Subject: declaring child survey in Stata
Posted by musti on Thu, 05 Mar 2015 16:04:04 GMT
View Forum Message <> Reply to Message

Dear Sir/Madam,

I am using 2003 and 2008 child data of Turkey Demographic and Health survey.

With the recommendation of my friend, I pooled 2003 and 2008 data sets. But there is a problem. The pooled data has only v000, v001, v002, v003, v005, v008 variables to declare the survey data in to stata. And when I tried to declare the survey data set to stata 11.2. i had the following results. Apperently it says "stage 1 is sampled with replacement; all further stages will be ignored."

. svyset v001 [pweight=v005], vce(linearized) singleunit(missing) v002 v003
Note: stage 1 is sampled with replacement; all further stages will be ignored

pweight: v005
VCE: linearized
Single unit: missing
Strata 1: <one>
SU 1: v001
FPC 1: <zero>

But i am sure the design of this survey has more than one stages. taking into consideration this fact, the following variables seem relevant:v004, v021, v022, v023. But i have no idea how to use them to declare the data set into stata. And which one of these variable i need to use to define the survey?

Stage 1 sampling units	strata	finite pop correction
Stage 2 sampling units	strata	finite pop correction
.....	,,,,,	.....
.....	.....	.....

What is the sampling units, strata and finite pop correction in each stage for DHS child data or does it change for child data and mother data?

2)the other question is whether or not i need to declare data set as survey in order to get correct results for regression, especially Difference in Difference and IV.

3) Another question is related to sample weight. I divided my sample weight variable with the correct denominator. But now i do not know how to use it?

in the section where stata define your survey data there is another section for sample weight. if i enter my sample weight there, will the problem be solved???

if i do not use survey command for calculation of regression or tables in final reports how should i use sample weight?

4)From guide to DHS statistics:

In SPSS using the WEIGHT command with the weight variable:
COMPUTE rweight = V005/1000000
WEIGHT by rweight.
b) In ISSA using the weight parameter
rweight = V005/1000000
x = xtab(table1, rweight).

How should i write the above commands in stata 11.2? for a regression and table with or without defining the survey data?

5) lets say i need to enter the V004, v021, v022, v023. are these variable standart and so i can directly copy pasted 20008 under 2003?

Thank you in advance for your valuable comments,

regards

Subject: Re: declaring child survey in Stata
Posted by Trevor-DHS on Thu, 05 Mar 2015 20:48:50 GMT
View Forum Message <> Reply to Message

1a) Why are you pooling the data? What benefit do you get from pooling the data over running two separate analyses? Are you trying to compare changes over time?
1b) You cannot use the PSU and stratum variables as they are. You need to create new PSU and stratum variables that are specific to each survey, e.g. egen newpsu = group(survey_year v001) and egen newstrata = group(survey_year v023)
1c) You probably have to denormalize the weights for your analysis. There are several posts on the forum about denormalization.
1d) your svyset command should look more like svyset newpsu [pweight=wgt], stratum(newstrata) where wgt is the denormalized weight. Even if you weren't denormalizing, you would need to first divide v005 by 1000000, but as you are pooling data you should be denormalizing the weights.

2) If you don't use svyset and svy: regress (or similar commands) then your tests of significance will be incorrect.

3) For the weight, see how it is used above in the svyset command. Whenever you use an svy: command Stata will refer to the weight in the svyset command.

4) Using the weight when not using the svy commands:
gen wgt=v005/1000000
tab var1 var2 [iw=wgt]
Using the weight with the svy commands:
gen wgt=v005/1000000
svyset v001 [pweight=wgt], stratum(v023)

svy: tab var1 var2
[this assumes the strata are in v023 - sometimes they are in v022 and sometimes they need to be created. See other posts concerning the strata variables to use.]

5) The variables you mention are standard variables, but I would never copy and paste data as you suggest. You should use the commands for combining datasets:
http://www.stata.com/manuals13/u22.pdf

Subject: Re: declaring child survey in Stata
Posted by musti on Fri, 06 Mar 2015 10:46:28 GMT
View Forum Message <> Reply to Message

Thank you.

1a) -I am pooling the data sets because I will use difference in difference method with Pseudo(pooled) cross section data. I want to see the effect of an education reform. the policy change happened in 1997 and therefore my control and treatment group are not affected in 2003 data. and in 2008 data, my treatment group affected and control group not affected.

So the data set i want to create in the end will look like:

number of observation ID variables year 2003 2008 independent variables
1 .... 2003 1 0 age sex .....
2 ... 2003 1 0 age sex .....
3 ... 2003 1 0 .....
....
....
25..... .. 2008 0 1 age sex .....
.....
45 2008 0 1.....
.....

- I will add the data set to the post, it is not in the format i wanted. Could you please check it? the date set attached to the mail is not in the format i wanted. i do not have the dummy variables. and instead of year veriable i have survey phase variable.

1b) Shoul i do 1b before pooling the data? By the way, voo1 in my data set is corrected before pooling the data. So my PSU, primary sampling unit is v001? If you notice, in the normal data set it is VOO1 but in mine it is v001. what i mean is that some of the variables not standard in both data sets before pooling. and therefore before pooling the data set they were corrected. But how should i include v023 in this data set?

should i ignore the other variables such as v021, v022 and v004.

1d) v005 was also different in the real survey but we corrected it like:

v005= V005/1000000. this is done at cross section level. then two data set is combined. should i change it at cross section level? or i can do it at pseudo panel level?

4) how do I understand strata is V022 format not in V023 format. Becuase i have both variable in the full data set of TDHS 2003 and 2008.

5)We combine the data set in spss. I will add the pooled format to the mail. Could you please check it?

File Attachments

1) 2003-2008.zip, downloaded 1234 times

Subject: Re: declaring child survey in Stata
Posted by musti on Fri, 06 Mar 2015 16:14:23 GMT
View Forum Message <> Reply to Message

1) By the way,
in 2008 TDHS V022(stratum number) has no observation.

when i browse V023(sample domain), the only value it has is national. it is shown as V023 national national national national national.....

V021(primary sampling unit) has values 101 101 101 102 102 102 103 103 103 104 104 104 104.....

And I have V024(region) west west west west..... south south ..... central central central central..... north north north.....
east east east.....

http://www.hips.hacettepe.edu.tr/eng/tdhs08/TDHS-2008_Main_R eport.pdf

When i check TDHS 2008 main report from the above link related to sample selection in the appendix page 209, it says staratification is done for region, then IT SAYS REGIONAL BREAKDOWN EXTENDED TO nuts 1,2,3 regions. Page 209 describes this. I put A few lineS of the table describing the stratums.

tab V024

Region Freq. Percent Cum.

West 651 16.88 16.88
South 497 12.89 29.76
Central 666 17.27 47.03
North 352 9.13 56.16
East 1,691 43.84 100.00

Total 3,857 100.00

tab V024

Region Freq. Percent Cum.

West 651 16.88 16.88
South 497 12.89 29.76
Central 666 17.27 47.03
North 352 9.13 56.16
East 1,691 43.84 100.00

Total 3,857 100.00

And in the appendix, table b1(page 209) of TDHS main report(the file size was too large so i gave the link to the file)

Table B1 list of strata by region, Nuts 1 region, residence, type and province, turkey 2008.

Stratum Region Nuts 1 region Type Province
1 west IStanbul urban/metropol istanbul
2 west istanbul rural istanbul
. ...
15 Cental west urban/metropol .....
anatolia
....
....
32 east Central east .....
anatolia
// /////
36 east South east anatolia ////////////////////////////

My question is whether or not V024 is strata in this case? and Psu is V001????

2) secondly, in my reasearch question, i will put the regions as independent variable. does this means FOR regional breakdown(strata) i need to use 12 sub regions. BECAUSE I WILL USE NUTS REGIONS, WHICH ARE MORE THAN 5 (V024 HAS ONLY 5 REGION AS GIVEN) So in this case, May i still use V024 as strata? or i need another variable?

tab V024

Region Freq. Percent Cum.

West 651 16.88 16.88
South 497 12.89 29.76
Central 666 17.27 47.03
North 352 9.13 56.16
East 1,691 43.84 100.00

Total 3,857 100.00

-For instance i will use the following variable as INDEPENEDENT VARIABLE SHOWING regional breakdown(the variable name is SREGION12). Assuming V024 is strata, Do i need a different starata for 12 regions. Because in the I WILL ASK STATA TO CALCULATE CLUSTERED STANDARD ERRORS?

tab SREGION12

Region of residence
(12) Freq. Percent Cum.

Istanbul 192 4.98 4.98
West Marmara 119 3.09 8.06
Aegean 215 5.57 13.64
East Marmara 216 5.60 19.24
West Anatolia 271 7.03 26.26
Mediterranean 497 12.89 39.15
Central Anatolia 252 6.53 45.68
West Black Sea 245 6.35 52.04
East Black Sea 159 4.12 56.16
Northeast Anatolia 400 10.37 66.53
Central East Anatolia 468 12.13 78.66
Southeast Anatolia 823 21.34 100.00

Total 3,857 100.00

Subject: Re: declaring child survey in Stata
Posted by musti on Fri, 06 Mar 2015 23:20:19 GMT
View Forum Message <> Reply to Message

One more question, if i use Difference in Difference method, do i still need to declara data as survey data to stata or i need declare it as panel data?

Subject: Re: declaring child survey in Stata
Posted by Trevor-DHS on Tue, 17 Mar 2015 01:27:26 GMT
View Forum Message <> Reply to Message

I'll respond to the 3 posts in order. First, though, are you using Stata or SPSS? The first code examples were in Stata, but the data files you sent were in SPSS.

First message

1a)
I don't understand how the 2003 data are not affected if the policy change took place in 1997.

You don't need to use year variables - you just need a variable that differentiates between the two surveys, so you can use your created variable v0 (phase). From this you can easily create your dummy variables.
gen d2003 = (v0==3)
gen d2008 = (v0==4)

1b) I don't understand your comment about VOO1 and V001. I only see v001 in your dataset. You can use the code I gave above but with v0 instead of year to create your new psu and new stratum variables.
egen newpsu = group(v0 v001)
egen newstr = group(v0 v023)
these can be done before or after pooling the data, but you need v023 in the dataset.

1d) I don't know what you mean by pseudo panel level, but I think your weight variable is ok after dividing by 1000000.

4) v022 provides a pairing or grouping of PSUs known as implicit strata that used to be for the calculation of sampling errors. We no longer recommend that approach, but rather to use the explicit strata that were defined for the survey and are found in v023.

5) You can do the pooling of datasets just as easily in Stata, using the append command.

Second message:

1) After looking at thew report, there are in fact 40 strata in the 2003 survey (see appendix B of the 2003 report), and I believe 36 in 2008 (I can't access the report due to a slow connection where I am currently). For the 2003 survey you can recode v0, sprovin and v025 to produce a variable with 40 categories that matches the strata given on page 169 of the 2003 report. Do something similar to produce the strata used for the 2008 survey.

Alternatively, you can use a more approximate definition of strata and just use v023. For the 2008 data you can create v023 as follows:
egen newv023 = group(v024 v025)
check that the coding of the resulting variable matches the codes used for 2003. You would then create a variable that separates these by survey using v0 as described earlier.
(I don't recommend this, but it probably won't make much difference in your significance test results).

2) You can include whichever region variables you wish to as independent variables. The variables used as strata and the variables used as independent variables do not have to match. See 1) just above about strata - it is not v024, but the 40 strata shown on page 189 (for 2003).

Third message:

DHS data are not panel data - the respondents, households, and clusters are not the same from one survey to the next - so I would not be declaring the data as panel data. You should be using the svyset and svy commands in your analysis.

Subject: Re: declaring child survey in Stata
Posted by musti on Wed, 01 Apr 2015 12:25:02 GMT
View Forum Message <> Reply to Message

Thank you very much.

One last question if you do not mind,
I have the results below from svyset decleration.

what should be the Method for variance estimation? I need to cluster standard errors according to region of birth(we have 12 regions).

svyset newpsu [pweight=v005], strata(strata) vce(linearized) singleunit(missing)

pweight: v005
VCE: linearized
Single unit: missing
Strata 1: strata
SU 1: newpsu
FPC 1: <zero>

Regards

Subject: Re: declaring child survey in Stata
Posted by Trevor-DHS on Thu, 02 Apr 2015 15:19:23 GMT
View Forum Message <> Reply to Message

That looks fine. The svyset command defines the sampling strata, not your domains of interest. How you produce your results is now up to you. Now use the svy commands with the results disaggregated by the 12 regions.

Subject: Re: declaring child survey in Stata
Posted by Reduced-For(u)m on Thu, 02 Apr 2015 17:07:55 GMT
View Forum Message <> Reply to Message

I think that musti wants to cluster standard errors from a single regression to get consistent standard error estimates and the standard in Diff-in-Diff is to cluster at the regional level at which

your exposure/treatment variable is defined.

In that case, one way to do it would be to change your svyset command to cluster at the region instead of the PSU. That said, you can't usually get consistent SE estimates from only 12 regions (you need like 30 or 40+ clusters for those to work). The usual way is some sort of fancy cluster bootstrap (Wild-t or something like that) - see Cameron, Gelbach and Miller "Bootstrap Based Improvements for Inference with Clustered Errors"

https://ideas.repec.org/a/tpr/restat/v90y2008i3p414-427.html

That is a bit technical, but one thing you could do is just set your svyset with region replacing PSU, and then use the T_10 (12 regions - 2) distribution for critical values. That is - you would need a t-stat of 1.812 for 90% confidence or 2.228 for 95% (two sided).

http://bcs.whfreeman.com/ips6e/content/cat_050/ips6e_table-d .pdf

---

Subject: Re: declaring child survey in Stata
Posted by habt_lancs on Sun, 08 Oct 2017 16:12:18 GMT
View Forum Message <> Reply to Message

Hello,

I have also the same problem.

I am working on Ghana DHS

And pooling the 1993, 1998, and 2003 surveys.
And i have to cluster at region level.

so, my question is do i still have to worry about weighting?

If so, should i de-normalise the weights.

In that case how can i do de-normalisation?

Can you also elaborate how the regression of this sort of analysis would go in stata?

Best,

Subject: Re: declaring child survey in Stata
Posted by Reduced-For(u)m on Sun, 08 Oct 2017 20:52:21 GMT
View Forum Message <> Reply to Message

Yes, you still have to worry about weighting, in the sense that if you want population level estimates of parameters (means/levels/distributions), and/or if you want to compare values/levels/coefficients from survey round to round (though with some caveats depending on what coefficient you might be interested in).

If the sample sizes are similar from survey round to round, you can get away with not adjusting the weights, but in general an easy way to deal with it is to add up the total sum of weights for each survey round and divide each individual weight by the sum of that survey's weights. The problem is just that the weights add up to something like the sample size, so if sample sizes change a lot you could end up weighting one survey a lot and another not very much. Maybe that makes sense (an observation is an observation) but it doesn't make sense in many contexts.

You will need to create new cluster-ID variables by, say, taking on a survey round identifier of some sort (so if the cluster variable value is 37 in the data, make it 199837 for cluster 37 in 1998, and 200337 for the 2003 cluster with the value of 37 (or whatever, that is just an example).

I have no idea what you are trying to do, so it is very hard for me to give you specific Stata advice, but I may be able to give some sort of guidance if you gave me more detail on what you were trying to accomplish.

Subject: Re: declaring child survey in Stata
Posted by habt_lancs on Sun, 08 Oct 2017 21:52:56 GMT
View Forum Message <> Reply to Message

Thanks so much. I understand the weighting and de-normalization process.

On the last point, I am working on a diff-in-diff estimator where the interest variable is measured at region level. So, i understand i have to cluster the standard errors by region level as a result of the specification. In the estimation i use regxfe as i include number of fixed effects.

But the sampling design also required me to cluster at the psu level, where i follow:
*tell Stata the weight (using pweights for robust standard errors), cluster (psu), and strata:
svyset [pweight=weight], psu(v021) strata(strata)

So my point is should i cluster only at region level, if so should i still define the svyset?

Or, Should i cluster both by psu and region ? I guess i may have to follow C.G.M method in this case.

And my last question is when you use commands like regxfe, does the svyset technique still the same?

Best,

Subject: Re: declaring child survey in Stata
Posted by Reduced-For(u)m on Mon, 09 Oct 2017 01:11:40 GMT
View Forum Message <> Reply to Message

You should just cluster by region. First you should make sure all the regions are the same in all 3 rounds...sometimes they change, and you'd have to adjust for that (just because two regions are "region 3" in two different surveys doesn't mean they are exactly the same region). Since no PSU spans two regions, you are implicitly clustering on PSU (you only have to cluster on the "higher" level). So set your clustering to region (with the caveat above about matching regions carefully), you don't need CGM stuff... (at least not for multi-way clustering....maybe for small number of clusters if you have less than, say, 40ish regions).

If the "svy:" prefix works, then it should work right. But I tend to use "xtreg" without the "svy:" prefix and set the clustering level and weighting myself. I haven't used the regxfe, but you should be able to set it all in the command itself you want (as options, not using 'svy').

Subject: Re: declaring child survey in Stata
Posted by habt_lancs on Mon, 09 Oct 2017 10:54:06 GMT
View Forum Message <> Reply to Message

Thank You So Much.

Subject: Re: declaring child survey in Stata
Posted by habt_lancs on Mon, 09 Oct 2017 23:05:27 GMT
View Forum Message <> Reply to Message

Thanks again.

But i am still a little unclear about one thing.

So when we use svy, we are not only implementing weighting but also taking in to account the sampling design/ the stratification.

But, as you suggest let us say i manually implement the regression using for example xtreg or regxfe by applying the new weight (constructed to take in to account the three round surveys i used) and clustering at region level.

should i still be concerned about the sampling design/ the stratification?

Best,

Subject: Re: declaring child survey in Stata
Posted by Reduced-For(u)m on Wed, 11 Oct 2017 17:26:57 GMT
View Forum Message <> Reply to Message

"should i still be concerned about the sampling design/ the stratification?"...

If you cluster and weight using xtreg/regxfe you ARE accounting for sampling design. That is all the "svy" prefix is doing too. You need to account for non-independence of observations (clustering and stratification) and non-randomness of cluster sampling (weighting). The "svy" command is just one way to tell the regression to cluster and weight - it is like putting in the options after comma in a regression code, it just lets you set that up once and then use the "svy" prefix instead of writing the code options directly into regression command.

But basically, they are doing exactly the same thing mathematically, there are just two ways to tell Stata to do that thing (with some very small caveats about the particular algorithms each version calls but which doesn't really matter much here).

Also just to be clear: "So when we use svy, we are not only implementing weighting but also taking in to account the sampling design/ the stratification."... that is true if you have included both the weighting and stratification/PSU information in the "svyset" command. You have to tell the "svy" prefix what to do first, but assuming you did that as recommended here, then yes, it covers both aspects of survey design corrections (the point estimate problem fixed with weighting; and the SE/p-val problem with stratification and clustering).

Subject: Re: declaring child survey in Stata
Posted by habt_lancs on Wed, 11 Oct 2017 20:56:42 GMT
View Forum Message <> Reply to Message

Thanks so much Once again.

Very helpful.

Subject: Re: declaring child survey in Stata
Posted by habt_lancs on Mon, 11 Dec 2017 12:57:48 GMT
View Forum Message <> Reply to Message

Continuing our discussion here, there are two things still bugging me:

So i want to extend my analytical sample by including the 1988 GDHS. So now i am pooling GDHS 1988, 1993, 1998, and 2003.

one problem with this is, while the rest of the surveys report the 10 regions in Ghana separately, the 1988 survey combines the 3 regions (upper west, east and northern) together and code them as region 8. As i told you i use region fixed effect in my regression and also standard errors clustered at region level. so my question is do you think treating three regions as one in 1988 and separately in the rest of the surveys create problems? or can i combine the three regions in the rest of the surveys to create consistency?

best,

Subject: Re: declaring child survey in Stata
Posted by Reduced-For(u)m on Mon, 11 Dec 2017 21:46:36 GMT
View Forum Message <> Reply to Message

In the past I have tried to regularize all my regions across survey rounds to have the most regions that are comparable... in this case I think (based on what you said) that that would imply merging the three regions in the later survey into one. That is fine - the definitions of regions are somewhat arbitrary anyway.

The only other thing you could do is define your own regions using the GPS data if it is available for all rounds, but I think in most cases that is overkill (and you have some small problems with the GPS displacement possibly messing with borders.

I'd just combine the regions in the 1988 survey and be good to go. You can always drop that survey as a robustness check to see if you get similar results, but it shouldn't really change much.

Subject: Re: declaring child survey in Stata
Posted by habt_lancs on Mon, 11 Dec 2017 21:52:21 GMT
View Forum Message <> Reply to Message

Thanks loads.

best,

Subject: Re: declaring child survey in Stata

Posted by habt_lancs on Tue, 12 Dec 2017 23:54:51 GMT
View Forum Message <> Reply to Message

Hello again,

I am also using the household members recode by pooling the surveys conducted in different periods. And in this case i have to de-normalilize the weight by using the following procedure :

HV005*=HV005x(total number of residential households in the country at the time of the survey)/(total number of households interviewed in the survey)

Unlike, the population of female 15 -49 data that can be obtained from UN, i do not know how to get data on "total number of residential households in the country at the time of the survey".

any help please...

Best,

Subject: Re: declaring child survey in Stata
Posted by Trevor-DHS on Mon, 18 Dec 2017 22:47:48 GMT
View Forum Message <> Reply to Message

Availability of data on numbers of households is limited, but you can find some sources available. For example, you could search google for "number of households by country". Wikipedia has a page of estimates at https://en.wikipedia.org/wiki/List_of_countries_by_number_of _households. The UN also has some estimates at http://data.un.org/Data.aspx?d=POP&f=tableCode:50. I believe the US census bureau has estimates for each country too, although at the time of writing I could not find them quickly. None of these data are great or provide data specifically for the survey years.

Another option is to use an approximation. Instead of using (households in the country)/(households in the survey), you could use an approximation of (women in the country)/(women in the survey). These two ratios should be very similar, and this might avoid working with the messier data about households.
