# Understanding Society User Support - Support #1894

## Weight for unbalanced and merged UKHLS and BHPS data

04/21/2023 03:30 PM - Yanan Zhang

| | | | |
|---|---|---|---|
| **Status:** | Resolved | **Start date:** | 04/21/2023 |
| **Priority:** | Urgent | **% Done:** | 100% |
| **Assignee:** | Olena Kaminska | | |
| **Category:** | Weights | | |

**Description**

Dear Sir/Madam,

I hope this message finds you in good health and high spirits.

I am currently working with individual-level data from the merged Waves 1-18 of the BHPS and Waves 1-8 of the UKHLS datasets. I have a couple of questions regarding the use of weights in my analysis. I would appreciate any guidance you could provide.

1. In my study, I am employing fixed effects estimates to analyze the relationship between two variables, x and y. Given this approach, is it necessary to apply weights to the analysis?

2. I have followed the guidelines and used the longitudinal weight provided in Wave 8 of the UKHLS. However, I understand that this weight is applicable only to those who have participated in all waves. Since many individuals have only participated in parts of the waves, I am unsure how to generate weights for these participants. Could you please advise on the appropriate way to handle this situation?

Thanks for your time!

### History

**#1 - 04/23/2023 11:09 PM - Understanding Society User Support Team**

*- Status changed from New to In Progress*

*- Assignee set to Olena Kaminska*

*- % Done changed from 0 to 10*

*- Private changed from Yes to No*

Thank you for your email.

We aim to respond to simple queries within 48 hours and more complex issues within 7 working days.

We are keen to hear about any data issues and experiences that you have as this will help us build the best possible knowledge database for the UKHLS and BHPS data sets.

Best wishes,
Understanding Society User Support Team

**#2 - 04/24/2023 01:20 PM - Olena Kaminska**

Dear Yanan Zhang,

Thank you for your question. Yes, you always need to use weights with UKHLS, regardless of analysis. To answer your second question please read information below.

You can pool the data however you want. There are three most important points to keep in mind:
1. Always take into account clustering within PSUs with UKHLS data. Taking into account clustering within a person (in case you have multiple entries per person) is optional and could be used in addition to clustering within PSUs. This implies that you don't need to use multilevel models while pooling – you could use the standard svy command if this suits your purpose.
2. When pooling information from multiple waves, especially BHPS waves and UKHLS waves you need to apply additional scaling to weights in order for each wave to contribute a similar level as all others. See question 19 in this document for how to implement it.
3. Define your population carefully. Unlike unpooled analysis, where population definition is straightforward, we find that many users get confused with the population definition in the pooled analysis. A few examples follow presenting the population definition and the data structure:
- Events, e.g. hospitalization occurrences (staying in a hospital for over 24 hours) observed in GB between 1991 and 2009 and UK between 2009 and 2020. In this situation hospitalization variable would be created and data is pooled from all waves and all people observed in each wave between 1991 and 2020. Note, you are studying events, not people, in this situation.
- Event triggered situations, e.g. happiness upon marriage observed in UK between 2009 and 2020. If you study the state after marriage – you could pool all the observations after marriage in the data from all the time points. Your data will consists of all marriages and relevant observations following

from all waves between 2009 and 2020. You are studying happiness following marriage, i.e. a state following an event (not people).
- A subgroup defined by a time point, e.g. 11 year olds living in UK between 2009 and 2020. You could pool information from 11 year olds from each wave and analyse them together in one model, which gives you more statistical power. In this situation you will have one observation per person as a person is 11 only once per lifetime (and wave).
- A subgroup defined by an event where event may happen multiple times, e.g. first year students studying in UK between 2009 and 2020. You could pool first year students from all the years we have in the study. Note, some people may have multiple occurrences of being a first year student. It then depends on your definition. If you want to study number of books read in a year by the first year students it may be appropriate to count all the multiple occurrences per person. In this situation you don't study people really but 'event triggered states'.
- Time variant state or characteristic, e.g. wellbeing observed in UK between 2009 and 2020. While wellbeing changes over time and it may be more appropriate to study it using a classic longitudinal analysing, there are situations, especially when studying very small subgroups, where pooling may add statistical power. In essence you are studying wellbeing states observed over a specific time period, (again not people). For this you just pool all the information on wellbeing from all the relevant waves.
- It does not make sense to study time-invariant states (e.g. eye colour) with pooled analysis. If you happen to do it, your effective sample size will not be any higher than in an unpooled analysis. So, technically there won't be any gain from pooling, and it would be easier and clearer to avoid it.

Pooling can be cross-sectional or longitudinal. Theoretically, you will be combining 'separate samples of events / states' each of which will have the corresponding weight.
If you are just interested in events (and what happened at the same time / wave) you are looking at pooling cross-sectional information. For this create a new weights variable new_weight, and give it a value of the cross-sectional weight from each wave (e.g. new_weight=a_indinus_xw if wave==1; new_weight=b_indinub_xw if wave==2 etc.)
Alternatively you may be interested in what happens before and / or after a particular event (e.g. studying work pattern for 3 years before birth of a first child and 3 years after for new mothers). In this situation you need to choose a longitudinal weight from the last wave in your analysis for each combination of waves (e.g. for birth at wave 3 where we observe waves 1-6, the weight will be f_indinus_lw; for birth at wave 4 with information in the model of waves 2-7 –it will be g_indinub_lw etc.).

Hope this helps,
Olena

#### #3 - 04/24/2023 07:31 PM - Yanan Zhang

Olena Kaminska wrote in #note-2:

> Dear Yanan Zhang,
>
> Thank you for your question. Yes, you always need to use weights with UKHLS, regardless of analysis. To answer your second question please read information below.
>
> You can pool the data however you want. There are three most important points to keep in mind:
> 1. Always take into account clustering within PSUs with UKHLS data. Taking into account clustering within a person (in case you have multiple entries per person) is optional and could be used in addition to clustering within PSUs. This implies that you don't need to use multilevel models while pooling – you could use the standard svy command if this suits your purpose.
> 2. When pooling information from multiple waves, especially BHPS waves and UKHLS waves you need to apply additional scaling to weights in order for each wave to contribute a similar level as all others. See question 19 in this document for how to implement it.
> 3. Define your population carefully. Unlike unpooled analysis, where population definition is straightforward, we find that many users get confused with the population definition in the pooled analysis. A few examples follow presenting the population definition and the data structure:
> - Events, e.g. hospitalization occurrences (staying in a hospital for over 24 hours) observed in GB between 1991 and 2009 and UK between 2009 and 2020. In this situation hospitalization variable would be created and data is pooled from all waves and all people observed in each wave between 1991 and 2020. Note, you are studying events, not people, in this situation.
> - Event triggered situations, e.g. happiness upon marriage observed in UK between 2009 and 2020. If you study the state after marriage – you could pool all the observations after marriage in the data from all the time points. Your data will consists of all marriages and relevant observations following from all waves between 2009 and 2020. You are studying happiness following marriage, i.e. a state following an event (not people).
> - A subgroup defined by a time point, e.g. 11 year olds living in UK between 2009 and 2020. You could pool information from 11 year olds from each wave and analyse them together in one model, which gives you more statistical power. In this situation you will have one observation per person as a person is 11 only once per lifetime (and wave).
> - A subgroup defined by an event where event may happen multiple times, e.g. first year students studying in UK between 2009 and 2020. You could pool first year students from all the years we have in the study. Note, some people may have multiple occurrences of being a first year student. It then depends on your definition. If you want to study number of books read in a year by the first year students it may be appropriate to count all the multiple occurrences per person. In this situation you don't study people really but 'event triggered states'.
> - Time variant state or characteristic, e.g. wellbeing observed in UK between 2009 and 2020. While wellbeing changes over time and it may be more appropriate to study it using a classic longitudinal analysing, there are situations, especially when studying very small subgroups, where pooling may add statistical power. In essence you are studying wellbeing states observed over a specific time period, (again not people). For this you just pool all the information on wellbeing from all the relevant waves.
> - It does not make sense to study time-invariant states (e.g. eye colour) with pooled analysis. If you happen to do it, your effective sample size will not be any higher than in an unpooled analysis. So, technically there won't be any gain from pooling, and it would be easier and clearer to avoid it.
>
> Pooling can be cross-sectional or longitudinal. Theoretically, you will be combining 'separate samples of events / states' each of which will have the corresponding weight.
> If you are just interested in events (and what happened at the same time / wave) you are looking at pooling cross-sectional information. For this create a new weights variable new_weight, and give it a value of the cross-sectional weight from each wave (e.g. new_weight=a_indinus_xw if wave==1; new_weight=b_indinub_xw if wave==2 etc.)
> Alternatively you may be interested in what happens before and / or after a particular event (e.g. studying work pattern for 3 years before birth of a first child and 3 years after for new mothers). In this situation you need to choose a longitudinal weight from the last wave in your analysis for each combination of waves (e.g. for birth at wave 3 where we observe waves 1-6, the weight will be f_indinus_lw; for birth at wave 4 with

information in the model of waves 2-7 –it will be g_indinub_lw etc.).

Hope this helps,
Olena

Subject: Re: Query on Weight Creation for BHPS and UKHLS Data

Dear Olena,

Thank you so much for your reply. In your response, you mentioned 'question 19 in this document', but I am unable to locate the referenced document. Could you please kindly provide the document for my reference?

Regarding my analysis, I am interested in using all observations (individual level) across wave 1-18 BHPS and wave 1-8 UKHLS. The provided weight is designed for a balanced dataset. However, I would like to create a tailored weight for my own analysis to increase the sample size. Do you have any suggestions or guidelines on how to create such a weight?

Thank you again for your time and assistance.

Best wishes,

Yanan

#### #4 - 05/10/2023 08:23 AM - Understanding Society User Support Team

*- % Done changed from 10 to 50*

#### #5 - 05/12/2023 12:49 PM - Olena Kaminska

Yanan,

On tailored weights we have a online workshop for you: https://www.understandingsociety.ac.uk/help/training/online/creating-tailored-weights .

The document is not published yet, but here is question 19:
19. Scaling weights
In pooled analysis and sometimes in other types of analysis you may need to apply an additional scaling to our weights. Our weights have a mean of 1 in each wave, which means that if combined in a pooled analysis the waves with smaller sample size will have a smaller contribution in your analysis. This includes BHPS waves and later waves (as sample size decreases with attrition). Ideally, when combining events / states over 30 years (for example) you want each year to have the same importance. To ensure this follow this example to calculate an additional scaling for your weights. For example, you are looking at job quality and therefore are pooling information from wave 2, 4, 6 & 8 as these are the waves when the questions are asked. Here is how to create a scaled weight for this analysis.

```
ge weightscaled=0
replace weightscaled=b_indpxub_xw if wave=2

ge ind=1
sum ind [aw=b_indpxub_xw] if wave=2
gen bwtdtot=r(sum_w)
sum ind [aw=d_indpxub_xw] if if wave=4
gen dwtdtot=r(sum_w)
sum ind [aw=f_indpxub_xw] if if wave=6
gen fwtdtot=r(sum_w)
sum ind [aw=h_indpxub_xw] if if wave=8
gen hwtdtot=r(sum_w)

replace weightscaled=d_indpxub_xw*(bwtdtot/dwtdtot) if wave=4
replace weightscaled=f_indpxub_xw*(bwtdtot/fwtdtot) if wave=6
replace weightscaled=h_indpxub_xw*(bwtdtot/hwtdtot) if wave=8

You can double check by looking at the sum of ind with weightscaled for each wave – it should be the same.
sum ind [aw=weightscaled] if wave==2
sum ind [aw=weightscaled] if wave==4
sum ind [aw=weightscaled] if wave==6
sum ind [aw=weightscaled] if wave==8
```

#### #6 - 05/23/2023 11:27 AM - Understanding Society User Support Team

*- Status changed from In Progress to Feedback*

#### #7 - 11/30/2023 01:01 PM - Understanding Society User Support Team

*- Category set to Weights*

*- Status changed from Feedback to Resolved*

*- % Done changed from 50 to 100*