Understanding Society User Support - Support #1472

Weighting youth analysis across waves 1–8

01/05/2021 02:33 PM - Marie Mueller

Status:	Resolved	**Start date:**	01/05/2021
Priority:	Normal	**% Done:**	100%
Assignee:			
Category:			

Description

Hello,

I will use youth data from waves 1 to 8 (not always from all 8 waves, depending on the analysis). I will also link data on household, mother, and neighbourhood. While I am planning to use data from multiple waves, my analysis will not necessarily be longitudinal. I am expecting a dramatic drop in N because I will only use data on children from Greater London. Therefore, I want to use all data available. While I will not run a growth model or similar, I will adjust for clustering in children (using a multilevel model). Some children will have 1 data point, some 2, some 3. Data on some children will come from the first waves, data on other children will come from the last waves. I wonder how to apply weights in this analysis. For example, if I use a cross-sectional youth weight from wave 8, this will not be available for respondents who were'youth' in earlier waves.

Thanks very much for any ideas/suggestions!

Best wishes,
Marie

History

#1 - 01/05/2021 02:35 PM - Alita Nandi

- Status changed from New to In Progress

- Assignee set to Olena Kaminska

- % Done changed from 0 to 10

- Private changed from Yes to No

Many thanks for your enquiry. The Understanding Society team is looking into it and we will get back to you as soon as we can.

We aim to respond to simple queries within 48 hours and more complex issues within 7 working days. While we will aim to keep to this response times due to the current coronavirus (COVID-19) related situation it may take us longer to respond.

Best wishes,
Understanding Society User Support Team

#2 - 01/11/2021 11:07 AM - Olena Kaminska

Marie,

Thank you for your question.
On weights: we do not have a specific weight for your analysis. Please do not use youth cross-sectional weight - this will be incorrect, because your analysis is technically longitudinal as you use information from different time points. Your suboptimal weight would be a longitudinal enumeration weight from the last wave of your analysis. Or you could use this as a base weight and additionally correct for the nonresponse between this weight and your analysis - thus creating a tailored weight for your analysis.
On multilevel modelling: please take into account PSU as the highest level of clustering, followed potentially by parent-child clusters, followed by an individual.

Hope this helps,
Olena

#3 - 01/13/2021 01:02 PM - Alita Nandi

- Status changed from In Progress to Feedback

#4 - 01/20/2021 05:34 PM - Marie Mueller

Hi Olena,

Thank you very much for your quick reply.

I was wondering if you could elaborate on what you mean by "enumaration weight from the last wave" or provide an example? Would such a weight be available also for respondents who were 'youth' in previous waves and therefore will not be available in wave 8 (at least not as 'youth')?

I have two additional questions about the clustering. First, I wanted to ask whether clustering in households (i.e. multiple children from one household) is automatically accounted for by PSU, strata, or weight? In other words, do I need to add clustering in households as a level? If yes, would that be possible with svyset? Second, what do you mean by 'parent-child clusters'?

Thank you very much in advance for your help!

Best wishes,
Marie

#5 - 01/20/2021 06:05 PM - Olena Kaminska

Marie,

If the last wave used in your analysis is wave 12 (i.e. you have at least one variable from wave 12 in your analysis) you would use psnen??_lw weight as your suboptimal weight. Enumeration weight is for all 0+ people who are present at a household at the time of household enumeration (response). So, some could be children, youth or adults (responding or not responding) but as long as their household responded they will have a positive enumeration weight.

On clustering: if using svyset just specifying PSU is sufficient - you don't need to specify lower level clusters. As you mentioned a multilevel model (as long as you don't mean just a regression with multiple predictors) - this type of model allows and requires specification of all clustering levels - hence my earlier advice.

Hope this helps,
Olena

#6 - 01/21/2021 09:48 AM - Marie Mueller

Hi Olena,

Thank you very much - this is very helpful!

Does this mean that the enumeration weight is the same for every member of a given household? So, the enumeration weight would basically account for attrition but not item or individual non-response, correct?

A follow-up question: Why do I not need to adjust for clustering in households when using svyset? How do I ensure that SEs are not impacted by non-independence of observations due to shared household?

I am still thinking about how to best build my model. As I will use information of multiple waves and, therefore, will have multiple observations per individual, a multilevel model may be the best option. However, I would need to include four levels: observations, individuals, households, neighbourhoods. This seems a bit excessive, especially because I only have a small number of observations (as my focus is on Greater London). Do you have any thoughts on this?

I do have a couple of questions about how to adjust for the complex sampling design in a multilevel model using mixed in Stata. 1) How do I account for stratification? Including the strata as a covariate is not ideal because there are so many. 2) I will use LSOAs as my neighbourhood variable. Can I "ignore" the PSUs as a level and simply use LSOAs as the neighbourhood level? 3) At what level would I place the enumeration weight? Observations, individuals, or households? Or would I even need to include different weights at different levels?

Thanks very much once again for all your help!

Marie

#7 - 01/21/2021 11:51 AM - Olena Kaminska

Marie,

The enumeration weight is the same for everyone in a household (with very few exceptions in EMB sample) only in wave 1. It is an individual weight - so it's meant to represent different types of individuals correctly - hence different values within a household. If you are interested in household-level variables we have a household weight - this one would be the same for everyone in a household.

About clustering with svyset: only the highest level of clustering is needed to be included for standard errors to be correct.

About thoughts on multilevel model: it completely depends on your substantive questions and what you are interested in. Indeed, one may ran out of degrees of freedom with small sample size - so a suboptimal option may be taken in that situation.

On other questions. 1) You can ignore stratification - your estimates will be conservative (so, still correct with slightly wider CIs than by design). 2) Use LSOA if they are higher level than PSU and if all PSUs are nicely nested within LSOAs. So double check this and use LSOA as your cluster variable. 3) Our enumeration weight is for individuals. But if your model resembles pooled analysis you may use different weights from each wave at an observation level.

Best wishes,
Olena

#8 - 01/21/2021 01:27 PM - Marie Mueller

Hi Olena,

Thank you very much for your detailed reply. This is super useful!

As for the weight, in a longitudinal design, I would apply the weight of the last time point to all time points, correct?

I will start trying out a few things now and may come back to you if something remains unclear.

Best wishes,
Marie

#9 - 01/21/2021 01:56 PM - Olena Kaminska

Yes, if the weight is at an individual level.

#10 - 01/25/2021 11:05 PM - Alita Nandi

- Assignee changed from Olena Kaminska to Marie Mueller

- % Done changed from 10 to 90

#11 - 01/26/2021 02:40 PM - Alita Nandi

- Assignee changed from Marie Mueller to Olena Kaminska

#12 - 01/27/2021 03:05 PM - Marie Mueller

Dear Olena,

I am now at the stage of trying to get a deeper understanding of what PSUs, strata, and weights actually are (i.e. what information they contain). I see that there is a detailed description of technical details of weights in the Mainstage User Guide, which I will use to better understand the enumeration weight. However, I have a question about PSUs and strata.

PSUs -> In Table 37 of the Mainstage User Guide, I can see that the level of geography of the PSUs depends on country and sample. Sometimes the PSU is the postal sector. In the UK, there are around 9,000 postal sectors. Sometimes the PSU is the LSOA. In the UK, there are around 35,000 LSOAs. Thus, the postal sector would be the higher level of clustering. Above you said I can use LSOAs as my highest level of geography if PSUs cluster in them. I'm not sure they always will, as the PSU seems to be the higher level. Nevertheless, as I am interested in area level variables at LSOA level and the LSOA can change over waves (other than the PSU), I assume that in a multilevel model I will always use LSOAs as my highest level of clustering (rather than PSUs). Do you have any thoughts on this?

Strata -> From Table 38 of the Mainstage User Guide, I find it difficult to understand what information exactly a stratum contains. For example, if strata correspond to groups of two or more PSUs, why is adjustment for clustering in PSUs not sufficient? In other words, what additional information does a stratum contain? What does it tell us?

Thanks very much in advance!

Marie

#13 - 01/28/2021 10:18 AM - Olena Kaminska

Marie,

Sample design is described here: https://www.understandingsociety.ac.uk/sites/default/files/downloads/working-papers/2009-01.pdf

Two things to note about clusters: there is no clustering in Northern Ireland. In GB PSUs are postal sectors (note they are postal sectors at the time of the sample selection - sometimes they change over time). Some postal sectors were combined if they were too small. Therefore in GB anything at a higher level than postal sector could be used as a PSUs. Any PSU higher than a household could be used in NI.

Stratification is also described in the paper. In UKHLS it is complex. But you don't need to understand it for clustering. A few years ago multilevel models could not take into account stratification - don't know if this has changed. Again, it is ok to ignore stratification - your estimates would be more conservative.

Hope this helps,
Olena

#14 - 02/01/2021 12:59 PM - Marie Mueller

Hi Olena,

Yes, thank you very much!

As for the clustering, I was wondering: when I am interested in neighbourhood effects at LSOA level and LSOAs may change over time (because households change address), I would want to include a level for LSOAs in my multilevel model. Now, LSOAs are not a higher geography than postal sectors (often the PSU). Therefore, I wonder if it is a problem not to adjust for PSUs as well. My multilevel model is quite complex already, so I would

rather not include PSUs as another level. What impact could this have on my results (e.g. on estimates and CIs)?

Best wishes,
Marie

#15 - 02/01/2021 01:11 PM - Olena Kaminska

Marie,

You do need to go with the highest level of clustering in the model (other levels are optional). But indeed things change over time and people also move. It is then a theoretical question which level is more appropriate to use - where people lived at the time of sampling or where people live now. I think it depends on your substantive topic, especially in a multilevel context (where you are specifically looking into clustering). So, it will be correct to use either current postal sector, or the old one. Or you can even use both in a more complex cross-classified multilevel model as both will have some effect, but this is too complex and I wouldn't suggest it for your analysis.
So, two points: always use the highest level, and you can use either modern or the old postal sectors / higher level - up to you.

Hope this helps,
Olena

#16 - 02/01/2021 01:21 PM - Olena Kaminska

Marie,

One more point. If you want to skip postal sectors and go with the lower LSOAs it is theoretically possible to have an estimate of the contribution of the higher level clustering to the standard error. What you need is to run a model with the postal sectors, LSOAs and individuals at the time of sampling (when they are nested) using variables from your model - in a multilevel context this should give you an estimate of the variance explained by postal sectors. If you can separate this you could calculate design effect due to it. This design effect can then be multiplied by your standard error in your final model - giving you a very good approximation of your correct CI in the final model.
Basically your CI in the final model needs to be a bit wider if you omit postal sector. Deff (design effect) would give you an estimate by how much. You may want to talk to multilevel experts on exactly how to estimate this deff (the interclass correlation will be part of it).
Hope this helps,
Olena

#17 - 02/01/2021 02:23 PM - Marie Mueller

Hi Olena,

Thank you so much for your detailed reply and suggestion for even more accurate analysis. I will probably start 'simple' and then see from there. Thank you once again for your rapid and thorough help.

Best wishes,
Marie

#18 - 03/02/2021 03:03 PM - Understanding Society User Support Team

- Status changed from Feedback to Resolved

- Assignee deleted (Olena Kaminska)

- % Done changed from 90 to 100