Subject: Sample weights and stratification - Nigeria 2008 and 2018
Posted by Goethe2014 on Mon, 11 May 2020 13:48:16 GMT
View Forum Message <> Reply to Message

Dear all,

Currently I am using DHS data in combination with Stata for the first time. I intend to estimate effects employing a Difference-in-Difference estimation on Nigerian DHS data from 2008 and 2018 (Individual/women recode). In this regard I would like to know more about the right way to weigh the data and account for the stratification process.

In literature I found that some scholars combine (append) two sets of data (DHS Year A and DHS Year B) and when running their regression account for the women's sampling weight by just including [pweight=v005]. As far as I understood from the DHS forum and manuals in this case we dont have to divide the sample weight by 1.000.000 as pweight can also handle it without doing so. My question now is whether it is that easy to just use the pweight command on the full/combined dataset as there are women from two distinct survey included whose sampling weight had been calculated for their original dataset (Year A OR Year B). Do I therefore have to reweigh the sample or is it really possible just to make use of [pweight=v005] as the data stems from different women and different year but the same country?

In addition I am also a bit confused whether I have to account for the stratification process which in the case of Nigeria was done by states and rural/urban. Some literature accounts for that fact, others ignores the stratification process.

Lastly, I struggle whether I have to make use of the svyset command at all when using DHS data. Again some literature just specifies the data as panel data using xtset command while others suggest svyset commands to account for the DHS survey characteristic.

In a paper which asks similar research questions, DHS data from two years from the same country is used and the authors also employ a Diff-in-Diff estimation. First, they define the data as panel data by using xtset command and then already run their regression model only including [pweight=v005] and vce(cluster v001) at the end.

I would really appreciate any help in order to generate the most robust results and understand DHS data better in general.
Greetings

Subject: Re: Sample weights and stratification - Nigeria 2008 and 2018
Posted by Bridgette-DHS on Wed, 20 May 2020 12:28:42 GMT
View Forum Message <> Reply to Message

Following is a response from DHS Research & Data Analysis Director, Tom Pullum:

These are good questions. I will give recommendations but there is some room for different approaches.

I would definitely combine the surveys into a single file, but when you do this, you need an identifier for the survey. I would probably just construct survey=1 or 2 for 2008 and 2018, respectively. You will need unique identifiers for strata and clusters. I won't take the time to look up the stratum identifiers for these two surveys but let's assume it was v023 in each of them. You construct the combined stratum identifier with "egen stratum_ID=group(survey v023)". You construct the combined cluster identifier with "egen cluster_ID=group(survey v001)".

In each survey, v005 has been normalized so that the sum of the weights is the total number of cases (times 1000000), and therefore the mean weight in each survey is 1 (times 1000000). (Stata automatically re-normalizes pweights and that gets rid of the factor of 1000000.) When you combine the surveys, the overall mean of v005 will also be 1000000. You do not need to do anything with v005. That is, you do not need to do any re-scaling or re-normalizing.

You then use svyset including adjustments for weights, clusters, and strata.

I recommend that you only use the combined file for looking at differences between the two surveys. The aggregate of the two survey (e.g. the CPR for the pooled surveys) is not meaningful. If, say, you were combining surveys from several countries for a pooled analysis then you might want to re-scale the weights to take account of differences in sample sizes and/or population sizes but I would be very cautious with such an analysis, and fortunately that's not what you are talking about. Successive DHS surveys in the same country are well-suited for a difference-in-differences approach.

Good luck.

Subject: Re: Sample weights and stratification - Nigeria 2008 and 2018
Posted by Goethe2014 on Wed, 20 May 2020 15:57:12 GMT
View Forum Message <> Reply to Message

Dear Tom,
Thank you for your reply (I assume the second reply was not meant to answer my question?).

If I understood correctly: First, I do not have to reweigh the sampling weights when combining the Nigerian DHS 2008 and 2018 IR data so I will stick to v005 as sampling weights.

Second, I have to create new stratum and cluster identifiers having combined/appended the surveys. I would code
gen survey=.
replace survey=1 if v007==2008
replace survey=2 if v007==2018

Regarding stratification: Both DHS surveys (according to the final reports) have been stratified by state and urban/rural. Now the issue is that in the DHS2008 v023 accounts only for the state. In the DHS 2018 v023 accounts for region, state and rural/urban. How could I deal with this difference in v023 definition? In general the surveys make use of the same stratification process

but v023 only indicates this stratification process completely in the DHS 2018. (i.e. "label list V023" for DHS 2018 shows that there exist 74 unique values of v023 (37 states either rural or urban) whereas "label list v023" for DHS 2018 shows that there exist 37 unique values (37 states only).

Regarding clusters: The DHS 2008 data lists 888 unique cluster, the DHS 2018 data lists 1400 clusters. These cluster are created based on the same National Census from 2006 (DHS cluster/PSU= 2006 Census Enumeration Area (EA)). If i now make use of the command "egen cluster_ID=group (survey v001) Stata now lists 2275 unique clusters (1-2275). This seems odd to me as the clusters should be the same.

I have seen that in published papers using the Nigerian DHS 2008 and 2013 data the authors do NOT generate any new variables and stick to v005 as sampling weight, v001 as clusters and do not account for the stratification process at all. Would that also be an option?

Thank you very much in advance!

Subject: Re: Sample weights and stratification - Nigeria 2008 and 2018
Posted by Bridgette-DHS on Wed, 20 May 2020 21:44:49 GMT
View Forum Message <> Reply to Message

Following is another response from DHS Research & Data Analysis Director, Tom Pullum:

Mostly I agree, but with these differences. First, the actual clusters are not the same in successive surveys. Whatever the original codes are for enumeration areas in the sampling frame, they are replaced with new numbers. If another researcher did not construct new distinct id codes, they made a mistake (although the impact may be small).

If you are using Stata and svyset, then yes, you should make the adjustments for weights, clusters, and strata. All three. It's painless. But include the "singleunit" option.

Yes, v023 does not always give the strata. Usually, since 2013 or so, but not necessarily earlier. The DHS website gives the correct stratum variable for every survey. I believe that for both of these surveys, the strata are the combinations of state and urban/rural, v025. I believe state is given by shstate. I recommend the following lines:

* In the 2008 survey
egen stratum_ID_2008=group(shstate v025)
gen tempvar=stratum_ID_2008

* In the 2018 survey
gen stratum_ID_2018=v023
gen tempvar=stratum_ID_2018

* Append and construct "survey" using v007

* In the combined file

egen stratum_ID=group(tempvar survey)
drop tempvar

egen cluster_ID=group(v001 survey)

svyset cluster_ID [pweight=v005], strata(stratum_ID) singleunit(centered)

Let us know if you have other questions or this is not clear.

Subject: Re: Sample weights and stratification - Nigeria 2008 and 2018
Posted by Goethe2014 on Thu, 21 May 2020 09:13:22 GMT
View Forum Message <> Reply to Message

Dear Tom,
Thanks a lot for the advice.

I applied the command first to the DHS 2008 data [egen stratum_ID_2008=group(ssstate v025)].If I tabulate stratum_ID_2008 it now shows me 74 distinct values. If I do the same in the DHS 2018 data [egen stratum_ID_2018=group(ssstate v025)] this also gives me 74 distinct values.

I then appended the DHS 2018 data to the DHS 2008 data [append using "MY_DHS_2018_FILE"]. If I now order the total appended dataset [order stratum_ID_2008 stratum_ID_2018] and browse I can see that of course there is no entry (.) for the stratum_ID_2018 for observations from DHS 2008 which makes sense. If I now make use of the command you suggested [egen stratum_ID=group(stratum_ID_2008 stratum_ID_2018)]this creates missing values only as either stratum_ID_2018 is missing for DHS 2008 data and vice versa stratum_ID_2008 is missing for DHS 2018 data. Therefore the newly generated stratum_ID variable in the appended/combined dataset has 0 entries.
Did I miss something or is there any solution for this issue?

Having solved this I would now have the weights (v005) and the stratification indicator (stratum_ID). If I want to define svyset I therefore know 2 of 3 needed values [svyset [pweight==v005], psu (???) strata (stratum_ID)]. You wrote that I could not assume the clusters (which are equal to the Primary Sampling Unit/psu v021) are the same in DHS 2008 and 2018. How would therefore the complete command for syset look like?

Thanks in advance.
Greetings

Subject: Re: Sample weights and stratification - Nigeria 2008 and 2018
Posted by Goethe2014 on Thu, 21 May 2020 15:39:00 GMT
View Forum Message <> Reply to Message

Dear Tom,

Thank you for the updated code. I applied it to the datasets in the following way (Note: The state variable goes by sstate not shstate in Nigerian DHS data):

1. Open DHS 2008 data file and command:

```
egen stratum_ID_2008=group(sstate v025)
gen tempvar=stratum_ID_2008
```

2. Open DHS 2018 data file, command and save:

```
gen stratum_ID_2018=v023
gen tempvar=stratum_ID_2018
save MYSTORAGEPATH, clear
```

3. In DHS 2008 file command:

```
append using MYDHS2018STORAGEPATH
gen survey=.
replace survey=1 if v007==2008
replace survey=2 if v007==2018
egen stratum_ID=group(tempvar survey)
drop tempvar
egen cluster_ID=group(v001 survey)
```

This results in the following (see screenshots of data - example of DHS2008 section, example of DHS2018 section, last entries of DHS 2008 and first entries of DHS 2018).
I have to admit I can not judge whether my result is now correct and looks how it should look like. The survey variable was correctly coded for sure but maybe you could tell me whether the other values also look like they are supposed to
(e.g. stratum_ID starting going up in odd numbers for 2008 data 1,3,5....147; stratum_ID going up in even number for 2018 data 2,4,6....148; v001 going from 1-888 for DHS 2008 and 1-1400 for DHS2018; cluster_ID ranging from 1-1763 and 2-2275 for DHS2018)

I would be very grateful if you could tell me whether my result looks how it should and I have done everything right or whether there is still an issue with the code applied.

Thank you in advance!

File Attachments

1) DHS2008DHS2018Border.JPG, downloaded 544 times
2) DHS2008Example.JPG, downloaded 543 times
3) DHS2018 Example.JPG, downloaded 555 times

Subject: Re: Sample weights and stratification - Nigeria 2008 and 2018
Posted by Bridgette-DHS on Thu, 21 May 2020 17:54:44 GMT
View Forum Message <> Reply to Message

Following is another response from DHS Research & Data Analysis Director, Tom Pullum:

Everything looks fine to me You can also use the "codebook" command to confirm that the number of distinct codes for cluster_ID and stratum_ID matches the sum of the respective numbers in the 2008 and 2018 files. From what you have said, the numbers do match. I believe you also do not have any "." codes for those two variables in the combined file. You should be ready to proceed. Remember to put "svy:" in front of estimation commands.

Subject: Re: Sample weights and stratification - Nigeria 2008 and 2018
Posted by Goethe2014 on Thu, 21 May 2020 21:59:40 GMT
View Forum Message <> Reply to Message

Dear Tom,

I checked using the codebook command and yes there are no missing values. In the future I will - as you suggested - then use the

svyset cluster_ID [pweight=v005], strata(stratum_ID) singleunit(centered)

command and svy: in front of all estimations in order to account for the survey characteristic. Hopefull all will turn out well!

Thank you very much again for your very helpful advice!