Understanding Society User Support - Support #1308

Cross-sectional weights 0 values

01/29/2020 10:33 AM - Pascale Bourquin

Status:	Resolved	**Start date:**	01/29/2020
Priority:	Urgent	**% Done:**	100%
Assignee:			
Category:	Weights		

Description

Hello,
We are doing crossectional analysis combining all waves from USoc and the BHPS and include all household members (and all regions of the UK whenever become available). We are basically taking information from different waves, for example, wealth from the waves with wealth modules and labour market outcomes/earnings for when people are in their mid 20s. Currently we are using the crossectional weights from the last wave we observe people so we have constant weights even though we use information from various waves for each unique observation (so depending on which wave this are (a-f)xewght (waves 1-6 BHPS), (g-h)xewghte (waves 7-8 BHPS), (i-j)xewtsw1 (waves 9-10 BHPS), (k-..)xewtuk1 (waves 11 onwarfdd BHPS), a_psnenus_xw (wave 1 Usoc), (b-h)_psnenub_xw (wave 2 onwards of USoc)). We are then adjusting them to account for selection of our sample based on observable characteristics.
I have two specific questions:
1) Are the weights we are using as our base weight (that is, we are using these weights to create our own adjusted weights) the right ones for this kind of analysis? and
2) Why is it the case that some of these weights are 0? I understand why longitudinal weights would be 0, however cannot quite understand why any of the cross-sectional ones would be. It is especially confusing that specific individuals have non-zero cross-sectional weights for some waves and not for others.
Any help answering the above 2 questions would be much appreciated.
Many thanks and best,
Pascale

History

#1 - 01/29/2020 11:41 AM - Olena Kaminska

Pascale,

Thank you for your question. Before I answer it I have a few clarifications. I am not sure what your analysis is. From your description is sounds like it may be longitudinal: in one model you use information from earlier and later waves. Is this correct? If this is the case and you wanted to use one of our weights you would need to use a longitudinal weight from the last wave in your analysis (note, not cross-sectional weight).
But I also understand that you want to make your own attrition adjustment. This is fine, but to advice I do need a few more details. If you use multiple models in your analysis you would need to create this adjustment possibly for each model separately if the models are based on different waves - this wasn't very clear.
On the other hand, if your analysis is truly cross-sectional and you ever use all the information only from one wave - you should use the cross-sectional weight for that wave. You can't use cross-sectional weight from a different wave to analyse information in a current wave - the results will be wrong.
You also mention your own adjustment. Is this adjustment for attrition - or is this a different adjustment? The way you described it sounds unclear, but I feel you didn't provide much information about it.
Let's say you are interested in a longitudinal effect of A on B. A is observed in 1991 (wave 1 of BHPS) and B is observed in wave 9 of UKHLS. In this situation you can use our 91_lw weight. Or model your own attrition correction. In this situation you can start with wave 1 of BHPS cross-sectional weight and model response to wave 9 of UKHLS conditional on response in wave 1 of BHPS. It is important in this situation to take into account: death, new birth, rising 16 (if you are using adult questions), mortality adjustment.
If you want to model attrition starting from a point later than wave 1, you should use either issue weight (_li in wave 2 or 6), or longitudinal weight. Using cross-sectional weight as a starting point in such situation will give you wrong results.
Our cross-sectional weights are zero for people who missed enumeration in one or more waves (though there are a few exceptions).

I will be able to advice you better if you provide me with more details on the waves you want to analyse and the type of analysis (longitudinal vs. cross-sectional and which instruments the information comes from).

Thank you,
Olena

#2 - 01/29/2020 02:49 PM - Pascale Bourquin

Hi Olena,
Thanks a lot for you quick response. Here some clarifications:
- Basically, what we are doing is selecting a group of people who are observed both at least once when they are young and living with their parents as well as at least once when they are in their mid/late 20s. We then calculate wealth their parents hold (so most recent wealth observation of parents as measured by any of the wealth modules) and want to investigate how levels of parental wealth differ by child outcomes such as education/ labour market outcomes etc. So essentially, as a final data set we have around 5000 people for whom we have a parental wealth observation (which is

pulled from whichever wave with parental wealth is most recent for the relevant individual) as well as education/ earnings and labour market outcomes (which will be pulled from the wave in which the relevant child was in their mid/late 20s). So really, across our selected sample, variables can be pulled from a variety of different waves (mainly from 2 different waves for each unique observation, though which two will ddiffer by individual). Does that make sense?
- Given the above, we believe the cross-sectional weights may be more appropriate than the longitudinal ones… as we are pooling individuals whose outcome variables are observed at different points in time (but we are not looking at for example, how an individual's earnings or education outcomes change over time). Please do let us know if we are wrong.
- The adjustment that we undertake is for attrition, as of course, the sample of people who are observed both when they are young with their parents (to get the wealth variable) as well as again later (to get labour market outcomes) is selected in the sense that homeowners are less likely to have attrited etc. So what we do is basically estimate a probit model to get the probability of being in our sample based on various observables and then create new weights by multiplying the inverse of the predicted probability that you land in our sample based on observables times the original weights.
- Regarding the 0-values for cross-sectional weights, can I just check that I have understood correctly: are you saying that those with a 0 value for USoc wave 8, for example, might have been enumerated in wave 8, but have a 0 value as they missed being surveyed in wave 2 or wave 3 (or any other previous wave) for example? So these are people who return after having missed a round?
Many thanks for your help, please do let me know if you require further clarifications.
Pascale

#3 - 01/29/2020 03:35 PM - Olena Kaminska

Pascale,

Your clarification helps a lot, and I think your situation is a little bit more complex and unique than I thought earlier. So a few points here:
1. Because you need observations on a person in two points of time (with and without parents) this because a longitudinal analysis (even if you want to know about them today - you need predictors from an earlier wave). If this is correct your analysis should be limited to OSMs.
2. Indeed you are dealing with joint response of parent-child pair. I am glad you are thinking to use probit model to add additional correction. Here is my suggestion on how to do it. Start with our longitudinal enumeration weight in wave 9 (if you use wave 9 for anyone). This will automatically delete TSMs (you don't need them as keeping them in will give you wrong results). This weight also would exclude ineligible correctly, so you won't need to worry about them.
Conditional on positive weight model people who are in your model as 1 and not in your model as 0. If you have item missiness this can be corrected at this step as well at the same time.
3. Yes, your understanding of my comment about zero weights is correct. It is in fact a bit more complex - so see my general reply to this below.
4. Finally you could almost avoid zero cross-sectional weights if you use your own tailored weighting (which you are planning anyways, but this becomes much more complex). So, you can start with for example wave 1 cross-sectional weight if you use only GPS and EMB samples, or with _li weight in wave 2 or wave 6 if you want to add BHPS and IEMB samples respectively. There are a number of things to remember then, like birth and death. But I suggest you follow our technical description of weights for this stage.

Should I worry about zero weights?
There are two main reasons for zero weights: there are zero weights by sample design, and there are zero weights as a result of fieldwork issuing rules.
We have around 1000 people with zero weights due to sample design. These people have zero design weights as well. They are 'TSMs from wave 1' selected through EMB and IEMB boosts. These are non-eligible people (e.g. white British) who are co-residents with eligible people (e.g. ethnic minorities or immigrants) at the first wave of sample selection. Their zero weights are correct and are related to the sample design. The sample design was implemented in the most cost efficient way to meet the need for analysing ethnic minorities, recent immigrants and the whole of UK population. Avoiding these zero weights would cost extra money but would not add much to the precision of UKHLS estimates.
There are also zero weights which result from a fieldwork issuing rules. If all nonresponding households that missed previous wave were not issued to the fieldwork and were dropped you would not observe zero cross-sectional weights other than 'TSMs from wave 1' zero weights. This would result in higher attrition rate initially, though likely to be lower attrition rates in later waves due to most of reluctant households having dropped in the first few waves. Importantly this would result in decrease in sample size much faster. UKHLS instead issues to the field households that missed a few consecutive waves which results in non-monotone household attrition. We believe this is an important approach, especially for a long-term future, when we expect that many types of analysis will be able to deal with non-monotone attrition through the analysis itself, or where special software may be developed to take into account analysis-tailored non-monotone nonresponse taking into account all the non-missing information.
Our longitudinal weights are developed for monotone attrition, i.e. requiring a response to a particular instrument in each wave. If you use some but not all combinations of waves (e.g. waves 1, 3 and 8) you can increase the number of respondents in your analysis via creating a tailored weight.
Our cross-sectional weights are based on almost monotone attrition – they are based on household participation in each wave except for waves 3-5. You can increase the sample size in your analysis by creating a tailored weight.

Hope this helps,
Olena

#4 - 01/30/2020 04:30 PM - Pascale Bourquin

Hi Olena,

Thanks a lot for your detailed response. Apologies that it has taken me a while to get back to you, I was discussing the above with my colleagues to decide how we can best implement this. We have decided to do the following:

1) We will take the "longitudinal person UKHLS+BHPS+IEMB inclusion weight" (*_psnenui_lw) as our base waits and adjust them as described in 2). These weights will be taken from the wave where the child outcome variables (homeownership/ earnings etc.) stem from.

2) For the adjustment coefficient: we will estimate a probit model with the left hand side being a dummy variable that = 1 if individuals have a parental wealth observation (from any of the wealth module waves) and 0 if not. We can then estimate for the full sample (pooling all waves) what the probability is that an individual has a parental wealth observation. One question we have on estimating this model is whether or not to weight the probit regression with the longitudinal weights? We think the answer is no (?), that is we should not weight at all when predicting the model. Is this correct?

3) Finally, we will create a new weight by multiplying the longitudinal weight of 1) by the inverse of probability calculated in 2) for the wave where we have our child outcome from

The idea behind creating these weights is that we want to account for 1) the probability that someone makes it to wave X of USoc (final wave observed in our case) conditionaal on observables AND 2) that they have at least one parental wealth observation (conditional on observables). Does this sound right to you?

Many thanks again for your help with this, it is much appreciated.

Best,

Pascale

#5 - 01/31/2020 03:41 PM - Stephanie Auty

- Status changed from New to In Progress

- % Done changed from 0 to 60

- Private changed from Yes to No

#6 - 02/03/2020 11:21 AM - Olena Kaminska

Pascale,

The solution that we have discussed would be correct in a simple situation (without pooling) where for example you use only people from waves 1 to 9, and everyone in your analysis have responded at wave 9. Because you want to pool information from across the waves the situation becomes slightly more complex. I have two questions to you:
1. Does your definition of your population depend on being observed in a later wave (and responded to the questionnaire for example)? In other words is it enough to be a child of a wealth parent to be in your analysis? Or do you have to also be employed, and maybe with a particular income (and thus having responded in a later wave) - I am asking here really about theoretical definition rather than how it is used in your analysis? It is important for weighting, because in the former situation you can start with an issue weight, in the latter you may need a weight that corrects for longitudinal nonresponse in a particular instrument.
2. Depending on the structure of your data it is likely you will need a separate probit model for each wave combination, or at least for each last wave of observation. So if you have a set of people last observed in wave 7 you will need a separate model for these, then for those observed in wave 8 etc. etc.

I appreciate this isn't straightforward and I am also not 100% clear about your data set up. It may be easier to talk to me via videochat - if you are interested please email usersupport@understandingsociety.ac.uk to set up an appointment. Please refer to the issue num 1308.

Thank you,
Olena

#7 - 02/28/2020 02:16 PM - Stephanie Auty

- Status changed from In Progress to Feedback

- Assignee changed from Olena Kaminska to Pascale Bourquin

- % Done changed from 60 to 80

#8 - 10/13/2021 12:03 PM - Understanding Society User Support Team

- Status changed from Feedback to Resolved

- Assignee deleted (Pascale Bourquin)

- % Done changed from 80 to 100