Understanding Society User Support - Support #561

Survey design and longitudinal weight choice in unbalanced panels

05/10/2016 03:48 PM - Dan Brown

Status:	Closed	**Start date:**	05/10/2016
Priority:	High	**% Done:**	100%
Assignee:	Dan Brown		
Category:	Weights		

Description

Hi all!

My questions are in the attached pdf file. Thanks very much for your help!

Cheers,
Dan

History

#1 - 05/10/2016 04:18 PM - Victoria Nolan

- Status changed from New to In Progress

- Assignee changed from Olena Kaminska to Dan Brown

- % Done changed from 0 to 80

Dear Dan,

Many thanks for your enquiry. In the first instance, please have a look at the worksheets on weighting in our online training course: https://www.understandingsociety.ac.uk/documentation/training/online. This will hopefully answer most of your queries.

It would also be worth looking through previous (closed) issues here on the user support forum to see if you queries have been answered in the past.

Just for future reference, please could you post your queries in the actual body of the message post rather than in a pdf attachment - this will enable other users to search for questions and answers in the forum, and also enables us to build up a better picture of frequently asked questions.

Many thanks,

Victoria

On behalf of the Understanding Society data user support team

#2 - 05/11/2016 05:49 PM - Dan Brown

Dear Victoria,

Thanks for your reply, and sorry for posting as an attachment! I have read the worksheets on weighting and sample design in the tutorial (e.g. 'Topic 8'), as well as the user guide, which were both very helpful in understanding the intuition behind the sample design. I've read the references given in the tutorial for these issues: Pitblado (2009), Wooldridge et al. (2013), and the relevant Stata help manuals, and I've been through the previous questions on the user forum under the headings 'survey design' and 'weights'. However, I am still unsure about the following questions. If I have missed the answers to these in the existing documentation/user forums, do please let me know where!

1. xtreg does not allow the use of svy in stata, so how should we take into account the sample design when running fixed effects regressions? So far I have used pweights and clustered my standard errors at the level of the psu. This deals with the different probabilities of selection into the sample (through the sampling weights) and the clustered sample design (by clustering standard errors at the level of the primary sampling unit), but it does not account for the fact that the sample design is stratified. How could I do that in these fixed effects regressions in Stata?

2. More generally, to check my understanding of the svy commands: suppose I am instead undertaking cross-sectional analysis, and the sample design is clustered but not stratified (unlike the UKHLS). If I use pweights and cluster standard errors at the level of the primary sampling unit, is this equivalent to first using svyset with that psu and sampling weight and then using svy commands?

3. The number of strata mentioned in the user guide seems to be fewer than the number in the dataset, or perhaps I have misunderstood. The user guide mentions 12 strata for England, Scotland and Wales, 1 for Northern Ireland and 4 in the EMB sample. These strata are split into sub-strata, but it sounds as though there are only approximately 100 or so sub-strata (e.g. 108 in the E/S/W sample). Yet in wave 1, for example, there are 1,776 strata (i.e. different values for the variable a_strata). What is the reason for this difference?

4. I am also estimating some descriptive statistics for a small sub-sample of the dataset in a single wave (approximately 1,300 individuals). I have dropped all the individuals who are not in my sub-sample of interest from the dataset, as it is considerably easier to set the data up after dropping observations that are not of interest (e.g. the size of the dataset is much smaller and so stata commands work faster). However, this means that I

cannot use the subpop option when running svy commands. It sounds as though using svy, subpop() to estimate descriptive statistics will produce different standard error estimates than if I drop the observations that are not in the sub-population of interest and just use svy. Pitblado (2009) presents a formula that demonstrates this difference, but I am not clear as to the intuition for why this difference exists?

5. As in question 4 above, I am calculating descriptive statistics from a small sub-sample of interest in a single wave (approximately 1,300 individuals). The vast majority of my strata then contain only a single primary sampling unit, and so stata cannot estimate standard errors when using svy commands. Now I don't think that any of the recommendations in the tutorial apply to this case. The ad hoc adjustments in the singleunit() option in stata do not apply – e.g. the observations are not certainty units. I can't just drop these cases from my sample of interest – as that is almost the entire sample of interest. Further, I cannot just merge strata with neighbouring strata that have multiple primary sampling units, as almost all strata have only a single primary sampling unit. So I am not clear what I can do to compute appropriately weighted descriptive statistics (e.g. means of particular variables) in this sub-sample?

6. For my fixed effects longitudinal analysis covering waves 1-5, I wish to include individuals who have left the sample in later years. I will therefore have an unbalanced panel. I understand that the longitudinal weights calculated in the UKHLS are only appropriate for balanced panel datasets. Suppose I have a way of dealing with attrition from my dataset in my model. Then I have been advised to use the weight from the first available wave in my dataset. Should this be the cross-sectional weight from wave 1 (which seems to be what you suggest in support number 414), and not the longitudinal weight from the first available wave (wave 2)? If I use the cross-sectional weight from wave 1, does this mean that I have accounted for differences in probability of selection at the start of the sample period, but have not account for subsequent attrition from the sample? And in terms of the population that I am claiming to 'represent', if I use the cross-sectional weight from wave 1, then should I interpret my estimates as representative of the UK population in that first wave only?

Sorry this is so long!

Cheers,
Dan

#3 - 05/12/2016 09:39 AM - Victoria Nolan

- Assignee changed from Dan Brown to Olena Kaminska

- Private changed from Yes to No

Dear Dan,

Many thanks for your follow-up. My colleague Olena, who is responsible for weighting, is currently on leave, but she is back in the office next week. I will assign this to her, and bring it to her attention on Monday so she can respond as soon as possible.

Best wishes, Victoria.

#4 - 05/23/2016 04:58 PM - Olena Kaminska

Dan,

Thanks for your questions.
To start with the issues with weights and strata+psu are separate and are not related.
The issue with having one psu per strata once you go to small subsample of UKHLS is known. One can deal with it manually (i.e. you would need to combine adjacent strata), but there is a syntax provided in the training manual that deals with this issue. This will give you an easy solution. Please let me know if you need help finding it.
The strata are indeed more detailed than most survey have - this provides benefits for precision. The strata takes into account not only reagions but also socio-demographic variables - thus improving precision most effectively. The number of original strata should be found on a_hhsamp.dta and will be higher than in any other files (especially later waves) due to non-response.
To answer the question about weights if your model can correct for attrition you should use longitudinal weight for the first wave of your analysis. But if you are using wave 1 (which does not have lw weight) you should use equivalent cross-sectional weight.
Your understanding of zero weights is correct.
If you want to use xt reg indeed it does not incorporate svy command. Basically you will have at least two levels of clusters in this situation. You should either use higher level clustering (as a conservative measure) or use a software such as MlWin which will allow you to take the full design into account correctly. The weight should be fine with xt reg. You can skip stratification - your results will be more conservative.
To understand how svy calculates standard errors please refer to Stata help - it has full formulae there.

Please let us know if we can be of further help,
Olena

#5 - 05/23/2016 05:48 PM - Alita Nandi

- Assignee changed from Olena Kaminska to Dan Brown

- % Done changed from 80 to 90

Hello Dan,

The training manual that Olena has asked you to look at to deal with single unit PSUs, is Example 6 of the online training course. To register for the course and download the training materials see https://www.understandingsociety.ac.uk/documentation/training/online/introduction-course

Best wishes,
Alita

#6 - 05/23/2016 07:19 PM - Dan Brown

Hi Olena,

Thanks for your reply.

I have two follow up questions:

1) You say 'basically you will have at least two levels of clusters in this situation'. Can I check what you mean by the 'two' levels? Do you mean the primary sampling unit and the household? I.e. you first cluster the design at the level of the postcode (you pick postcodes from which to sample households - these postcodes are the primary sampling units), and then you cluster at the level of the household (you pick households from which to sample individuals)? I want to check that I have not missed a more aggregate level of clustering than the primary sampling unit in your sampling design?

Obviously my error term (in my regression analysis) could be correlated across observations within a more highly aggregated region than the primary sampling unit, but I want to check that in terms of the sampling design itself the most aggregate level of clustering is the primary sampling unit? (I was just a little confused by your reference to 'two').

2) I have read Example 6 about the single PSU strata, but as I mentioned I am not 100% sure that this applies to my case for the following reason: The solution given in Appendix B suggests merging adjacent strata, which works well because there are not many single PSU strata. In the example in Appendix B, the first iteration of this process does not solve the problem, which I believe is because two single PSU strata were adjacent - and so, as outlined in that example, we need to do another iteration before all single PSU strata have been matched with a **multiple** PSU strata. The problem is that I am looking at a very specific sub-sample of individuals (about 1,200 individuals in a single wave), and about 80% of the strata are single PSU strata (I use svydes as described in the example, and the vast majority of strata have only 1 unit). So then I would need many iterations (perhaps 20 or so eyeballing the data) until all single PSU strata have been matched with a multiple PSU strata. Is that acceptable? Or would it be more sensible to pair adjacent single PSU strata?

Thanks again for your help,
Dan

Olena Kaminska wrote:

> Dan,
>
> Thanks for your questions.
> To start with the issues with weights and strata+psu are separate and are not related.
> The issue with having one psu per strata once you go to small subsample of UKHLS is known. One can deal with it manually (i.e. you would need to combine adjacent strata), but there is a syntax provided in the training manual that deals with this issue. This will give you an easy solution. Please let me know if you need help finding it.
> The strata are indeed more detailed than most survey have - this provides benefits for precision. The strata takes into account not only reagions but also socio-demographic variables - thus improving precision most effectively. The number of original strata should be found on a_hhsamp.dta and will be higher than in any other files (especially later waves) due to non-response.
> To answer the question about weights if your model can correct for attrition you should use longitudinal weight for the first wave of your analysis. But if you are using wave 1 (which does not have lw weight) you should use equivalent cross-sectional weight.
> Your understanding of zero weights is correct.
> If you want to use xt reg indeed it does not incorporate svy command. Basically you will have at least two levels of clusters in this situation. You should either use higher level clustering (as a conservative measure) or use a software such as MlWin which will allow you to take the full design into account correctly. The weight should be fine with xt reg. You can skip stratification - your results will be more conservative.
> To understand how svy calculates standard errors please refer to Stata help - it has full formulae there.
>
> Please let us know if we can be of further help,
> Olena

#7 - 05/23/2016 07:20 PM - Dan Brown

And thanks Alita - have found and read that example!

Alita Nandi wrote:

> Hello Dan,
>
> The training manual that Olena has asked you to look at to deal with single unit PSUs, is Example 6 of the online training course. To register for the course and download the training materials see https://www.understandingsociety.ac.uk/documentation/training/online/introduction-course
>
> Best wishes,
> Alita

#8 - 05/24/2016 10:59 AM - Olena Kaminska

Dan,

1) xtreg assumes you have clustering / nesting (this does not include psu - some nesting related to analysis). If you don't - you probably should be

using a simpler version of regression. You should know which clustering it is - I won't know it as it depends on your analysis. From sampling perspective yes, use PSU.
2) Either way is good but automatic is probably simpler (and less prone to error). They are also likely to give very similar results.

Hope this helps,
Olena

#9 - 06/06/2016 09:50 AM - Victoria Nolan

- Status changed from In Progress to Closed

- % Done changed from 90 to 100

Files

Additional_questions_for_ISER.pdf	77.3 KB	05/10/2016	Dan Brown