Subject: Weighting Data for Pooled Indonesian DHS Dataset
Posted by adelia on Fri, 17 Jun 2016 04:07:21 GMT
View Forum Message <> Reply to Message

Hi!

I have several questions regarding the weighting and survey set of Indonesia DHS dataset 2007 and 2012

So currently I want to pooled the BR dataset for both period and would like to generate pooled wealth index using PCA in order to examine the change in wealth between the years, in which I need to weight the data. I'm using STATA for the analysis, and from what I understand I must do the following before I merge the dataset

1. De-normalizing the sample weight

gen wgt=v005/1000000
gen wgt_denorm=wgt*total number of female population age 15-49 in the particular period / number of female age 15-49 interviewed in the survey

My question about this part is whether it is necessary to use the total number of female population age 15-49 in the particular period or it is sufficient to do the following command?

scalar TOTWT=1000000
quietly summarize v005
scalar T=r(sum)
gen wgt_denorm=v005*TOTWT/T

If it is indeed necessary to use the total number of female population age 15-49 in the particular period, will it be okay for me to estimate this number by multiplying the total country population projected for that particular period with the percentage of female age 15-49 in the household population as reported in the DHS final report on that year?

2. Correcting the PSU and strata to make it specific for the particular year e.g. by adding 10000 for PSU and strata in 2007 and 20000 for PSU and strata in 2012

Question for this part, I generate the strata manually by using the following command

egen strata=group(v024 v025), label

Is this the correct strata to use for Indonesia DHS 2007 and 2012? Or should I use v022 instead?

3. Setting the survey design (after merging the dataset) with the following

svyset [pweight=wgt], psu(v021) strata(strata)

Are these steps correct and do I need additional steps to correct the weighting?

Thank you very much beforehand

Subject: Re: Weighting Data for Pooled Indonesian DHS Dataset
Posted by Bridgette-DHS on Mon, 20 Jun 2016 12:13:25 GMT
View Forum Message <> Reply to Message

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

When you pool the surveys, I recommend that you construct a variable called "survey" with survey=1 for the 2007 survey and survey=2 for the 2012 survey. There are other ways to distinguish between the surveys but this an easy way.

You then construct a unique cluster code with "egen cluster=group(survey v001)". Note that v001 and v021 are the same and you can use v021 here if you prefer.

You construct a unique stratum code with "egen stratum=group(survey v024 v0025)". In the 2012 survey, v022 and v023 are identical and they are a crossing of v024 and v025. In the 2007 survey, both v022 and v023 are defective and the ONLY way to get the strata is by crossing v024 and v025. (The 2007 survey also has a defective label for v007, year of the survey.) When you pool surveys, the clusters and strata id codes must be different for the surveys. The strategy of adding 1000 or 2000, etc., to the codes is obsolete; "egen group" is much better, and as you note it can include labels.

The label for v024 (region) is called HV024 in the 2007 survey and hv024 in the 2012 survey. There are some differences in the labels that I think are misspellings or changes in official names. You must be very careful when pooling surveys because codes can change, especially for country-specific variables. The labels for the last survey in the append sequence will over-write the previous labels and you will not be alerted to any differences or changes.

Apparently you want to re-weight the surveys in such a way that the weighted sample sizes are proportional to the number of women in the population. You can do that either before or after the appending or pooling. Or, if you just want the weighted number of cases to be the same in each survey, you can do that. Your proposed recodes should work, but you need to confirm at the end that the total weights have the property you were trying to achieve. You also want to make sure that the average weight has lots of digits before the decimal point. The mean weight of the original hv005 or v005 will be close to 1000000. You definitely do not want the weights to end up with only one or two digits before the decimal place. Preferably 6 to 8.....

Subject: Re: Weighting Data for Pooled Indonesian DHS Dataset
Posted by adelia on Mon, 20 Jun 2016 13:12:31 GMT
View Forum Message <> Reply to Message

Thank you very much for your reply.

Regarding the weighting, I did the re-normalization as follow for IDHS 2007 BR, is this correct?

. gen wgt=v005/1000000

. gen wgt_denorm=v005*64140000/32895

. gen wgt_denorm_des=wgt_denorm/1000000

. sum v005 wgt wgt_denorm wgt_denorm_des

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        v005 |    84,726    955475.8     1167341     21872    7968061
         wgt |    84,726    .9554758    1.167341    .021872   7.968061
  wgt_denorm |    84,726    1.86e+09    2.28e+09   4.26e+07   1.55e+10
wgt_denorm~s |    84,726    1863.025    2276.129   42.64691   15536.45
```

My intention is to use this weight to generate a pooled wealth index using IDHS 2007 and 2012. I think for this, in the pooled data the weighted sample sizes need to be proportional to the number of women in the population, no? Or the same weighted number of cases in each survey is enough?

Many thanks beforehand

Subject: Re: Weighting Data for Pooled Indonesian DHS Dataset
Posted by Bridgette-DHS on Mon, 20 Jun 2016 14:12:03 GMT
View Forum Message <> Reply to Message

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

As I understand it, you are putting the 2007 and 2012 surveys of Indonesia into a single data file in order to analyze changes and differences. This is convenient for data processing and I would do the same thing.

The question you are asking is only relevant if you want to calculate a mean for the two surveys, and I would definitely not recommend doing that. For example, I would be interested in the CPR in each survey and the change in CPR between the two surveys, but I hope you are not planning to calculate the mean CPR of the two surveys combined. Are you? If so, why? How could you interpret such a number? (I mention the CPR just as an example.)

If you want to do what I think you want to do, that is, to estimate a change from 2007 to 2012, and calculate a confidence interval for the difference, or test whether the change is statistically significant, then you will get exactly the same result if you leave the weights alone OR if you

re-scale them so the total weight is the same in each survey OR if you re-scale them so that the total weight in each survey is proportional to the population size. I encourage you to compare these options. Let me know if they are NOT all the same, in terms of point estimates, standard errors, confidence intervals, and test statistics.

Subject: Re: Weighting Data for Pooled Indonesian DHS Dataset
Posted by adelia on Mon, 20 Jun 2016 14:34:16 GMT
View Forum Message <> Reply to Message

Thank you for your reply.

So what I am trying to do is to decompose the change in inequality (measured using Erreygers Index) of infant mortality and under-5 mortality, whether the change is due to the change in the association (the coefficient) of wealth and mortality or due to the change of wealth (i.e. more wealthy people in 2012 and 2007). This analysis are usually done using income instead of wealth index, thus the change in income can be analyzed easily since income has absolute value. However, wealth index is based on rank, has no absolute value, and the distribution of its value will always be the same as well so I cannot really examine the change in wealth of the population. I try to analyze the change in wealth by computing pooled wealth index using 2007 and 2012 dataset and then see whether there are more people with higher wealth index in 2012 than 2007. In order to calculate this pooled wealth index, I think I need to use the re-normalize sample weight is it correct? Do the weighted sample sizes need to be proportional to the number of women in the population or the same weighted number of cases in each survey is enough for this?

Many thanks

Subject: Re: Weighting Data for Pooled Indonesian DHS Dataset
Posted by Bridgette-DHS on Thu, 23 Jun 2016 14:55:41 GMT
View Forum Message <> Reply to Message

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

Normally I would not recommend that you apply the wealth index from one survey to another survey. However, that may be best here. That is, I suggest that you look up (on www.dhsprogram.com) the formula for the wealth index in the 2007 survey, and then apply it to the 2012 survey, with the same cut points for the quintiles (of equal size in 2007 but not in 2012). You can then do a decomposition of the change in mortality, with components for change in the wealth-specific mortality rates and change in the distribution of wealth (and, depending on how you do it, a component for the interaction between the two).

For that purpose I am pretty sure that there is no need to re-normalize the weights, but that may depend on what decomposition procedure you use. I would first do the analysis with the original weights. I would then repeat it with the weights in one survey changed by some arbitrary factor, such as 2, just to see whether that affects your decomposition. If it does, then re-weighting should

be considered. The options are somewhat arbitrary. The total weighted number of births in each survey, rather than the number of women, could be what you want to match with the population, but then you get into the potential role of changes in wealth-specific fertility as a source of change in mortality. This is really your decision and I am not comfortable making a recommendation.