Understanding Society User Support - Support #357

cluster variable

04/11/2015 02:49 PM - Elisa Macchi

Status:	Closed	**Start date:**	04/11/2015
Priority:	Normal	**% Done:**	100%
Assignee:	Elisa Macchi		
Category:	Data analysis		

Description

I am working on the BHPS dataset. My aim is to check whether the change in the Long Term Care policies in 2002 had an effect on precautionary savings at an individual level.
To do so, I need to use the BHPS as a panel: I saw that there are longitudinal weights that are to be used in order to make longitudinal studies on the data. In addition to that, do you think I should use also cluster and strata variables when declaring the dataset? I understood the dataset is has a clustered design but as far as I saw, such cluster (PSU) and strata variables are contained in the HHsamp but not in the Indresp. So I thought that maybe, if the analysis is done at an individual level the longitudinal weights are already taking everything into account.

kind regards,
Elisa Macchi

History

#1 - 04/13/2015 09:06 AM - Redmine Admin

- Status changed from New to In Progress

- % Done changed from 0 to 50

The weights are needed for the point estimates, while the PSU and strata are needed for the confidence intervals. More on this can be found in the Understanding Society user guides and training course materials.
https://www.understandingsociety.ac.uk/documentation/mainstage
https://www.understandingsociety.ac.uk/2015/03/12/stata-training-course-online
On behalf of the team,
Jakob

#2 - 04/13/2015 10:50 AM - Alita Nandi

Most statistical packages assume that the data is a simple random sample. But if it is not, as in the case of the BHPS, that needs to be specified in order to correctly estimate the standard errors (and hence confidence interval). As these are household level variables these are provided in the whhsamp file. But these should be used in individual level analysis as well.

So, to summarize, you should merge the wpsu and wsrtata variables from whhsamp file into the windresp file and then specify this information in your analysis (e.g., in the svyset statement if using Stata).

Depending on you sample, you may want to additionally consider that individuals are clustered within households.

#3 - 04/14/2015 08:45 AM - Elisa Macchi

Dear Alita,

Thank you for your answer.

It is not totally clear to me how this procedure of weighting using longitudinal data will work in case I would use a subsample of the data. Would it still work?
The meaning of the longitudinal weights is to make "as if" there were no attrition?

In addition to that, as you mentioned, it seemed to me that there have been 3 stages in the selection of the final respondents set. Yet, I can only find one "strata" variable. Which are the others?

thank you in advance,
Elisa

#4 - 04/15/2015 11:20 AM - Alita Nandi

If in the documentation it has been specified that a specific weight is for the Original "Essex" Sample - it means you must use the Original "Essex" Sample or its sub-samples (such as only men, only 16-59 year olds etc) only. You cannot use the Extension/Boost samples with this weight. See Table 25 of the Volume A User Guide to see which weights are to be used with which sample.

"The meaning of the longitudinal weights is to make "as if" there were no attrition?"
Yes, basically it makes the estimates representative of the population who were alive and residing in UK during the sample period.

"In addition to that, as you mentioned, it seemed to me that there have been 3 stages in the selection of the final respondents set. Yet, I can only find one "strata" variable. Which are the others?"

First Postcode sectors were selected, then from those approx 33 addersses were selected, within these upto 3 dwelling units were randomly selected (if there was more than 1 dwelling unit -rarely) and from each dwelling unit upto 3 households were selected (if there was more than 1 HH - rarely). So, using the wpsu represents this clustering. If you are doing individual level analysis, in households with more than one person, there will be additional clustering - the household identifier represents the clustering at HH level.

The wstrata variable represents stratification not clustering.

Alita

#5 - 04/16/2015 10:22 AM - Elisa Macchi

Dear Alita,

I am using the entire sample (with all the boosting) and I guess I will consider a subsample (age 40+). Since the analysis is longitudinal, I guess I should use longitudinal weights: I had a look at table 25 and it is not clear to me what does "wLRWTUK1 from latest wave in longitudinal sequence" means. In addition to that, I noticed that several of the longitudial weight have value=0, is this ok?

This is the svyset command I am using:

vyset psu [pweight=xrwtuk1], strata (strata)

Is this ok, or since I am doing an analysis at individual level, should I specify "hid" like this?

svyset psu [pweight=longitudinalweights], strata (strata) || hid

as a result I got this:

. svyset psu [pweight=finlwght], strata (strata) || hid
Note: stage 1 is sampled with replacement; all further stages will be ignored

```
pweight: finlwght
            VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: psu
        FPC 1: &lt;zero&gt;
```

thank you for your help.

Elisa

#6 - 04/16/2015 01:12 PM - Elisa Macchi

In addition, my variable of interest is "save": sometimes the variable takes value -7 since this specific obs was a proxy and the question was not answered.
My first attempt was to drop all the save==-7 observation. Now I am wondering if the longitudinal pweight would take this into consideration: I mean, I could simply leave the save==-7 in the sample and applying the longitudinal weight they would eliminate this values.

Is this true?

Kind regards, Elisa

#7 - 04/21/2015 02:02 PM - Alita Nandi

Response to #5

If you are using 40+ year olds of all the 4 samples taken together then use wLRWTUK1. If the last wave of data you are using is wave 18, then the weight you should use is rLRWTUK1. But note that this longitudinal weight, rLRWTUK1, is non-zero for all those who responded continuously from the wave 11 to 18, zero otherwise. So, anyone who did not respond between these waves even once will have a zero weight. This is also zero for proxy and telephone respondents. Yes, if you left proxy and telephone respondents in (save=-7) then they will be effectively "dropped" from the analysis as their weight=0. But it may be better to drop them from your sample before you start the analysis so as to not get sample descriptives of the wrong sample. See section V of the Vol. A User guide.

hid is not unique across waves and so cannot be used in its current form in longitudinal analysis.

More on "Analyzing Correlated (Clustered) Data" http://www.ats.ucla.edu/stat/stata/library/cpsu.htm

#8 - 04/29/2015 11:48 AM - Elisa Macchi

Dear Alita, reading the volume A of the BHPS I have noticed that the sample for the BHPS was obtained through a 3 stages selection procedure. Yet, I am aware of just one sampling unit variable (PSU) which I believe to be the stage 1 sampling unit.

So, as I wrote, I declared to be data to be a survey in this way:

svyset psu [pweight=finxwght], strata (strata)
Note: stage 1 is sampled with replacement; all further stages will be ignored

pweight: finlwght
VCE: linearized
Single unit: missing
Strata 1: strata
SU 1: psu
FPC 1: <zero>

This works in principle but if I run regressions using svy, stata does not compute standard errors and gives this warning "Note: missing standard error because of stratum with single sampling unit". I am not fully sure but I guess that the problem comes from the fact that I could not specify the fact that the survey was a 3 stage design.

So, my question is: where can I find the stage 2 and stage 3 sampling units? And are there any variables for stages finite population correction?

kind regards,
Elisa

#9 - 05/12/2015 12:18 AM - Alita Nandi

"Note: missing standard error because of stratum with single sampling unit". This happened because you included Northern Ireland boost sample. This is the only BHPS sample to have a simple random sample design. So, for all those in this sample psu=-8 and strata=-8. Stata interprets this as a stratum with a single PSU and hence cannot compute se.

Option 1: replace psu=hid to trick stata into thinking that there are many psu in one NI stratum
Option 2: use the singleunit option of svyset

#10 - 05/22/2015 12:29 PM - Redmine Admin

- Status changed from In Progress to Closed

- % Done changed from 50 to 100