Understanding Society User Support - Support #1520

Pooling cross-sectional data of UKHLS - Waves 1 to 7

03/08/2021 11:02 AM - Marie Mueller

Status:	Resolved	**Start date:**	03/08/2021
Priority:	Normal	**% Done:**	100%
Assignee:			
Category:			

Description

Hello,

This issue relates to issue #1472.

Summary

I will use UKHLS youth data of Waves 1 to 7. My outcome of interest is the SDQ. I will use data on youth living in Greater London only. I am **not** interested in change in SDQ over time. My **main** goal is to retain as many observations as possible, as the number of observations drops dramatically due to my focus on Greater London. Therefore, I would like to use all the observations available across Waves 1 to 7. Depending on age at study entry, some participants complete the youth self-completion questionnaire only once, others twice, and a few even three times. Also, some participants complete the youth self-completion questionnaire in the earlier waves, others in the later waves. In other words, I have an unbalanced panel design in that some (but not all) participants have data at multiple time points. The **main** problem: finding an appropriate weight.

In #1472, we came to the conclusion that using a longitudinal person enumeration weight at the last time point would be the suboptimal weight to use (here: g_psnenus_lw). However, looking at the data, I find that with this approach I would lose too many observations. Of 2,056 individuals, 749 have a missing weight and 551 have a weight of zero, leaving me with only 756 individuals included in my analysis.

Ultimately, I want to use **all** observations of **all** 2,056 individuals. I assume the only solution would be to **pool cross-sectional information** of all seven waves, correct?

If yes, I did see point 12 in *Weighting FAQs*, however, I can't really follow. I also had a look at #1374 and #1257 which seem to be related. From these two I take that I do not have to worry about *scaling* the weights. I was wondering if you could give some advice about how to approach the pooled analysis.

Questions

1. How do I best pool my data? Do I run seven separate models (one for each wave using the according youth cross-sectional weight of that wave) and pool my estimates afterwards? Or do I combine data of all seven waves in one analysis, using a long format? From the issues linked above, I take that the latter is the right way. Then I would have seven rows for each individual and my long weight variable would contain the corresponding cross-sectional weights (i.e. row 1 for individual 1 would contain the cross-sectional weight of individual 1 at the first wave, row 2 for individual 1 would contain the cross-sectional weight of individual 1 at the second wave...). Is this correct?

2. If the above is correct (i.e. I combine all information of all individuals across all waves in one data set, transformed into long format), how do I then analyse these data? I was planning to run a mixed model to adjust for clustering in LSOAs, households, and individuals. (Note that I may omit the household level because I have too many levels for the relatively small number of observations.) However, I already anticipate the problem that an individual will have different weights at different time points, which will lead to an error in mixed and, I assume, svyset too. How do I go about this problem? My idea would be to include only one observation per individual which would lead to loss of data but would solve the problem of inconsistent weights within the individual (and would also allow me to omit the individual level from my multilevel structure). I guess then I would not need a long format after all because I would only have one observation per individual and I do not actually need to know what wave this was taken from. Is there a better way to do this and to actually keep all observations of all individuals?

Thank you very much in advance!

Best wishes,
Marie

History

#1 - 03/08/2021 06:40 PM - Understanding Society User Support Team

- Status changed from New to In Progress

- Assignee set to Olena Kaminska

- % Done changed from 0 to 10

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 - 03/08/2021 06:41 PM - Understanding Society User Support Team

- Private changed from Yes to No

#3 - 03/09/2021 10:22 AM - Olena Kaminska

Marie,

Thank you. Your problem is not about weights or even data structure but with the definition of the population that you want to study. Once you define your population - other questions will be easy to answer. I will give you an example.
Let's say you want to study 12 year olds. You can then pool all 12 year olds we have in UKHLS. This way you will have dataset comprising of 12 year olds from different waves. At each wave weighted 12 year olds represent the 12 year old population in that year. So you can have many 'samples' of these 12 year olds from different waves and you could just put them together and analyse them in one model.
The weight choice here is straightforward:
- if all the information comes from the concurrent wave when they are 12 - use cross-sectional weight;
- if some information comes from earlier (or later) waves than you have to use longitudinal weight corresponding to the last wave that the information comes from;
- scaling may be useful especially if you go back to BHPS data and use it together with UKHLS data - otherwise some years will have a large impact on your estimates (if you don't care though that you have 10 times more 12 year olds from 2000s than from 1990s then you don't need any scaling).
- the weight choice for each subsample of 12 year olds follows simple rules described in FAQs. To combine you just put these samples together and give the chosen weight to respondents.

Note, I would recommend to use one person once if you are interested in people as described in the example above. Pooling multiple observations per person is useful when you want to study change - i.e. events, not people.

Hope this helps,
Olena

#4 - 03/09/2021 03:36 PM - Marie Mueller

Dear Olena,

Thank you very much for your rapid and detailed reply.

I will use the UKHLS youth data. My population of interest is **10- to 15-year-olds.** Some individuals have multiple observations. This depends on their age at study entry, determining how often in Waves 1 to 7, they are categorised as 'youth' (i.e. are between 10 and 15 years old). Also, some individuals are 'youth' at earlier waves, others at later waves, again depending on their age at study entry. Due to my focus on Greater London, my sample is small to begin with. **So my main aim is to retain as many observations as possible in my model, ideally all observations of all individuals across Waves 1–7.**

Approach 1

My first approach was to use the longitudinal person enumeration weight of the last time point (here Wave 7). This resulted in a loss of too many observations – so I will not take this approach.

Approach 2

My second approach would be to pool the data of all waves. However, because those individuals with multiple observations will have different weights at different waves, I would not be able to just combine all the data and run a mixed model. You suggested I should only use one observation per individual. However, I was wondering how to select one out of multiple observations for one individual? **For example, if one individual was 'youth' at Waves 1, 2, and 3 (at ages 10, 12, and 14 years), which of these observations would I choose?**

Approach 3

A third approach would be to look at age groups 10, 11, 12, 13, 14, and 15 separately (i.e. run six separate models). Some individuals (those with multiple observations across waves) would then appear in more than one of these models. The sample size in each of these models would be quite small, which is why I would prefer to use all observations across all ages in one model. However, if running separate models was the best approach, then I would – for the age 12 analysis – combine the data of all 12-year-olds across waves into one dataset and each individual would be given the cross-sectional weight of the wave at which their observation was measured (i.e. the wave at which they were 12 years old), correct? **So it would be okay to mix cross-sectional weights of different waves in one model?**

Summary

To summarise, my population are 10- to 15-year-olds; some have multiple observations. I need to retain as many observations as possible in my model(s) because I am working with a very small sample due to my focus on Greater London. I am not interested in change over time but would be happy to use multiple observations per individual if that meant I could retain more observations.

So I guess my main question is: **What is the best approach to retain as many observations as possible in my model(s)?**

Thanks very much in advance!

Marie

#5 - 03/09/2021 03:54 PM - Olena Kaminska

Marie,

Your analysis as you describe is only legitimate if you study something time variant - something that changes with time. In this situation you indeed may have 5 different values per person over 5 waves (e.g. attitudes).
If you are studying something time invariant (e.g. eye colour) then this approach is statistically wrong and violates iid assumption.

Before I can answer weighting question you need to decide on what you study (i.e. your superpopulation) and let me know.

Thanks,
Olena

#6 - 03/09/2021 04:15 PM - Marie Mueller

Dear Olena,

I will try to describe my problem/issue in a different way:

- My study population is young adolescents – here 10- to 15-year olds.
- I am interested in all ages 10 to 15 years (but not in change over time).
- My outcomes were measured at Waves 1, 3, 5, and 7.
- I could select one wave and run a cross-sectional analysis using the cross-sectional youth weight.
- However, I would like to use data of all four waves 1, 3, 5, and 7 to increase my sample size.
- Some individuals have data at multiple waves, which seems to make pooling a bit more complicated.

How do I best approach these data (without losing too many observations)?

I hope this is a bit clearer?

Thanks very much!

Marie

#7 - 03/12/2021 11:41 AM - Marie Mueller

Dear Olena,

To make it (hopefully) even clearer:

- My outcomes are measured at waves 1, 3, 5, and 7
- In every wave, I have a snapshot of youth at the age of 10 to 15 years
- This means that I have four snapshots of interest for my analysis
- In theory, if every individual would only appear in one of these snapshots, I could pool the data of all four snapshots (using cross-sectional weights)
- However, this is not the case: some individuals appear in more than one snapshot
- Therefore, pooling data and using cross-sectional weights is not possible (because some individuals would have different weights at different waves)

I see three possible solutions for my problem and am wondering which one would be most valid (or if you could suggest a better, more valid approach that I am missing):

1. I could use data of **one** wave only (i.e., wave 1, 3, 5, **or** 7). The problem: the sample size would be (too) small.
2. I could use data of **all** four waves (i.e., wave 1, 3, 5, **and** 7). The problem: some individuals would appear more than once (as described above), and I would need to select only one observation per individual. I am not sure this is a valid approach/how to decide which observation to select.
3. I could use data of **all** four waves (i.e., wave 1, 3, 5, **and** 7) and run six separate models, one for each age 10, 11, 12, 13, 14, and 15 years. Then every individual would only appear once (if at all) in each model, and I could make statements about separate age groups. The problems: I would have a large number of models. Further, I didn't have initial hypotheses justifying this approach (i.e., looking at every age separately). Finally, the sample size in each model would probably be too small.

Note: In all three approaches, I would use a pooling approach (using cross-sectional weights).

I hope my problem is clearer now, but please let me know if not.

Thank you very much for all your help and apologies for so many (long) posts from my side.

Best wishes,
Marie

#8 - 03/15/2021 09:40 AM - Olena Kaminska

- Assignee changed from Olena Kaminska to Alita Nandi

#9 - 03/15/2021 12:23 PM - Understanding Society User Support Team

- % Done changed from 10 to 50

Hello Marie,

You are right (1) will result in small sample sizes and there are no avaialble weights to use with approach (2).

You could do (3), or estimate one model with data pooled from across all 6 waves: as individuals are likely to give different responses across waves as their life circumstances change (including age), you could pool the waves and estimate multivariate models where you control for different aspects of their lives that are changing including age. In these cases use cross-sectional weights. Here you are assuming that the individuals who appear more than once are different individuals where their past does not matter, or to the extent it does it is accounted for by the variables included. Alternatively, you could include information about their past explicitly in the model - in this case you should use longitudinal weights.

The remit of this forum is to advise you on aspects about the survey, the data and the appropriate weights to use for the analysis you would like to conduct. But the choice of analysis method is yours as it depends on your research question, what assumptions you are ok to make etc. Hope this helps.

#10 - 03/15/2021 12:23 PM - Understanding Society User Support Team

- Status changed from In Progress to Feedback

#11 - 03/15/2021 01:24 PM - Marie Mueller

Dear Alita,

Thank you very much for your quick and very helpful reply. Yes, of course, I understand that I need to make my own choices. I just wanted to make sure that I understand the data structure and the weights correctly and that I draw the right conclusions. Please accept my apologies if I did not make this clear in my previous posts.

Thank you very much for your suggestion for how to pool the data. If I understand correctly, you suggest that I could ignore the clustering of data in the individual (which would solve the problem that some individuals would have more than one weight). I did not consider this option before because I thought adjusting for dependence of observations was crucial for the accurate calculation of standard errors. I will consider this option now.

As for pooling in general, it sounds straightforward: prepare all the waves and combine them into long format where the seven (or rather four, as my outcomes are only measured at waves 1, 3, 5, and 7) cross-sectional youth weights will be combined into one weight variable (i.e. into one column). Then use these data in my mixed model (ignoring the individual level).

Thank you very much once again for your help!

Best wishes,
Marie

#12 - 03/15/2021 02:45 PM - Understanding Society User Support Team

- Assignee changed from Alita Nandi to Olena Kaminska

About clustering on individuals: If you pool data across waves of this data, you are right the error is not independently distributed as the same person may appear more than once. So, you should account for both the clustering due to design (primary sampling unit) and multiple observations for the same individual (psu & pidp). My understanding is that once you account for clustering of error due to PSU, the additional clustering of individuals is minimal and could be ignored. I will assign this to Olena for her views on this.

If you are using Stata use SVY suite of commands to tell Stata that the data is from a sample that is clustered and stratified: the variables for these are PSU STRATA in the XWAVEDAT file.

About data structuring - yes that is the correct method. Just remember that the names of the cross-sectional weights are different across waves as the samples included are different, so you will need to create a new weight variable which equals the youth cross-sectional weight variable from that wave.

#13 - 03/15/2021 03:58 PM - Marie Mueller

Thank you, Alita!

Note: As I am interested in neighbourhood influences measured at LSOA level, I thought a multilevel model would make sense to account for clustering in neighbourhoods (i.e. LSOAs, not PSUs). LSOAs may be more accurate in representing shared exposure / influences because LSOAs change over time (whereas PSUs don't). So, for example, at wave 7 LSOAs may reflect better the shared exposure of children living in the same LSOA than PSUs. However, if it would be accurate (- and I am not sure about this -) to only account for PSUs (and not LSOAs), svyset would be a good option.

#14 - 03/16/2021 12:23 PM - Olena Kaminska

Marie,

You need to account for the highest level of clustering as a minimum. Our sample is clustered within PSUs - so taking them into account is the easiest way. If all our PSUs are nested within LSOAs you can just take into account clustering within LSOAs. But if clustering is not perfect (if for example one PSU belongs to two LSOAs) the correct model to use would be a cross-classified multilevel model.

Hope this helps,
Olena

#15 - 03/16/2021 01:03 PM - Marie Mueller

Hi Olena,

Yes, thank you!

From our various exchanges, I conclude the following:

Using **svyset**, accounting for clustering in the highest level, i.e. the PSU, would be sufficient. Using svyset, I could therefore *ignore* clustering in *lower* levels, i.e. households (for those children that share a household) and individuals (for those children that have more than one observation).

Using a **multilevel model (mixed)**, I should account for all the levels available, i.e. neighbourhoods, households, and individuals (as far as possible). If clustering of PSUs in LSOAs is not perfect, I would need to run a cross-classified multilevel model.

In a **pooled** analysis, using svyset or mixed, I would simply combine data of all waves, combining the cross-sectional weights of all waves into one weight variable (i.e. one column).

Best wishes,
Marie

#16 - 03/16/2021 01:40 PM - Olena Kaminska

Marie,

Yes, this sounds correct.

Best of luck,
Olena

#17 - 03/31/2021 04:16 PM - Understanding Society User Support Team

- % Done changed from 50 to 90

#18 - 08/05/2021 01:56 PM - Understanding Society User Support Team

- Status changed from Feedback to Resolved

- Assignee deleted (Olena Kaminska)

- % Done changed from 90 to 100