Understanding Society User Support - Support #1221

Using UKHLS at the quarterly level

07/31/2019 11:33 AM - Tom Waters

Status:	Resolved	**Start date:**	07/31/2019
Priority:	Normal	**% Done:**	100%
Assignee:	Tom Waters		
Category:	Weights		

Description

Hi,

I am trying to look at various outcomes by quarter in a dataset which pools multiple UKHLS waves together with another dataset (the FRS).

However I understand that there is non-random sampling within wave. For example, in wave C Northern Ireland is only sampled in the first year (2011). I was sent a document from someone at Essex which discusses pooling data from different waves in UKHLS. It says, "The weights provided currently are not designed for pooling as they are scaled to a mean value of 1.0 within each wave, and therefore produce different weighted sample sizes in each wave. As a result, cases from later waves will be under-represented."

I have three questions:
1. Why is it that later waves would be under-represented? I should've thought that if they had the same average weight, then cases would be equally represented when you weight?
2. What steps should I take to ensure that quarterly level outcomes are representative?
3. Are there any additional steps I should take to take account of the fact that we are pooling with the FRS? Currently we are making sure that, within a financial year, the FRS observations and UKHLS observations have the same average weight. Is that sufficient?

Many thanks for your help.

Tom

History

#1 - 07/31/2019 11:58 AM - Alita Nandi

- Status changed from New to In Progress

- Assignee set to Olena Kaminska

- 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.

Best wishes,
Understanding Society User Support Team

#2 - 07/31/2019 01:28 PM - Olena Kaminska

Tom,

Thank you for your question. Here is the note that we are preparing to answer your question (see below). It should answer the first two questions of yours.

To clarify on your first question, because the weights are scaled to the mean of 1, the total of weights is equal to the number of respondents. With attrition this number goes down, but it is most important if you use BHPS with UKHLS - because BHPS has a much smaller number of respondents than UKHLS. If not adjusted for this earlier years that are represented by BHPS for example will be underrepresented in your analysis and the results will be heavily dominated by UKHLS years. This is only relevant if you are putting different waves together.

The third question: the only thing to keep in mind is the difference in the population definition: because UKHLS is a longitudinal study we are missing some immigrants. See the last FAQ on this which describes the population you can represent. This means that immigrants from FRS need to be upweighted to compensate for not enough in UKHLS. Remember that we represent cross-sectional population including immigrants in 1991, 2009 and 2015, so it is immigrants in-between that you need to take into account.

Can I run analysis on a calendar year / month?
Yes, it is possible to run analysis using a calendar year / month with a few extra adjustments. The UKHLS sample is designed such that each month of issue is a random representative (once weighted) sample of a population with some exceptions:
- NI is only present in issue months 1-12 (first year of issue)
- BHPS is only present in issue month 1-12 (first year of issue)

Because of this please use us_lw weight in your analysis, including for cross-sectional estimates. This weight correctly excludes BHPS.
Please also note that if you use months 13-24 you are excluding NI from your analysis. If you use months 1-12 NI has too high values of weights. You need to adjust for this. Here is the Stata syntax for NI adjustment if you use month 1-12:

```
gen adj=1
replace adj=0.5 if x_country==4
gen weight=x_xxxxxus_lw*adj
```

We suggest that you use month / year of issue rather than calendar month /year. Each month of issue has interviews from 3-4 months, but the majority of interviews come from the calendar month coinciding with the month of issue. The few interviews that come in later calendar months are of hard nonrespondents that require a number of attempts for a successful interview. Our weights are designed for a whole month of issue to represent the population. If you omit the interviews from the calendar months following the issue months you are excluding late respondents –a category of people who tends to be very different to earlier respondents.
If you still want to restrict your analysis to the calendar months there are two ways you can adjust for the late respondents:

- Create a tailored adjustment to our weight (see Appendix 1)
- Use late respondents from another issue months with our weights (see below).

Let's say you are interested in studying December of 2014. You optimal option with the highest sample size will be to combine data from:

- Wave 5 issue month 24 (only respondents in December of 2014, excluding late respondents)
- Wave 5 issue month 23 (only respondents in December of 2014, to compensate for late respondents missing from issue month 24)
- Wave 5 issue month 22 (only respondents in December of 2014, to compensate for late respondents missing from issue month 24)
- Wave 5 issue month 21 (only respondents in December of 2014, to compensate for late respondents missing from issue month 24)
- Wave 6 issue month 12 (only respondents in December of 2014, excluding late respondents)
- Wave 6 issue month 11 (only respondents in December of 2014, to compensate for late respondents missing from issue month 12)
- Wave 6 issue month 10 (only respondents in December of 2014, to compensate for late respondents missing from issue month 12)
- Wave 6 issue month 9 (only respondents in December of 2014, to compensate for late respondents missing from issue month 12)
- Use e_xxxxxus_lw weight for wave 5 and f_xxxxxus_lw weight for wave 6 respondents. You should combine these into one weight variable. No NI adjustment is needed. No extra nonresponse adjustment is needed as late respondents are compensated for by bringing them from previous issue months. But you will need a scaling factor (see a note on Pooling data from different waves for cross-sectional analysis).
- Use psu and strata variables from xwave .dat to take into account clustering and stratification.

Note if you want to study January 2014 for example, the information will come from 3 waves, because to compensate for missing of late respondents from wave 5, issue month 1, you will need to get them from wave 4, issue months 22-24. The rest will follow the above example.
If you use respondents from a calendar months / year just from one wave you will need extra adjustment for late respondents and for NI.

Pooling data from different waves for cross-sectional analysis

The weights provided are not designed directly for pooling data across waves as they are scaled to a mean value of 1.0 within each wave, and therefore produce different weighted sample sizes in each wave. The result is that cases from later waves will be under-represented. This matters because each monthly sample is not a random subset of the total. In the example of box 1, sample months 1 to 12 will be under-represented (as we are taking their data from wave 3, rather than wave 2 for months 13 to 24) . To overcome this, we should scale the weights for these cases to give the same weighted total that this sample had at wave 2. (Or we could equivalently scale the weights for the months 13 to 24 sample to equal their weighted total from wave 3.) Thus, the syntax becomes that in box 2.
This rescaling becomes even more important when pooling data from more than one 12-month period (e.g. two calendar years). In that case, in addition to the imbalance between the 24 monthly samples, the relative contribution to the estimate (weighted sample size) will also tend to be less for the later year(s) unless rescaling is done, such that each year contributes equally to the estimate. This is achieved by scaling all of the weights to the relevant weighted totals from one common wave.
Box 2: Example syntax for pooled analysis for cross-sectional estimation relating to calendar year 2011, with weight re-scaling

```
use "\\....\b_indresp.dta", clear
merge 1:1 pidp using "\\....\c_indresp.dta"

ge jbstat2011=0
replace jbstat2011=b_jbstat if b_month>=13 & b_month<=24
replace jbstat2011=c_jbstat if c_month>=1 & c_month<=12

ge weight2011=0
replace weight2011=b_indpxub_xw if b_month>=13 & b_month<=24
ge ind=1
sum ind [aw=b_indpxub_xw] if b_month>=1 & b_month<=12
gen bwtdtot=r(sum_w)
sum ind [aw=c_indpxub_xw] if c_month>=1 & c_month<=12
gen cwtdtot=r(sum_w)
replace weight2011=c_indpxub_xw*(bwtdtot/cwtdtot) if c_month>=1 & c_month<=12

ge psu2011=0
replace psu2011=b_psu if b_month>=13 & b_month<=24
replace psu2011=c_psu if c_month>=1 & c_month<=12

ge strata2011=0
replace strata2011=b_strata if b_month>=13 & b_month<=24
replace strata2011=c_strata if c_month>=1 & c_month<=12

svyset psu2011 [pw=weight2011], strata(strata2011) singleunit(centered)
svy: proportion jbstat2011 if weight2011>0
```

Which population can I represent with UKHLS?
You can represent the population residing in Great Britain (England, Scotland and Wales) in 1991, residing in the UK in 2007-2008 and in 2015-2016.

You can also represent the population residing in the Great Britain between 1991 and 1999, and in UK since 2001 excluding recent immigrants. If you use data between 1991 and 2009 you exclude immigrants to the GB since 1991, and to the NI since 2001. If you use data between 2009 and 2015-16, you exclude immigrants since 2009, if you use data collected since 2015-16 you exclude immigrants since 2015/16.
You can represent GB longitudinally since 1991, and NI or UK longitudinally since 2001 with BHPS data. The larger sample (UKHLS) starts from 2007/8 – starting at this time point is useful if you are looking at small subgroups.

#3 - 07/31/2019 02:13 PM - Tom Waters

Dear Olena,

Thank you for your prompt response. I am still thinking about the other points, but I had one immediate question on the below

Olena Kaminska wrote:

> To clarify on your first question, because the weights are scaled to the mean of 1, the total of weights is equal to the number of respondents. With attrition this number goes down

Suppose that wave A had 1,000 respondents in calendar year 2009 and 1,000 respondents in calendar year 2010. Suppose that all of those sampled in 2010 attrit, and all but one of those sampled in 2009 do. The one that doesn't attrit is then sampled in calendar year 2010, in wave B.

So when I want to look at the distribution of some outcome in calendar year 2010, I pool together wave A's 1,000 observations in 2010, and wave B's 1 observation in 2010. As far as I can tell, the code you pasted would cause wave B's one observation to have half the weight of the resulting 1,001 observation sample. Is that correct? If so, that seems like a very strange outcome: surely for estimating the distribution of this variable in 2010, any observation is 'as good' as any other (conditional on its original weight); the fact that it comes from wave A or B is essentially arbitrary. Am I misunderstanding something?

#4 - 07/31/2019 02:48 PM - Olena Kaminska

Tom,

Let's change the number to 800 in 2009 and 400 in 2010 after attrition (1000 each before). In this situation you have twice as many people in 2009 as in 2010 - so 2009 will be twice more important in your analysis than 2010. You want the years to be of equal importance - so you need to upweight each person in 2010 by 2. That's what the syntax does: the weight from 2010 would be multiplied by 800/400.

If you want your example of 1000 and 1, then you want 1 person to represent 1000 (so it is the same as in the other year). So you will need to multiply it by 1000/1 - that's what the syntax would give you.

If you have multiple years with n_j each it may be easier to calculate the average of all totals (n_av), and each weight should be multiplied by n_av/n_j. For example if you have 4 years with totals of 100, 200, 300 and 400, the average is 250. For each year to represent the same amount you would need the weight to be multiplied by 250/100 , 250/200, 250/300 and 250/400 respectively.

Hope this helps,
Olena

#5 - 07/31/2019 03:58 PM - Tom Waters

Olena Kaminska wrote:

> Tom,
>
> Let's change the number to 800 in 2009 and 400 in 2010 after attrition (1000 each before). In this situation you have twice as many people in 2009 as in 2010 - so 2009 will be twice more important in your analysis than 2010. You want the years to be of equal importance - so you need to upweight each person in 2010 by 2. That's what the syntax does: the weight from 2010 would be multiplied by 800/400.
>
> If you want your example of 1000 and 1, then you want 1 person to represent 1000 (so it is the same as in the other year). So you will need to multiply it by 1000/1 - that's what the syntax would give you.
>
> If you have multiple years with n_j each it may be easier to calculate the average of all totals (n_av), and each weight should be multiplied by n_av/n_j. For example if you have 4 years with totals of 100, 200, 300 and 400, the average is 250. For each year to represent the same amount you would need the weight to be multiplied by 250/100 , 250/200, 250/300 and 250/400 respectively.
>
> Hope this helps,
> Olena

Thanks Olena. I am probably misunderstanding, but this seems to be doing something slightly different. One issue is - if we are looking over multiple years, we want each calendar year to have the same weight in our analysis. I agree with that. But the box that you pasted (and the example I gave with 1000 and 1) is specifically looking at a *single calendar year*. The code in the box implies that observations sampled in the same calendar year should get different weights according to whether they are in wave B or C, when you pool together. That is what seems strange to me. Suppose that the sample sizes were as follows:

Calendar year 2009: Wave A, 800 obs; Wave B, 0 obs
Calendar year 2010: Wave A, 800 obs; Wave B, 400 obs
Calendar year 2011: Wave A, 0 obs; Wave B, 400 obs

Suppose I am just interested in calculating the distribution of some variable in *calendar year 2010*. Then the code above implies that I should give wave B observations twice the weight of wave A observations. That is what seems curious because, as I said above, the fact that an observation comes from wave A or B is essentially arbitrary.

Sorry if I'm being slow.

Tom

#6 - 08/07/2019 11:26 AM - Stephanie Auty

- Category set to Weights

- % Done changed from 0 to 60

#7 - 08/07/2019 11:26 AM - Stephanie Auty

- % Done changed from 60 to 50

#8 - 08/08/2019 10:14 PM - Olena Kaminska

Yes, your understanding is correct. But you multiply the weight from wave B by 2 only because in the first place it is twice smaller than it should be. If you want to give the same importance to people in wave A as in wave B in your analysis you need this adjustment.
Note that our weights are scaled to the mean of 1 - this means that the total of weights in wave A in 2010 is 1600, and in wave B is 800. So, if you put the two waves together without any adjustment, wave B will be twice less important. The adjustment corrects for this.
Now, this becomes very important in UKHLS because year 1 and year 2 within a wave are different: Northern Ireland is only in year 1, and BHPS is only in year 1 (it has boost of Scotland and Wales, which would mean that even if you exclude BHPS and NI from your analysis, GB will not be scaled correctly within a calendar year). There are a few ethnic minority boosts that are higher in year 2 as well. So the scaling factor will correct for this.
Having said this and double checked the syntax we found a mistake - you need to get the total of the whole wave for weights. Please change into the following syntax (month 12 into 24):

```
ge ind=1
sum ind [aw=b_indpxub_xw] if b_month>=1 & b_month<=24
gen bwtdtot=r(sum_w)
sum ind [aw=c_indpxub_xw] if c_month>=1 & c_month<=24
```

Hope this helps,
Olena

#9 - 09/13/2019 10:28 AM - Gundi Knies

- Status changed from In Progress to Resolved

- Assignee changed from Olena Kaminska to Tom Waters

- Target version set to X M

- % Done changed from 50 to 100