Understanding Society User Support - Support #1558

Cross-sectional vs longitudinal weights

06/28/2021 01:15 PM - Carolin Schmidt

Status:	Resolved	Start date:	06/28/2021
Priority:	Normal	% Done:	100%
Assignee:			
Category:			

Description

Hello,

I am working with Understanding Society for the first time and have a few questions.

My coauthors and I are analysing immigrants' age at immigration and their housing outcomes (own or rent). We'll merge household data with individual (i.e. respondent) data so that we can use the person's age at migration and other variables together with household data. My questions mainly refer to the weights that we should use.

Since we're interested in an invariant variable (age at migration) we cannot use household fixed effects and would therefore run regressions in the pooled sample in our main regressions. We classify a household as an immigrant household if the respondent is an immigrant. If I'm not mistaken, that means that we should use cross-sectional (i.e. a_, b_, c_, …) household (i.e. hhden_) weights, is that correct? That is, I'll have „a_hhden_" for all cases in wave 1, „b_hhden_" in wave 2, etc. If we run robustness checks within age-at-migration subgroups, we can make use of the panel structure again. I suppose we should use the longitudinal weights then?

We are not only interested in immigrants from the five subgroups mentioned in Q4 of the FAQ help file, but also in immigrants from other countries. If, in the pooled sample, we first use data from wave 1 onwards and run additional checks using only data from wave 6 onwards, am I right that we'd have to use us_ weights (wave 1 onwards) and ui_ weights (wave 6 onwards), and that it would be wrong to use data from wave 1 onwards with us_weights and an „if wave >= 6" qualifier?

Thanks a lot in advance.

Best wishes,
Carolin

History

#1 - 06/29/2021 03:00 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

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 - 06/29/2021 03:36 PM - Olena Kaminska

Carolin,

Your problem is not weights but how you are using longitudinal dataset. By default the longitudinal aspect means that the dataset does not represent immigrants since wave 1 (or since the boost). If you pick wave 5 you will have only immigrants who came to the country before wave 1 - it does not represent immigrants who came to the country in the previous 5 years. You need to carefully keep this in mind.
Your best option is to just select wave 6 and analyse it separately with ui_xw weights. It has a boost of immigrants and is representative of the country population (including all immigrants) at that point of time.
Pooling across waves needs to be done extremely carefully (or better avoided) because the definition of population of immigrants is different in each wave. If you are interested in studying only immigrants who immigrated to the country before 1991 you have no problems- you can pull all the datasets in the usual way.
Wave 1 with us_xw weights are also representative of the population, including immigrants at that time. Let me know if you want to pool the data - it won't be straightforward. But first think conceptually if you think immigrants at the time of wave 1 are similar to those at wave 6.

Hope this helps,
Olena

#3 - 06/29/2021 04:21 PM - Carolin Schmidt

Dear Olena,

Thanks a lot for your help. I did read the relevant sections on immigrants in the help files but did not completely understand the implications as it seems. Your answer helped me to clarify this.

So, if we run cross-sectional regressions in wave 6 only, we'd use the us_xw weights, and then we could run panel regressions in waves 6-10 using longitudinal ui_ weights (those that refer to wave 10), is that correct?

Carolin

#4 - 06/30/2021 01:37 PM - Olena Kaminska

Carolin,

Use ui_xw weight for wave 6, and use ui_lw weight if your longitudinal analysis starts at wave 6 or later (remember you don't have new immigrants since wave 6 in your analysis).

Olena

#5 - 06/30/2021 01:51 PM - Carolin Schmidt

Hi Olena,

Thanks for the quick reply. One last question (at the moment, at least): the sample is representative of immigrants in wave 6, but is it for the following waves too, when we look at it from a panel perspective?

Thanks,
Carolin

#6 - 06/30/2021 02:59 PM - Olena Kaminska

Carolin,

Yes, in a sense that all immigrants who were present in the country are represented longitudinally (some would have left the country and this is somewhat captured in the data). So at wave 10 you can study all immigrants who were in the country continuously for the last 4-5 years (since wave 6), and immigrated to GB at any time since wave 6 (check if IEMB covers NI).

Hope this helps,
Olena

#7 - 07/01/2021 11:40 AM - Carolin Schmidt

That does help a lot. Thanks, Olena.

#8 - 07/04/2021 08:41 PM - Carolin Schmidt

- File Screenshot 2021-07-04 at 20.29.51.png added

Hi Olena,

I'm having a few more questions now.

Like I said, we'd like to classify entire households as immigrant households if the person who answers the household questionnaire (let's say the household reference person) is an immigrant.

a) I can't seem to identify all immigrants correctly. I can use information from the variables f_ukborn and f_ff_ukborn: if either of them says that the person was not born in the UK, they're an immigrant. However, we'd like to know the country of birth of these immigrants. For many cases I can get further information from the variable f_plbornc but it's not always populated for those that were not born in the UK. Am I missing another variable, e.g. one that is fed forward and that contains the COB of immigrants that I'm currently missing?

I'm attaching a screenshot to show some examples.

Screenshot 2021-07-04 at 20.29.51.png

The variable "immigrant" is 0 when either f_ukborn or f_ff_ukborn says that the person was born in the UK. It is 1 if it says the contrary. I have a few missings that I can't identify (e.g. row 740).

The variable "cob" is basically a copy of f_plbornc for those people that were born abroad. I set it to 1 for those that are born in the UK (if immigrant 0). Here, too, I can't populate all cells. For instance, for the cases in row 741 or 742, I know that they were not born in the UK (from f_ff_ukborn) but f_plbornc doesn't contain any information on their country of birth.

b) This question refers to the use of the household reference person when we combine individual with household data. The dependent variable in our model is whether the household owns or rents its primary residence. We'd like to add individual (i.e. reference person) data to the regression, e.g. age or sex. When we combine data like this, my understanding is that we should use household and not individual weights, is that right?

c) I have identified each household reference person (hhref = 1, 0 otherwise). Basically, anyone who's not the household head isn't of interest for us. Can I just drop all cases where hhref 0 or would that mess up the weights?

Thanks a lot!
Carolin

#9 - 07/05/2021 03:51 PM - Olena Kaminska

- Assignee changed from Olena Kaminska to Alita Nandi

#10 - 07/06/2021 04:50 PM - Understanding Society User Support Team

- Assignee changed from Alita Nandi to Olena Kaminska

- % Done changed from 10 to 50

#11 - 07/07/2021 12:28 PM - Olena Kaminska

Carolin,

Here are replies to b) and c):
If you are analysing households, and you want to represent households in the population then use household weights (and one observation for each household);
If you want to represent reference people of households (whatever it may mean) or heads of households (this may be a different person to that who gave a household interview) use individual weights of that person (again, you may have only one observation per household).
If you want to represent individuals but to use household level variables in your model use individual level weights.

Someone else from the team will respond to question a).
Hope this helps,
Olena

#12 - 07/07/2021 12:28 PM - Olena Kaminska

- Assignee changed from Olena Kaminska to Alita Nandi

#13 - 07/07/2021 01:26 PM - Understanding Society User Support Team

- Status changed from In Progress to Feedback

- % Done changed from 50 to 80

About (a)

UKBORN (Whether born in E, S, W, NI or outside UK) & its follow-up PLBORNC (Country of birth if said they were born outside UK) is asked the first time a person is interviewed as an adult. So, if you look at the variable in any specific wave file it will have valid values only for those who are being interviewed for the first time in that wave. After the first wave, this set will only include the new entrants, and those who were not interviewed before. So, you will find a lot of missing values. These variables are picked up from the wave that they were reported and combined - these are then made available in the file XWAVEDAT. This file is a individual level file created by us, which includes stable information for all sample members collected from multiple sources and waves. So,it is best to pick up these variables from xwavedat and then merge that into your dataset using pidp.

Only for the IEMB sample in Wave 6, information about whether a household member was born in the UK or not was collected as part of the household grid. So, only for this sample and in that wave this information is available for all household members even if they are not adult respondents. This is the variable BORNUK. The UKBORN variables from across all waves and the BORNUK variable in this wave were combined to form a variable BORNUK_DV (born in UK or no) and is made available in XWAVEDAT. So, there will be less missing values for BORNUK_DV than UKBORN. So, BORNUK_DV in XWAVEDAT s the best variable to identify immigrants.

As information about the migrant status of children was not asked in any other wave, we suggest using the information on their date of birth and the date of the parents' arrival in the UK to identify whether they were born in the UK or not (this requires making the assumption that the parents were living continously in the UK after they arrived).

#14 - 07/21/2021 02:14 PM - Carolin Schmidt

Dear Alita,

Thanks, that helps a lot.

I have another follow-up question (probably fo Olena?). It may sound strange, but I can't find the longitudinal weights for our analysis of the period up to wave 10. If I'm not mistaken, we should use j_hhdenui_lw, given that we use waves 6-10 and want to represent households. I've gone through all files and found a number of cross-sectional weights, but for some reason I'm not able to locate the wave I need for my panel regressions.

Can you please tell me where to look or let me know if I'm supposed to use different weights?

Thanks,
Carolin

#15 - 07/22/2021 06:02 PM - Olena Kaminska

Carolin,

Thank you for your question. UKHLS does not think of households as longitudinal entities as they change with time. For this reason the design is set up to follow individuals longitudinally, and only individuals get longitudinal weights.
Theoretically it would be possible to define a weight for a longitudinal household if you have a definition for what this would mean in your situation. If you still decide to proceed with a longitudinal definition of a household and would like to specify it to me (with examples of what happens when a household splits, someone enters, someone is born, dies etc.) then I should be able to advice on the weights.

Hope this helps,
Olena

#16 - 08/04/2021 12:51 PM - Carolin Schmidt

Dear Olena,

Thanks for the details. I'll meet my coauthors soon and will discuss the definition of households with them. For now, I think it would be the best if we continued with wave 6 only.

I have another question here: household wealth is collected in waves 4 and 8. We can't use wave-4 wealth because we want to exploit the IEMB sample. Could we use wave-8 wealth to proxy wave-6 wealth (I am aware that we'd have to make strong assumptions here)? What would happen to the weights? Is this even possible?

Thanks a lot,
Carolin

#17 - 08/05/2021 01:49 PM - Understanding Society User Support Team

- Assignee changed from Alita Nandi to Olena Kaminska

Those aspects that are not expected change frequently are asked less frequently including wealth. So, it is expected that wealth information in the waves it was not asked will be imputed using the information collected in other waves - but whether it is ok to impute information collected later is ok depends on your research question.

I will assign this to Olena to answer the weighting question.

#18 - 08/05/2021 06:21 PM - Olena Kaminska

Carolin,

I can't comment on the substantive idea of using later wave income. I will only mention that I wouldn't use household income as it is at wave 8 to infer it for wave 6 - instead I would use individual income parts and reconstruct household income. This is because household income would change a lot for specific categories of people: e.g. those who married between wave 6 and 8 or got divorced.

In terms of weights as you would be using wave 6 and wave 8 information you should use wave 8 longitudinal weight and you'll be fine.

Hope this helps,
Olena

#19 - 08/06/2021 10:09 AM - Carolin Schmidt

Hi Olena,

Thanks for the quick reply. We are looking at households and so, there are no longitudinal weights in this case. Is there something else we could do?

Carolin

#20 - 08/06/2021 11:08 AM - Olena Kaminska

Carolin,

If you tell me how you define your longitudinal households I would be able to suggest the best option for you. For example what happens if two people from a household split up?

Thanks,
Olena

#21 - 08/06/2021 01:25 PM - Carolin Schmidt

Hi Olena,

Our research question doesn't depend on the wave number and we were wondering if we could just use wave 8. Earlier you recommended to use wave 6 to analyse immigrants but (perhaps another stupid question): is wave 8 as representative of (i) immigrants and (ii) the general population in that wave as wave 6? That would solve basically all our problems because then, we would have controls for wealth and debt, which are strongly correlated with housing outcomes.

Concerning your previous question: we look at immigrant households which we classify as "immigrant" if the household reference person (the person who pays the rent/mortgage) is an immigrant. If that immigrant HRP happens to divorce their spouse, we would drop the ex-spouse's new household (if it's even in the dataset) and treat the newly divorced immigrant as the same immigrant household as before. If the immigrant HRP marries or moves in with someone, we'd treat them as the HRP in the new household as well, even though the new partner may pay the rent.

Getting weights for this case would be brilliant as it would allow us to run a panel regression using waves 8 and 12 (that is, if using wave 8 makes sense in the first place).

Thanks so much,
Carolin

#22 - 08/06/2021 02:48 PM - Olena Kaminska

Carolin,

Wave 8 is representative of the population with one exception: it does not include immigrants that entered the country since wave 6.
Wave 6 includes all immigrants, including those who immigrated that year.
The way you describe your research question sounds like you can study individuals with some household attributes. So if you study household reference people with all the household and individual characteristics and you could potentially run this as an individual analysis with individual level weights.

Hope this helps,
Olena

#23 - 08/06/2021 03:16 PM - Carolin Schmidt

Hi Olena,

Thanks, this helps a lot, again!

Yes, we are using information from the household head (age, age at migration, country of birth etc.) and the household (household income, hopefully household wealth, household size, region, marital status). If you think we can use individual weights and waves 8 and 12, this would be fantastic! Out of curiosity, is there a rule, or rule of thumb, to decide if individual weights can be used even though we treat the individual as an entire household? What if we used many more household-level variables than individual ones?

So, we would then likely only use wave 6 first (in order to retain the maximum number of immigrants) with wave-8 wealth and longitudinal w8 weights, and then we could use waves 8 and 12 (with a lower sample size) and longitudinal w12 weights.

Carolin

#24 - 08/09/2021 08:41 AM - Olena Kaminska

Carolin,

If you represent individuals and therefore run individual-level analysis you can use as many household attributes as you like (no limit). You will be describing individuals who are immigrants (for example) who live in such and such households. This is an individual-level analysis and would require individual-level weights.

Hope this helps,
Olena

#25 - 10/12/2021 02:32 PM - Understanding Society User Support Team

- Assignee deleted (Olena Kaminska)

- % Done changed from 80 to 100

#26 - 10/12/2021 02:32 PM - Understanding Society User Support Team

- Status changed from Feedback to Resolved

Files

Screenshot 2021-07-04 at 20.29.51.png	193 KB	07/04/2021	Carolin Schmidt