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

I have set up the weights and adjustments for DHS surveys using the following syntax:

gen wgt=v005/1000000
egen stratum=group(ADM1_CODE* v025)
svyset [pw=wgt],psu(v021)strata(stratum)

Am I correct that the svyset command, when called with svy: at the start of a regression,
produces robust SE clustered at the cluster level? or are these robust standard errors? Is it
possible to correctly use the svyset command and cluster the SE at the ADM1(province level)or is
it better to not call the svy: command and do the following:

regress y a b c...cluster(ADM1_CODE)

*please note: in the 2007 DRC DHS the v024 variable which is typically used in DHS adjustments
contains both numeric and character values for the same province (see below). Therefore, I used
ADM1_CODE and not v024 when making unique strata values by region/urban-rural

province	Freq.	Percent	Cum.
kinshasa	106,141	4.86	4.86
bandundu	231,557	10.60	15.47
bas-congo	86,226	3.95	19.41
equateur	251,785	11.53	30.95
kasai-occidental	153,546	7.03	37.98
kasai-oriental	187,879	8.60	46.58
katanga	214,694	9.83	56.41
maniema	89,496	4.10	60.51
nord-kivu	104,015	4.76	65.28
orientale	224,819	10.30	75.57
sud-kivu	104,501	4.79	80.36
20	33,727	1.54	81.90
30	44,293	2.03	83.93
40	51,749	2.37	86.30
50	33,135	1.52	87.82
61	40,326	1.85	89.66
62	44,224	2.03	91.69
63	40,950	1.88	93.57
70	47,891	2.19	95.76
80	48,764	2.23	97.99
90	43,851	2.01	100.00

```
       Total |  2,183,569      100.00
```

Subject: Re: Clustered Standard Errors
Posted by Bridgette-DHS on Wed, 02 Mar 2016 16:57:26 GMT
View Forum Message <> Reply to Message

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

Yes, if you construct svyset as you say, and then put "svy: " in front of an estimation command, you will get robust standard errors.

You can make the cluster adjustment in two ways. One is with svyset, the other is with regress y x, cluster(v021). (I think you meant to put "cluster(v021)" rather than "cluster(ADM1_CODE)".)

You can make the weight adjustment in two ways. One is with svyset, the other is with regress y x [pweight=v005].

You can also do regress y x [pweight=v005], cluster(v021).

You cannot, however, include the stratum adjustment within an estimation command. That adjustment can ONLY be made with svyset.

Note that v001 and v021 are exactly the same.

Something has gone wrong with v024 in your DRC 2007 data file. I expect that it has been recoded incorrectly. A conspicuous warning that something was wrong is the huge number of cases in the lines with labels.

I don't know which of the DRC 2007 files you are using, but here are the codes for the PR file

. label list hv024
hv024:
        10 kinshasa
        20 bas-congo
        30 bandundu
        40 equateur
        50 orientale
        61 nord-kivu
        62 maniema
        63 sud-kivu
        70 katanga

80 kasa-oriental
90 kasa-occident

In the PR file, the total household population is 48,291. You are way off.

Subject: Re: Clustered Standard Errors
Posted by analyst_till on Wed, 03 Feb 2021 16:09:19 GMT
View Forum Message <> Reply to Message

Dear all,

I am operating a diff-in-diff analysis at the district-level with the Ethiopian DHS surveys of 2000 and 2011 and have a similar issue. Mr. Pullums answer was very helpful, however, I am still wondering about how I can adjust the clustering in the svyset specification.

More precisely, I want to cluster standard errors at the district-level (since this is my pseudo-panel cohort), which is not possible in a "svy:" regression.

If I now regress: "reg y i.post##i.Treatment i.district [pweight=v005], cluster(district)", I can account for weights and cluster standard errors at district level. However, in this case I do not take the stratification into account.
Would it in general also be okay to ignore the stratification in my analysis to be able to cluster SE at district-level?

Or, could you tell if there is a code to directly incorporate the fact that I cluster SE at district level e.g. through?:

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

Thanks in advance!

Subject: Re: Clustered Standard Errors
Posted by Bridgette-DHS on Fri, 05 Feb 2021 14:32:00 GMT
View Forum Message <> Reply to Message

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

Your Stata code, using "egen group", looks fine. The only way to include the stratum adjustment is with svyset.

I don't quite understand the role you want for "district". The purpose of svyset is to adjust for the design effect in multi-stage samples. For DHS data, the clusters are the Primary Sampling Units (PSUs). The strata are essentially subpopulations within which separate samples are drawn. The adjustments for clusters and strata work in opposite directions, but only affect the standard errors of the estimates. Districts are administrative units that usually have no direct role in the design of the sample.

However, you can conceptualize a multi-level analysis in which, say, respondents are level 1, clusters are level 2, and districts are level 3. The justification would be that individuals within the same district are more similar than individuals in different districts, and you have some district-level covariates. We encourage the use of spatial covariates, but at the finest level of aggregation, which would be the cluster rather than the district. Please clarify.

Subject: Re: Clustered Standard Errors
Posted by analyst_till on Fri, 05 Feb 2021 16:01:20 GMT
View Forum Message <> Reply to Message

Dear Mr. Pullum,

Thanks a lot for your answer!

To clarify first: The first question in this thread (mentioning the egen command) was not posed by me. I just posed my question in this thread, since I thought to have a similar issue.

You are right, I have to be more precise about what I plan to do:
I want to evaluate the impact of a large employment programme in Ethiopia on certain outcomes, using diff-in-diff. Since the programme was implemented at the district-level, my identification strategy would consist in comparing changes between districts where the programme has been implemented and where it has not been implemented. Therefore, my unit of analysis is the district. For this purpose, I assigned district names to DHS clusters, using the DHS geographic dataset and GIS software (I am aware of possible issues with representativeness at the district-level).

Now, I want to run a diff-in-diff regression, aggregating individual observations at district-level and also cluster the standard errors at district-level.
Also, I wanted to account for stratification and weighting, following exactly the instructions of another thread:

**in the 2000 sample

egen stratum_ID_2000 = group(v024 v025)
gen tempvar=stratum_ID_2000

**in the 2011 sample

```
egen stratum_ID_2011 = group(v024 v025)
gen tempvar=stratum_ID_2011
```

**in the appended panel of both rounds

```
gen post =.
replace post =1 if v007 == 2003 & !missing(v007)
replace post =0 if v007 == 1992 & !missing(v007)
```

*construct unique identifiers for strata and clusters

```
egen stratum_ID = group(tempvar post)
drop tempvar

egen cluster_ID = group(v001 post)

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

In general, I was wondering if this procedure is still correct if I run my analysis at district-level? Or, should I code "egen Cluster_ID = group(district post) instead?

I would be very grateful, I you could give me some advice on how to proceed with the DHS weighting and stratification procedure in my case.

Thanks in advance!
Till

Subject: Re: Clustered Standard Errors
Posted by Bridgette-DHS on Mon, 08 Feb 2021 13:17:48 GMT
View Forum Message <> Reply to Message

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

It's great to hear that you are using DHS data in this way. We have done some related work (https://www.dhsprogram.com/pubs/pdf/WP142/WP142.pdf). Unfortunately I don't have time to give a detailed response, but I can make some suggestions. First, I recommend the usual svyset for pooled surveys. Second, and most important, I would recommend working with individual-level data rather than aggregating. You construct two binary variables. The first one is S, which is 0 for the first survey and 1 for the second survey (pre-and post-intervention). The other is (say) A, which is 0 in a control area and 1 in an intervention area (area=district). The difference-in-differences approach is equivalent to assessing the significance of the interaction between A and S. If you have a binary outcome Y, then in the pooled file you do a logit regression of Y on A, S, and AS=A*S. You can include other controls, because interventions are not usually assigned at random. Then look at the sign and significance of AS. That's what we did in WP142, with the Uganda 2011 and 2016 surveys. I also applied this approach to the 2005 and 2010 surveys in Rwanda (https://www.ghspjournal.org/content/2/3/342/tab-supplemental) . If you

collapse the individual-level responses and use districts as units of analysis you get into various statistical issues that can be avoided with the individual-level data.

Subject: Re: Clustered Standard Errors
Posted by analyst_till on Tue, 09 Feb 2021 10:04:50 GMT
View Forum Message <> Reply to Message

Dear Mr. Pullum,

thanks a lot again, the study you mentioned will probably be interesting for me to look at.

I am sorry, my formulation was not well put. In fact, I just wanted to say that I will compare observations in the same districts across time, with districts being my panel variable. However, I do not aggregate my individual observations at district level in the dataset, such that the final outcome Y is the outcome of individual i in district j at time t. Also, I proceeded exactly as you mentioned, such that my regression equation looks as follows (with "post" instead of "survey"):
svy: reg y i.post##i.Treatment i.district

However, I have have two small questions left:

1. If you recommend sticking to "svyset": how can I make sure that I cluster standard errors at district-level then? It does not seem to be possible within the "svyset"/"svy:" framework, or am I wrong?

2. Including robustness checks, I will probably use 4 different DHS rounds for Ethiopia. I noticed that the strata variable v023 was not always coded in the same way (e.g. in one round 22, in the other round 21 strata). Is it best in this case to rely on "egen strata=group(v024 v025)" in each of the rounds, to make sure that the strata variable is indexed in the same way across surveys and afterwards proceed as noted above in the pooled data?

Best,

Till

Subject: Re: Clustered Standard Errors
Posted by Bridgette-DHS on Tue, 09 Feb 2021 13:49:07 GMT
View Forum Message <> Reply to Message

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

The clustering is by PSUs, not districts, so svyset does not need to include any reference to clusters. You can refer to a posted file (survey_strata.do) that describes the strata in all DHS surveys. In each survey, name the identifier "stratum_ID". Then construct the combined ID as, for example, stratum_ID_all, within "egen stratum_ID_all=group(survey stratum_ID)". Also "egen cluster_ID_all=group(survey v001)". That should work.

I'm not sure that your regression needs to include "i.district". That term could overlap too much with "i.post##i.Treatment". I would try regressions with and without that term. Good luck.

Subject: Re: Clustered Standard Errors
Posted by analyst_till on Tue, 09 Feb 2021 17:24:50 GMT
View Forum Message <> Reply to Message

Thanks for the tips concerning the Strata.

With respect to the districts: Although my unit of analysis is the individual, my identification strategy relies on changes at the district-level. Therefore, I include district fixed effects in my regression framework and as far as I know, these are reflected by "i.district".
Please correct me if I am wrong, but I guess this term is crucial to guarantee that I compare only observations of the same districts over time?
As far as I know, STATA breaks the arising collinearity between i.district and i.treatment by dropping one of the district fixed effects.

Also, I am still a bit confused about the clustered standard errors. According to Duflo et al. 2004, one can tackle the issue of serial correlation in a diff-in-diff by clustering standard errors at group-level, allowing for autocorrelation within the groups.
Since the group/cohort in my panel is district, shouldn't I cluster standard errors at district level?
If standard errors are clustered at PSU-level through "svyset", I would not cluster SE at the level of my cohort.

Is it possible that I have to code?:
svyset districts [pweight=v005], strata(stratum_ID) singleunit(centered)

Sorry if it was not clear, I hope you understand what I mean!
Best,

Till

Subject: Re: Clustered Standard Errors
Posted by Bridgette-DHS on Wed, 17 Feb 2021 14:00:29 GMT
View Forum Message <> Reply to Message

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

Sorry for the delay with this reply. Yes, you can make a case for using districts as clusters, even though they are not PSUs. But you may run into a problem because the weights vary within districts. The weights (v005) are constant within v001, but they are not constant within districts. Stata will not like that. Perhaps you can think of a way to get around that, but I don't know what it

would be. Re-calculating the weight to be the within-district average of v005 will lead to some inconsistencies. That is, analyses using the original v005 will not agree (exactly) with analyses using the within-district average of v005.

Subject: Re: Clustered Standard Errors
Posted by analyst_till on Mon, 22 Feb 2021 14:50:48 GMT
View Forum Message <> Reply to Message

Thanks again for the answer.

Indeed, when I try to run districts instead of clusters in the "svyset" framework, STATA is giving me missings for the F-statistics.

To sum up my issue: I have districts as groups and clusters as PSUs in my framework. I also understood from the literature that one should first of all cluster SE at the PSU-level, corresponding to the survey design.
But is it not an issue to cluster standard errors at another (lower) level than the groups that I use in my diff-in-diff?

May I finally ask, how you would proceed in this case? I read your abovementioned paper about the SMGL in Uganda and if I am not wrong you also compare changes in districts over time. So, I guess you clustered SEs at the PSU-level here and this is in general a valid approach to proceed (although groups are districts) ?

Thanks a lot for your time!
Till

Subject: Re: Clustered Standard Errors
Posted by Bridgette-DHS on Tue, 23 Feb 2021 12:08:30 GMT
View Forum Message <> Reply to Message

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

For the SMGL analysis in Uganda, and for an earlier analysis in Rwanda, the svyset adjustments were for PSUs even though the units for the difference-in-differences model were districts. I really don't think there is a problem with this approach. Say that you had any binary predictor, such as urban/rural, and two surveys, and you wanted to test whether the change (in some outcome) between the two surveys was the same in both urban and rural areas. You would be testing for the significance of the interaction term. Or maybe your predictor was not spatial at all, for example a binary version of education (e.g. <=primary and > primary). You would test the significance of the interaction term. Isn't your situation equivalent to that? Your binary predictor is whether the person/household is in or is not in an intervention area, and you want to test whether the change between the two surveys was the same in both the intervention areas and the control

areas.

The svy adjustments are intended to compensate for how a DHS sample deviates from a simple random sample. If the district ID is not relevant to the sampling design, I don't see why you need to include district ID in the svyset command.

Yes, the above described situation fits my case and I understand why we adjust for the sampling design.

My confusion was more specifically about at what level I should cluster standard errors. Abadie et al. (2017) argue that clustering SEs is justified either by the sampling design (the case for DHS), or by the experimental design, when clusters of units are assigned to the treatment.

For my analysis the latter case also applies, since districts were assigned to the programme I evaluate. Therefore, I was wondering whether sampling process or experimental design have to be addressed in the first place.
So, I guess it is the sampling design that has to be addressed in the first place, as you mentioned, and therefore standard errors should be clustered at PSU-level here.

Thanks a lot for your help again!

Thanks a lot!

I have a last concluding question: Since we discussed the clustering of standard errors here, I was confused about how the svy-adjustments directly relate to the clustering of standard errors? In this thread above, you mentioned that one can make cluster adjustments either via svy or via "reg, cluster(v021)". Could you tell me which part of "svyset" defines that I will cluster SEs at a certain level, is it "svyset PSU...." Or the "singleunit(centered)" option at the end?

Here, since I am working with repeated cross-sectional data over several years, you recommended to include a grouped PSU variable, namely "egen cluster_ID = group(v001 survey)".
Does this mean that I cluster standard errors at PSU-level, or at the PSU-time level (since I interact PSU with survey) here? And if yes, does it always has to be the PSU-time level at which I cluster standard errors?

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

The generic form of svyset is this:

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

In this syntax, the first variable after "svyset" (here "cluster_ID") is the PSU. "singleunit(centered)"
is related to the "strata(stratum_ID)" term. It keeps the program from crashing if it encounters only
one PSU within a stratum. There are a couple of alternatives to "centered" but I have done
comparisons and the results are indistinguishable for the different options. You are usually ok
without the singleunit option but I usually include it because I hate crashes.

Below I will paste an example from one time when I was pooling two surveys from 2008 and 2018.
The two surveys had different specifications of strata. You can see how "egen group" was used.
Note that "egen group" does NOT combine or pool. It does just the opposite. For example, say
that in the PR file you wanted to construct a variable for all combinations of urban/rural (hv025=1
or 2) and male/female (hv104=1 or 2). You would use "egen place_sex=egen(hv025 hv104)" to
get a four-category variable for the combinations of hv025 and hv104. This can be handy for
making tables or interpreting interaction terms. In the example below, it basically distinguishes
the designs of the 2008 and 2018 samples. Hope this helps.

* 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)

Subject: Re: Clustered Standard Errors

Posted by analyst_till on Wed, 24 Feb 2021 14:42:42 GMT

So actually "group" helps us to distinguish the (different) designs of the surveys, but I would still say that I cluster standard errors at PSU-level/obtain PSU-level robust standard errors?

I guess that this is something I would always have to include.
So when I would like to cluster standard errors at district-level, I first run "egen district_id = group(district survey)" and then "reg y x, cluster(district_ID)"?

Thanks a lot!

Subject: Re: Clustered Standard Errors
Posted by Bridgette-DHS on Wed, 24 Feb 2021 15:09:33 GMT

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

You would specify svyset as I described, and then run "svy: reg y x". This will provide all the adjustments.

I still don't understand why you want to treat districts as clusters. Perhaps you are thinking of a multi-level model, in which individuals are nested in clusters and clusters are nested in districts. I just don't think that's necessary or helpful. Perhaps other users will contribute some thoughts, but I cannot add anything to what I have already said.

Subject: Re: Clustered Standard Errors
Posted by analyst_till on Wed, 24 Feb 2021 15:38:03 GMT

I am interested in clustering at district-level since the policy I evaluate was introduced at this level. Econometric literature (such as Abadie et al. (2017)) suggests that clustering is justified either to account for the sampling design, or to account for the experimental design/treatment assignment. But I have to look at this again.

My last question was rather referring to the technical adjustments, if I always have to inlcude the "group" adjustment to cluster.

Thanks a lot for your help again!

Subject: Re: Clustered Standard Errors
Posted by id709nvz on Thu, 24 Mar 2022 23:57:07 GMT
View Forum Message <> Reply to Message

Dear,
I am into exactly your problem. Did you manage to find a solution, please?
I am doing a Diff-in-Diff regression and the reform was passed at the regional level, unlike yours
(district) and hence SE must be clustered at the regional level. But I find it really difficult to
understand the procedure with pooled Dhs data.

Please, can you kindly forward your solutions?

Best regards,

Subject: Re: Clustered Standard Errors
Posted by analyst_till on Sat, 26 Mar 2022 09:30:40 GMT
View Forum Message <> Reply to Message

Hi,

Following recommendations of my supervisor, I finally did not adjust for the sampling/survey
design in my specific case.

So, for a regression I could simply do:

reghdfe y post##treatment, cluster(districts) absorb(districts)

Sorry that I can not come up with a proper solution to the abovementioned problem.

Subject: Re: Clustered Standard Errors
Posted by id709nvz on Thu, 01 Sep 2022 16:52:00 GMT
View Forum Message <> Reply to Message

Dear Till,
Could you please kindly gudie me on how you managed to join Ethiopian district names to the
GPS coordinates provided by DHS? Just a highlight would be enough.

Please kindly suggest me.
Thank you in advance!