Understanding Society User Support - Support #658

Weights, BHPS and USoc pooled

11/17/2016 09:09 PM - Min Zhang

Status:	Closed	**Start date:**	11/17/2016
Priority:	Immediate	**% Done:**	100%
Assignee:	Min Zhang		
Category:	Weights		

Description

Dear Support team,

I have two questions with weights. Our data is based on a sub-population (respondents whose parents had non-missing interview records on their parents, that is, respondents' grandparents), while BHPS and USoc are pooled together. We are using BHPS wave 1-18, USoc wave 1-5. We have read through both BHPS and USoc user guides and weights-related posts on this forum but could not find the best solution.

Our research is on social class and education.

1. Class:
I use non-missing social class variables at the last, or most recent, waves that respondents attended. For example, in the case in which the last wave that a respondent was interviewed was BHPS wave 12, the non-missing social class recode at BHPS wave 12 would be taken as the measure of social class. Say, if the record at BHPS wave 12 is missing but the one at BHPS wave 11 is non-missing, the social class variable at wave 11 would be taken as the measure of social class.

This means that our data is drawn from different waves of BHPS and USoc. Can you please suggest which weights serve the best for our purpose.

2. Education
We use the highest educational qualifications that respondents have ever achieved. Which weights should we use in this case?

We incorporated gender, birth cohort, age, race (a roughly binary white vs non-white variable though), regions and data sources (a binary BHPS vs USoc variable) as the control variables in our analyses. We also take into clustered errors using household ID. We wonder whether it is statistically acceptable if there are no perfect weights used in our analyses.

We highly appreciate your supports.

Best wishes.

History

#1 - 11/17/2016 09:22 PM - Min Zhang

I am sorry that I forgot to mention our data structure which may further complicate the questions regarding weights.
If a respondent has non-missing records on both maternal and paternal grandparents, this respondent would show up twice in the data. If a respondent has non-missing records on only one side, either maternal or paternal grandparents, this respondents would show up once in the data.

Would you suggest that it would be OK not to use weights in our analyses?

Many thanks,
Min

#2 - 11/22/2016 01:32 PM - Victoria Nolan

- Category set to Weights

- Status changed from New to In Progress

- Assignee set to Victoria Nolan

- % Done changed from 0 to 10

- Private changed from Yes to No

Dear Min,

Many thanks for your enquiry - it has been passed on to our weighting team who will respond as soon as possible.

Best wishes, Victoria.

On behalf of the Understanding Society Data User Support Team

#3 - 11/22/2016 02:24 PM - Peter Lynn

Hi. Could you clarify:

1. I understand the analysis units are the "respondents", yes? (the information about the respondent's grandparents, provided by the respondent's parents, is essentially treated as an attribute of the respondent)

2. Which respondents are included? Is the availability of grandparents' social class the only restriction? (so it does not matter in which waves the respondent responded, provided they responded at least once?)

3. Is your analysis longitudinal or pooled cross-sectional?

Thanks,

Peter

#4 - 11/22/2016 02:44 PM - Min Zhang

Hello Peter,

Many thanks for your reply.

1. Yes the analysis units are the respondents.

2. Sorry that I did not make it clear. The sample is confined by the availability of grandparents' class and parents' class, education and income (respondents with single parents are included). For education, the sample is confined to the respondents over the age of 22; for class, the sample is confined to those over the age of 25.

3. The analysis is pooled cross-sectional.

I am concerned particularly with the issue that I mentioned at the #2 post. Respondents would be observed twice in the sample if the information on both paternal and maternal grandparents is available. Respondents would be observed only once if the information on only one-side grandparents(i.e., either paternal or maternal, but not both) is available.

A study on the relationship between grandparents and grandchildren used cross-sectional data (not Understanding Society survey) but did not use any weights. The author said that this is because the data was drawn from different waves.

Many thanks for your support again.

Kind regards,
Min

Peter Lynn wrote:

> Hi. Could you clarify:
>
> 1. I understand the analysis units are the "respondents", yes? (the information about the respondent's grandparents, provided by the respondent's parents, is essentially treated as an attribute of the respondent)
>
> 2. Which respondents are included? Is the availability of grandparents' social class the only restriction? (so it does not matter in which waves the respondent responded, provided they responded at least once?)
>
> 3. Is your analysis longitudinal or pooled cross-sectional?
>
> Thanks,
>
> Peter

#5 - 11/24/2016 04:20 PM - Peter Lynn

So, take for example your analysis based on all respondents who have at least one interview at which they were aged over 22. You can identify all such people in the data. For a pooled CS analysis of such people, I would suggest that you use the cross-sectional weight from the first wave at which they responded and were aged over 22.

However, you have considerable further selection, as you then drop people for whom the relevant information about grandparents is missing. I would suggest that you should adjust the weights to account for this. Basically, all the respondents identified at step 1 above should have two records in your analysis data set - one for each pair of grandparents - but one or both of those records may have to be dropped from the analysis due to unobserved grandparent data. Create the data set in which each respondent appears twice and create a 0/1 indicator of whether or not the record can be used in your analysis. Model this indicator based on relevant respondent characteristics (e.g. a logistic regression). This will give you a predicted

probability for every respondent of the grandparent information being present. Call this P. You then need to adjust the weight from step 1 above by multiplying it by 1/P for all the records that can be included in your analysis.

This deals simultaneously with some respondents being present twice as you describe, so you don't need to do anything else.

If this is too complicated for your purposes, you could just use the cross-sectional weight as described in step 1 above, and then multiply it by 0.5 for all the respondents who appear twice in your analysis data set. However, this makes no adjustment for selection based on observation of grandparent details. I would imagine this is highly selective (related strongly to parents' age and maybe also to age at which respondent left parental home, in addition to usual non-response predictors), so it might be dangerous to not make such an adjustment.

Hope this helps.

#6 - 11/28/2016 11:15 AM - Victoria Nolan

- Status changed from In Progress to Feedback

- Assignee changed from Victoria Nolan to Min Zhang

- % Done changed from 10 to 80

#7 - 11/29/2016 08:04 PM - Min Zhang

Dear Peter,

Many thanks for your suggestions. They are very helpful. I apologize for my late reply. I spent a few days testing weights with my analyses following your instruction. After a few tries I have a couple of questions as follows:

1. Although you did not mention how I scale weights while pooling BHPS and US together, I guess that the weighting would not correct if the BHPS weights and US weights are simply appended together. If I do need to scale weights, my question is how I should do this correctly in my data.

2.You suggested that "you use the cross-sectional weight from the first wave at which they responded and were aged over 22." I also found similar solutions to the questions of cross-sectional weights in the online support. Why use the weights from the first wave at which they respondents instead of using the weights from the wave from which respondents' class was drawn? For example, a respondent was first interviewed at BHPS wave 3 and his or her latest record of class was available at BHPS wave 9. In this case, why did you suggest that I use the weights from wave 3 instead of wave 9?

3. I am not an expert of weights. I have read the official user guides but am not confident if I have used weights correctly. Could you please confirm if the weights I used are correct?

For BHPS and its extension in US:
from wave 2 onward to wave 10, xrwght;
from wave 11 onward, xrwtuk1.
US wave 2: b_indinbh_xw
US wave 3: c_indin01_lw
US wave 4: d_indin01_lw

For US wave 1 - wave 5: w_indpxub_xw w_indinub_xw

In the BHPS, proxy and telephone respondents have zero respondent weights but positive enumerated individual weights. US provides both main interview weights and main and proxy weights. My data was drawn from both main interviews and proxy weights. Would you suggest that we switch to enumerated weights for BHPS?

4. For the BHPS and US, the means of the weights I use are equal to one. Would this be a problem?

I am grateful for your suggestions. Your suggestions are very important to me. Many thanks for your time.

Best wishes,
Min

#8 - 12/13/2016 03:43 PM - Peter Lynn

Min,

Apologies for the slow reply.

1. You should not need to do any further scaling, as each weight is independently scaled to a mean of 1.0. Thus, each wave/year will be (approximately) equally represented in pooled analysis.

2. Yes, you should use the weight for whichever wave the data relevant to your analysis come from. I may have misunderstood the relevance to your analysis of being a respondent at age 22. If you only use data for a respondent from wave 9, and not wave 3, you should use the wave 9 weight.

3. For pooled CS analysis of data from the individual or proxy interview, for Understanding Society I think you should use:
Wave 1: a_indpxus_xw
Wave 2: b_indpxub_xw
Wave 3: c_indpxub_xw
Wave 4: d_indpxub_xw

Wave 5: e_indpxub_xw

(If you use indin instead of indpx, you will find that all the proxy respondents have zero weight and are therefore dropped from your analysis.)

4. No, this is probably what you should prefer conceptually (see answer 1. above).

Peter

#9 - 01/04/2017 01:29 PM - Victoria Nolan

- Status changed from Feedback to Closed

- % Done changed from 80 to 100