm on Sun, 16 Aug 2015 21:17:37 GMT
View Forum Message <> Reply to Message

Thanks for the reply. Can you please suggest how should the stratification and survey design be set properly in stata so that I can use the svy command when running regressions?

Do I have to specify the survey design and stratification? Would the following command be okay?

svyset v001 [pweight=v005]|| v002 [pweight=v005]

Please let me know.

Thanks
R

Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by Reduced-For(u)m on Mon, 17 Aug 2015 00:59:24 GMT
View Forum Message <> Reply to Message

See this thread:

http://userforum.dhsprogram.com/index.php?t=tree&goto=240&S=501bdf45b666e5230f2eab8ece102b2d#msg_240

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

Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by Bridgette-DHS on Wed, 26 Aug 2015 11:37:54 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Stata Specialist, Tom Pullum:

You may be in a situation that I have been in before, where I want to run a model in which some variables are weighted and some are not. The collapse command is either all weighted or all unweighted--at least, I have not figured out how to have some variables weighted and some unweighted. The syntax you proposed, with a weight of 1 for all cases will be exactly the same as no weight.

If this is what you need, then you can do the collapse twice--once weighted and once unweighted. Just before doing the first collapse you would save a working file as file A. Then you would do the collapse with weights and save as file W. Then go back to A, do the collapse without weights, and save as file U. Then merge U and W.

Hope this helps.

Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by user_rm on Tue, 01 Sep 2015 12:34:48 GMT
View Forum Message <> Reply to Message

thanks for your reply. I have been looking at the variable v005 and I have a query relating to it. The weights are the same for all individuals in the district e.g for district Ahmadabad, the sample weight (v005) is 1148330 for all kids in that district. I wanted to ask if I should apply these weights for every individual (ie kid here) or does this number represent the total population for the entire district? I ask because I was using the collapse command to get the data at district level. I used

collapse(rawsum)v005/1000000 , by (district) for the sample weight v005 as well to get 1 figure for each district. But in that case, if there are 25 kids in district, the collapse (rawsum) amount adds the weight 25 times. I was wondering if that is right? Or should I be using just 1148330 as the total weight for all kids in the same district?

Just to be clear I am running regressions at district level. So like you suggested in your previous email, I use collapse commands for controls and outcome variables. I am unsure about how I should use the survey weights? should I collapse(rawsum) v005 or collapse(mean) v005? I apply these weights to my district level regression. Also when running regression should I use aweights or pweights? When I am collapsing data initially then should I use aweights or pweights?

Please clarify.

Thanks
R

Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by Bridgette-DHS on Tue, 01 Sep 2015 16:41:48 GMT
View Forum Message <> Reply to Message

Here is another response from Tom Pullum:

Sampling weights are inflation or deflation factors that have an average value of 1. DHS moves the decimal point six places to the right, that is, multiplies the weight by 1,000,000 (one million). Thus "1148330" is actually "1.148330". The reason for doing this is the same as the reason why percents have a factor of 100, fertility and mortality rates often have a factor of 1000, and maternal mortality rates have a factor of 100,000. It's just to have many significant digits without having to worry (or worry very much) about a decimal point. A weight of 1.148330 for each child in the cluster just means that that the probability that a child in that cluster would appear in the sample was a little less than the average for the whole country, and therefore, to compensate, those children get a weight greater than 1. It doesn't really have anything to do with the population of the district. It is impossible, just from the weights, to estimate the population at any level of aggregation.

When you collapse, you are usually calculating a mean. For example, if you wanted to calculate the mean of some variable x in the district, you would use "collapse (mean) x [iweight=v005/1000000]" . This would give you an estimate of the mean of x in the district, corrected or adjusted for the weights. If you did not use the weights, the mean of x would be biased toward the oversampled children. (With collapse, the default is the mean, so you would get the same thing with just "collapse x [iweight=v005/1000000]".)

I don't know what you are doing with "collapse(rawsum)v005/1000000 , by (district)". You have omitted something, and I don't just mean spaces.

"(rawsum)" indicates a sum, not a mean, and it ignores the weights. I think "sum" is the only statistic that can be prefaced by "raw". There is no "rawmean", for example.

I think you may be making this more complicated than necessary. I would say that you need to use the weights for any collapses and for any estimation commands. Avoid having the numbers come out 1,000,000 times larger than they should be. That's all.

Let me know if you still have questions.

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

Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by Bridgette-DHS on Thu, 10 Sep 2015 10:09:46 GMT
View Forum Message <> Reply to Message

Here is another response from Tom Pullum:

What I like to do is to try out any procedure that I suggest to someone else. In this case, you have done a lot of file preparation, and I cannot take the time to do the merging and recoding to get the data file you are working from.

The weight for the cluster should be the total weight of the children in the cluster, just as you said. You can get that total weight by defining wtd_n=1 and then collapse (sum) wtd_n [pweight=weight], by(whatever). That will give you the sum of the weighted cases If you want an unweighted sum then you could have unwtd_n=1 and then collapse (sum) wtd_n, by(whatever). If you want to combine weighted and unweighted sums on the same records, I think the easiest way is to do those two runs and then sort and merge on "whatever".

By the way, you should not use aweight. It is almost never appropriate.

My bigger question is this: why are you aggregating the data so much? If you know where the intervention occurred, just use that as a binary variable at the lowest level possible. The haz score, for example, is an outcome at the level of the individual child. I would do the analysis at that level, and not collapse at all. I strongly advise against aggregating when you do not have to.

You may be able to use a multilevel model, i.e. a mixed effects model, with the children as level 1 and the village as level 2. Have you tried to do that?

In any case, this looks like an interesting project. Let me know if you have other questions.

Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by user_rm on Fri, 11 Sep 2015 10:12:30 GMT
View Forum Message <> Reply to Message

Thanks for your email.

1. Just to double check - you mean that the weight of each child (e.g 1.148330) in a cluster should be added? So 5 children in the cluster would have a weight of 5*1.148330?

2. The code you suggest, I tried it sometime before, but I think this method would work only if the number of kids in each cluster are the same, which is not the case here. I am not entirely sure the exact nature of the issue I encountered (as it was a while back I was trying various things), but I will try again and see if it works this time. The code I sent should probably do the same thing I think?

3.a) aweights should not be used with the aggregated regressions equations either?
b) Also stata is giving me the option of only aweight or fweight only for making the mean and sd table using 'tabstat' command.
Is there another way to get this table then?

4. My analysis is at the individual level. The aggregated weighting I have done just to check if the results differ from individual ones. It's just a check. Ideally the coefficients shouldn't change, but i suppose the standard errors would.

Many thanks
R

Subject: Re: Weighting variables in DHS India data (1992 and 1998)
Posted by Bridgette-DHS on Fri, 18 Sep 2015 18:17:11 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Stata Specialist, Tom Pullum:

For #1, yes, you would add the weights. For #2, I'm not sure what the issue is. What I described did not require that the number of kids is the same in every cluster. I would never have assumed that.... For #3a, I recommend that you never use aweights. Look at the help for aweights. It has been years since I need them. For #3b, you can use svyset and then svymean, etc. Or you can

use v005 as an fweight. The means and standard deviations will be fine. If you need standard errors, just multiply the calculated standard errors by 1000 (the standard errors are inversely proportional to the square root of the sample size). For #4, I agree that the individual-level and aggregated analyses will agree if the outcome is at the individual or aggregated level and all the covariates are at the aggregated level. They will not agree if any of the covariates are at the individual level.

#2 is still unclear to me.

If you have more questions about that, please send some Stata code.... Good luck!
