Subject: Pooled Cross sections
Posted by cbdolan on Tue, 14 Jun 2016 13:35:27 GMT
View Forum Message <> Reply to Message

I am using the 2007 and 2013/14 DRC Birth Recode files.

Using DHS forum guidance, I have de-normalized the weights using process outlined below. Then I appended the 2007 and 2013/14 files to start the analytic data set. I have the following specific questions:

1. To get the sample size of women 15-49 I used the individual recode. Is this typical or is there another file/source researchers use for that N?

2. There is a debate on what to do with these de-normalized weights and the use of the stata sub-pop command. I don't know if I can just use the weights as they are (de-normalized) or if I have to do something else before I can use the subpop command at the cluster level (ie. running rural and urban sub-pop analysis).

3. I have been told that if I am using individuals as the units of observation in a regression, then we don't use the sampling weights at all. The thought is that other individual level controls should likely pick up any of the differences in weights. Is this assumption correct?

4. Additionally, some researchers are more comfortable using using microeconomic methods (appropriately clustering standard errors) instead of these survey weights, particularly when using data from more than one survey. Have you come across any references in the literature (or even forum posts) that discusses these advantages/disadvantages with the DHS data?

Thanks in advance for your time.

```
*I'm adding in code to use for pooled weights
use "Y:\Data\4_DHS_BirthRecode\CDBR61FL.dta"
*Original weight in DHS : v005 (which should preferably be divided by 1000000)
generate n_v005=(v005/1000000)
*note this is the population of 15-49 in DRC (2013) from United Nations, Department of Economic and Social Affairs, Population Division (2015). World Population Prospects: The 2015 Revision, custom data acquired via website.
generate P1549=16167000
*note this is the sample size from the individual recode file of women 15-49 interviewed
generate n1549=18827
*Country specific weight :CSW= P1549/n1549 (population aged 15-49 in the country / sample size of )
generate CSW=(P1549/n1549)
*New weight
generate NW=n_v005*CSW
file Y:\Data\4_DHS_BirthRecode\n_CDBR61FL.dta saved
clear
```

use "Y:\Data\4_DHS_BirthRecode\CDBR50FL.dta"
*Original weight in DHS : v005 (which should preferably be divided by 1000000)
generate n_v005=(v005/1000000)
*note this is the population of 15-49 in DRC (2013) from United Nations, Department of Economic and Social Affairs, Population Division (2015). World Population Prospects: The 2015 Revision, custom data acquired via website.
generate P1549=13201000
*note this is the sample size from the individual recode file of women 15-49 interviewed
generate n1549=9995
*Country specific weight :CSW= P1549/n1549 (population aged 15-49 in the country / sample size of)
generate CSW=(P1549/n1549)
*New weight
generate NW=n_v005*CSW
file Y:\Data\4_DHS_BirthRecode\n_CDBR50FL.dta saved
clear

*generate weight: see code at top
*make unique strata values by region/urban-rural)
egen stratum=group(v024 v025)
*tell stata the weight (using pweights for robust standard errors, cluster (psu), and strata
svyset [pw=NW],psu(v021)strata(stratum)
*prefix regrss with "svy:stata will now know how to weight your data and compute the right standard errors */

Subject: Re: Pooled Cross sections
Posted by Reduced-For(u)m on Wed, 15 Jun 2016 19:48:30 GMT
View Forum Message <> Reply to Message

I can "weigh in" (get it!?) on a couple of these....

3. I don't think this makes sense. Weighting is about adjusting for the probability of showing up in the sample. But what might make sense in this description relates to causal effects. If the effect of some variable of interest is common across all people, then you don't need to weight because one observation is as good as any other one. But, if there is unobserved heterogeneity in treatment effects across people, then the average treatment effect you estimate would be biased, because it would over-weight the treatment effect some people got (those with a low sampling weight) and under-weight others (those with a large sampling weight). So if your model is wrong in certain ways, weighting can maybe make it less wrong.

4. Clustering standard errors and using weights address two different problems. Weights will affect both your p-values and (to a lesser extent) your standard errors. Clustering will ONLY affect your standard errors (and CI/pval). My simulations suggest that clustering at too small of a level when using pooled DHS rounds can lead to SE that are way, way too small. The appropriate level to cluster at depends on exactly what you are doing. One good paper on thinking about that,

which is relevant to the DHS context, is the famous "How much should we trust difference-in-difference" paper:

http://economics.mit.edu/files/750

If you tell me what you are trying to estimate (in general) I can maybe give some guidance on how to cluster. But the standard econometric thinking on the matter seems to apply fairly well to the DHS (if you think of appending multiple DHS like appending multiple rounds of the CPS or whatever other household survey is popular in your field).

Subject: Re: Pooled Cross sections
Posted by cbdolan on Thu, 23 Jun 2016 14:22:05 GMT
View Forum Message <> Reply to Message

Thanks for the detailed follow up and paper link. Both were helpful.

I think I have something wrong with the way I constructed the pooled weights based on the results of my descriptives. The N's shouldn't be this dissimilar. (Note: I previously merged the files with the spatial files so I'm using ADM1_CODE as the province level variable).

I annotated the code to clarify the steps. Please let me know if I've missed a step.

tab ADM1_CODE[iweight=NW]

ADM1_CODE	Freq.	Percent	Cum.
Bandundu	329,552,409	16.16	16.16
Bas-Congo	90101292.7	4.42	20.58
Equateur	264,708,100	12.98	33.56
Kasai-Occidental	173,724,497	8.52	42.07
Kasai-Oriental	231,513,280	11.35	53.42
Katanga	215,915,193	10.59	64.01
Kinshasa	182,595,002	8.95	72.96
Maniema	72641458.3	3.56	76.53
Nord-Kivu	135,628,336	6.65	83.18
Orientale	189,207,489	9.28	92.45
Sud-Kivu	153,913,256	7.55	100.00
Total	2.0395e+09	100.00	

. tab ADM1_CODE

ADM1_CODE	Freq.	Percent	Cum.
Bandundu	272,736	12.63	12.63

Bas-Congo	118,666	5.49	18.12
Equateur	300,174	13.90	32.02
Kasai-Occidental	195,252	9.04	41.06
Kasai-Oriental	234,033	10.84	51.90
Katanga	259,660	12.02	63.92
Kinshasa	156,412	7.24	71.17
Maniema	132,302	6.13	77.29
Nord-Kivu	140,773	6.52	83.81
Orientale	203,672	9.43	93.24
Sud-Kivu	145,920	6.76	100.00
Total	2,159,600	100.00	

I did the following to set up the pooled weights:

```
use "Y:\4_DHS_BirthRecode\CDBR61FL.dta"
*Original weight in DHS : v005 (which should preferably be divided by 1000000)
generate n_v005=(v005/1000000)
*note this is the population of 15-49 in DRC (2013) from United Nations, Department of Economic and Social Affairs, Population Division (2015). World Population Prospects: The 2015 Revision, custom data acquired via website.
generate P1549=16167000
*note this is the sample size from the individual recode file of women 15-49 interviewed
generate n1549=18827
*Country specific weight :CSW= P1549/n1549 (population aged 15-49 in the country / sample size of )
generate CSW=(P1549/n1549)
*New weight
generate NW=n_v005*CSW
file Y:\4_DHS_BirthRecode\n_CDBR61FL.dta saved
clear

use "Y:\4_DHS_BirthRecode\CDBR50FL.dta"
*Original weight in DHS : v005 (which should preferably be divided by 1000000)
generate n_v005=(v005/1000000)
*note this is the population of 15-49 in DRC (2013) from United Nations, Department of Economic and Social Affairs, Population Division (2015). World Population Prospects: The 2015 Revision, custom data acquired via website.
generate P1549=13201000
*note this is the sample size from the individual recode file of women 15-49 interviewed
generate n1549=9995
*Country specific weight :CSW= P1549/n1549 (population aged 15-49 in the country / sample size of )
generate CSW=(P1549/n1549)
*New weight
```

```
generate NW=n_v005*CSW
file "Y:\4_DHS_BirthRecode\n_CDBR50FL.dta saved
clear

use "Y:\4_DHS_BirthRecode\n_CDBR61FL.dta"
append using "Y:\4_DHS_BirthRecode\n_CDBR50FL.dta"


*generate weight: see code at top
*make unique strata values by region/urban-rural )
egen stratum=group(ADM1_CODE v025)
*tell stata the weight (using pweights for robust standard errors, cluster (psu), and strata
svyset [pw=NW],psu(v021)strata(stratum)
*prefix regrss with "svy:stata will now know how to weight your data and compute the right
standard errors */
```

Subject: Re: Pooled Cross sections
Posted by Bridgette-DHS on Thu, 23 Jun 2016 19:18:57 GMT
View Forum Message <> Reply to Message

Following is a response from DHS Stata Specialist, Shireen Assaf:

When you compute the NW the magnitude of the weight is increased by a factor of 1000 because of the CSW. It seems you need to divide the NW by 1000, i.e.:

```
gen NW2=NW/1000
tab ADM1_CODE [iw=NW2]
```

If you use svy with pweights the percentage estimates would not be affected by using NW or NW2 just the frequencies. So if you just report the percentages this shouldn't matter and no need to divide the NW by 1000, just use svy and pweights.

I would also suggest not using NW and simply use the weights for the appended data as follows.

```
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
```

You will notice that the percentages are quite close to those produced when using svy and NW or

NW2.

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

Subject: Re: Pooled Cross sections
Posted by cbdolan on Fri, 24 Jun 2016 13:33:43 GMT
View Forum Message <> Reply to Message

Thanks for the reply. Most helpful.

If I follow the suggestion below and eliminate NW then I'm not denormalizing and creating the country specific weight (CSW) that I thought was the recommended process for pooled surveys? Did I miss something obvious?

Thanks again.

Carrie

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

Subject: Re: Pooled Cross sections
Posted by Bridgette-DHS on Thu, 30 Jun 2016 13:46:01 GMT
View Forum Message <> Reply to Message

We already corresponded with the user by email.

Subject: Re: Pooled Cross sections
Posted by joemer on Mon, 22 Aug 2016 06:56:52 GMT
View Forum Message <> Reply to Message

Hi DHS,

The syntax you posted is actually helpful for me since I am using pooled cross sections in the Philippines.

Now my question is: Will this be quite different if I will just used selected population (let's say adolescents) in IR.

Thank you.

Subject: Re: Pooled Cross sections
Posted by CKAllen on Fri, 26 Aug 2016 14:28:38 GMT
View Forum Message <> Reply to Message

Quote:If I follow the suggestion below and eliminate NW then I'm not denormalizing and creating the country specific weight (CSW) that I thought was the recommended process for pooled surveys? Did I miss something obvious?

Thanks again.

I have the same question... I have created de-normalized weights after creating unique strata and clusters for 6 appended surveys (three countries, two surveys each). However when I run my logistic regression my 'population size' is gigantic (see below). I've read through most of the threads on this site and cannot seem to understand why my weighting is wrong. However the post above suggests that I do not need to de-normalize the weights? There are so many differing opinions and methods on this, but I think I'd like to de-normalize my weights in order to account for very different country sizes. This is my first analysis I've done with DHS, so I realize there may be something simple I am missing. Any help is appreciated! thank you!

number of obs = 36,349
population size = 11,882,122,587,765

Subject: Re: Pooled Cross sections
Posted by Bridgette-DHS on Mon, 29 Aug 2016 14:05:11 GMT
View Forum Message <> Reply to Message

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

The word "denormalize" means nothing to me, even though it is commonly used on the forum.

When you use svyset, the type of weight is always pweight. Stata always normalizes pweights so that the mean pweight in the data file is 1, and the sum of the pweights--the weights that Stata is using in the calculation, regardless of what you say--is the same as the number of cases in the file. You can divide v005 by 1000000 or multiply it by whatever you want, and it will have no effect on your results.

However, you can re-weight or RE-normalize PARTS of a data file. When you are putting several countries or surveys into a single file, you can internally re-allocate in exactly the way you say. There have been many postings on this. You can re-weight so that the weighted number of cases for a survey is proportional to the country's population or part of the population (e.g. women 15-49) , as you do. OR you can re-weight so that each file counts equally--that is, if there are k files, the sum of the weights for each file is 1/k times the total number of cases in the pooled file. The danger of proportional allocation is that a country such as India or Nigeria will sompletely dominate the results.

If you are calculating survey-specific estimates or looking at differences between surveys, there is no need to re-weight. You can put the surveys into a single file for data processing efficiency, but leave the weights untouched. The kind of re-weighting you are talking about is only relevant if you want to combine surveys to produce a single estimate, e.g. the contraceptive prevalence rate in West Africa. But then I would ask, is it really meaningful to calculate something like that? The surveys are done at different times, so what is the reference time point of your pooled estimate? Do you have all the countries in West Africa? (The answer of course is "no".) Can't you just give the CPR, say, for the countries and dates that you have, for which no re-weighting is needed? A lot of the forum discussion about weighting in pooled files is based on the desire to do a type of analysis that maybe shouldn't be done at all....

Subject: Re: Pooled Cross sections
Posted by id709nvz on Wed, 07 Oct 2020 01:45:56 GMT
View Forum Message <> Reply to Message

Dear,
I have issues which I believe is related to your response.
I am using combined Ethiopian KR file (2000, 2005, 2011, and 2016). I went through DHS forum to get some insight on how to weight the data.
To start with my research question:
I want to analyze the effect of X on Y where X started in 2011 but the survey were collected few months before, so in general 2011 is also a kind of pre treatment. My analysis is a sort of Diff-in Diff. could you please how to use weights accurately. to be precise I used the following command.

*After opening KR file of 2000 survey
gen wgt=v005/1000000
gen weight=(wgt*FEMALE POPULATION)/FEMALE SAMPLE -For KR 2000; FEMALE POPULATION=14619 and FEMALE POPULATION=15367
gen survey=1

*Same procedure for 2005, 2011 and 2016. Except that I used different total number of female population, female sample and with gen survey 2, 3, and 4 respectively. Then I appended the data starting from the latest survey (2016). Afterward, I did the following.

egen cluster=group(survey v021) *V021 is PSU and survey identifies year of survey (coded 1, 2, 3 and 4)
egen stratum=group(survey v022) *v022 is sample strata for sampling error
svyset cluster [pw=wgt], strata(stratum) singleunit(centered)

Is it correct to use: tab x1 x2 [iweight=wgt] or tab x1 x2 [iweight=weight]
how about svy: reg Y x1 x2 (Will this understand the complex nature of the survey).
Added to this: svy is not supported with ivreg2 stata command is there an alternative to this? I simply used ivreg2 with cluster option at v001.

Thank you in advance for any comments.
Kind regards!

Subject: Re: Pooled Cross sections
Posted by Bridgette-DHS on Fri, 09 Oct 2020 15:57:03 GMT
View Forum Message <> Reply to Message

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

This looks fine to me. The use of egen group is appropriate. Your adjustment to the weights is appropriate, but I believe that for some analyses it will have no effect (that is, the results with adjusted weights will match the results with unadjusted weights).

Yes, using iweight with tab is fine.

I am not familiar with ivreg2 (I see it is in an econometrics package you can add to Stata). Does it allow pweights and the cluster option? If so, I would recommend using them (pweights and the cluster option). Then not being able to use svy will only mean you are dropping the stratification adjustment.

In general, when a complex estimation procedure does not allow some option (such as svy) I would recommend pairing it with a simpler procedure that DOES allow svy. Check them against each other to see what is gained or lost by using the more complex procedure. Then it's a judgment call about which one to use. I would usually go with the one that is more conservative, that is, has fewer significant coefficients, and NOT with the one that is more favorable for your research hypothesis.... Good luck!

Subject: Re: Pooled Cross sections
Posted by id709nvz on Fri, 09 Oct 2020 16:04:11 GMT

Dear,
Thanks, pweight and cluster works with ivreg2. That is the only option I have thus far.

My worry was if anything other than svy doesn't account for complex nature of the survey.

Have a nice weekend.