Understanding Society User Support - Support #1786

Weights for longitudinal study

10/17/2022 05:41 PM - Connor Gascoigne

Status:	Resolved	**Start date:**	10/17/2022
Priority:	Normal	**% Done:**	100%
Assignee:	Olena Kaminska		
Category:	Weights		

Description

Hi Olena,

I am currently looking to see the effect of government policy on the mental health of people in England, Scotland, and Wales. I drop Northern Ireland (NI) since a dataset I combine with at a later stage does not include information for NI. As part of that, I am performing a longitudinal analysis using the Waves 1-11 from the UKHLS to produce an estimate for each individual in the survey. I wish to take the individual level estimates and aggregate them to produce national and regional level estimates. To make sure the aggregated estimates properly account for the survey design, non-response, and any additional stratification, I plan on using the survey weights to produce the initial estimates.

I have seen there are two types of weights I can use: longitudinal and cross sectional.

For the longitudinal weights, I believe I would take the most recent surveys weight. For me, this would be k_indinus_lw. I then attached this weight to all the individuals for all waves (i.e., the weight for an individual is their k_indinus_lw weight for all waves). From this arises my first questions:

(Q1). I believe the weighting is weighted to include those individuals in NI as well. Since I do not consider these individuals in my analysis, will the fact I remove these individuals from the data set affect the weighting? Alternatively, do I have to alter the weights when I remove the respondents from NI?

(Q2). If a respondent lives in, say, Scotland, then I will include their response in the survey. If between waves 6 and 7, they move to NI, then due to the way I sort the data I will remove their responses from wave 7 onwards. Much like in (Q1), do I need to account for this by altering the longitudinal weights.

A main benefit of the longitudinal weights is the creation of a balanced dataset. From (Q2) (and similar examples where an individual's change in response means I drop them), I create an unbalanced dataset. This got me wondering if I would be better to use the cross-sectional weight and then pool them to create my own set of weights. This is because it would be useful to still include an individual's response even if they do not respond to all the surveys and this is essentially what I am doing for the example in (Q2). If this is the case:

(Q3). Because the naming convention changes for Wave 1, 2-5 and 6+, could I confirm if the weights I would need to make a pooled weight would be a_indinus_xw, b_indinub_xw, c_indinub_xw , d_indinub_xw, e_indinub_xw, f_indinui_xw, g_indinui_xw, h_indinui_xw, i_indinui_xw, j_indinui_xw, and k_indinui_xw'

(Q4). If these are the correct cross-sectional weights, what is the best way to go about making a pooled weight? I work in R and have my data organised into long format where each individual has one row per wave and a column for each of the above weights. Due to this, in each weighting column there is an NA for all rows except the rows relating to the wave for that weight.

I apologise for such an involved question, and I hope I have managed to explain myself in a manner that is understandable - if I need to better explain myself, then I am more than happy to do so. If you can give me any guidance at all, I would really appreciate it!

Thank you in advance for any help!

Kind regards,
Connor

History

#1 - 10/18/2022 02:30 PM - Olena Kaminska

Connor,

Thank you for your question. The answers are below:
Q1: Weighting represent any subgroup of the population - you don't need to do anything additionally. See Q1 and Q2 in https://www.understandingsociety.ac.uk/sites/default/files/downloads/general/weighting_faqs.pdf
Q2: The weights represent any subpopulation. It's up to you how you define it, i.e. whether you want to include or exclude those who lived in NI at

some point of time. However you define your subpopulation weights will work fine with it.
Q3 & Q4: Pooled analysis is an option but is used for studying different sampling units to non-pooled analysis (e.g. events). See below for details.
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.).

Finally, make sure to use strata and psu alongside weights.
Hope this helps,
Olena

#2 - 10/19/2022 12:57 PM - Understanding Society User Support Team

- Status changed from New to Feedback

- % Done changed from 0 to 90

- Private changed from Yes to No

#3 - 11/30/2023 11:08 AM - Understanding Society User Support Team

- Status changed from Feedback to Resolved

- % Done changed from 90 to 100