Understanding Society User Support - Support #2167

Using weights for 17 BHPS and 13 UKHLS waves for pooled cross-sectional analysis.

10/23/2024 11:20 AM - Nico Ochmann

Status:	Resolved	**Start date:**	10/23/2024
Priority:	Urgent	**% Done:**	100%
Assignee:	Olena Kaminska		
Category:	Weights		

Description

Dear Peter,

I would like to use all waves of the BHPS (except for the first wave) and all waves of UKHLS for pooled cross-sectional analysis. For that purpose, I need to use weights for each wave.
I am wondering whether the following code would make sense to construct my weights and use in my Stata regression [pw=newwgt].
Please note that wave=1 is bb_ and wave=18 is a_ .

```
gen newwgt = indin91_xw if inlist(wave,1,2,3,4,5,6,7)
replace newwgt = indin99_xw if inlist(wave,8,9)
replace newwgt = indin01_xw if inlist(wave,10,11,12,13,14,15,16,17)
replace newwgt = indinus_xw if wave==18
replace newwgt = indinub_xw if inlist(wave,19,20,21,22)
replace newwgt = indinui_xw if inlist(wave,23,24,25,26,27,28,29,30)
```

I would appreciate a reply from you.
Best,
Nico

History

#1 - 10/23/2024 11:57 AM - Olena Kaminska

Nico,

Yes, this all looks good. Please note, that until 2001 BHPS represented GB excluding NI, and from 2001 NI is included.
Also we suggest to scale weights as descried on p.10 here: https://www.understandingsociety.ac.uk/wp-content/uploads/working-papers/2024-01.pdf .

Hope this helps,
Olena

#2 - 10/24/2024 09:54 AM - Nico Ochmann

Dear Olena,

Your suggestions helped very much, thanks for your prompt and great reply.
I have a bit of a follow-up question if you do not mind me asking.
If I decided to collapse the data by local authority district and year, I would need to apply the weights I have constructed.
The question I have for you is the following:
Would I need to take the raw sum the constructed weights by local authority district and year?
So basically:

```
bysort lad year: egen weightscaled_sum = sum(weightscaled)
```

And then apply:

```
collapse (median) wage [aw=weightscaled_sum], by(lad,year)
```

I would appreciate a reply in case you see what I am trying to do.

Best,
Nico

#3 - 10/24/2024 12:39 PM - Understanding Society User Support Team

- Assignee changed from Peter Lynn to Olena Kaminska

#4 - 10/24/2024 07:57 PM - Olena Kaminska

Nico,

Please explain more what you mean by collapsing? Does it mean that your unit of analysis are not people but lad? If that's the case, how do you create measure for each lad? Technically, if for example you average a measure within each lad and then analyse each lad, you may need weights at both levels separately to first create the average correctly, and then to represent lads. And what do you mean by collapsing by year?
Sum of weights won't be correct here.

Olena

#5 - 10/25/2024 08:19 AM - Nico Ochmann

Hi Olena,
Thanks for your reply, I really appreciate your time.
Let's say I wanted to compare the gender median wage gap in a lad-year panel, I would do the following:
gen wage_m = wage if female==0
gen wage_f = wage if female==1
collapse (median) mwage_m = wage_m mwage_f = wage_f [aw=weightscaled], by(lad year).
This should give me the median for each gender in a given lad and year. I basically have a lad-year panel then.
One question arises now:
1. What weights do I use to analyse the data within the lad-year panel.
Thanks for your help.
Best,
Nico

#6 - 10/25/2024 10:04 AM - Olena Kaminska

Nico,

Yes, year does create an issue as our sample is not random within each year. There is a year of issue and calendar year. To start please read question 13 here:
https://www.understandingsociety.ac.uk/wp-content/uploads/working-papers/2024-01.pdf

Once you created medians per lad, do you just report them, or are you trying to represent lads? The weights would be different if the latter.

Hope this helps,
Olena

#7 - 10/25/2024 12:03 PM - Nico Ochmann

Hi Olena,

Thanks for you reply.
I read question 13, but I think I do not want to go there as it gets too messy.
If I were to replace years with waves to have a lad-wave panel, would that help me?
If yes, I would then just weigh by my previously constructed weightscaled variable to find the medians in a lad-wave panel?
And yes, if I wanted to run some regression based on the medians of other variables in my lad-wave panel, the weights will be different, but how would I construct them?
My apologies for asking all these questions, and please let me know if you cannot answer them. I might then have to post something on the Stata forum.
Best,
Nico

#8 - 10/28/2024 10:41 AM - Olena Kaminska

Nico,

Yes, wave-level analysis is much simpler and does not need additional adjustment for a year-level analysis. I would recommend it.

On your second question. Our weighted data represents people. You can represent people living in all LADs too. And you can analyse LADs as a characteristic of people too.

If you want to run a LAD-level analysis where you want to represent LADs (and not people), possibly with measures of median within each LAD, you need an additional scaling factor to correct for the fact that LADs with higher population are overrepresented in comparison to LADs with lower population. You need then to find a population within each LAD from external official data source (let's call it x). Your newweight=UKHLS_weight/x . The newweight will represent each LAD equally then and will be suitable for LAD level analysis. Note, you can't use newweight for person-level analysis.

Hope this helps,
Olena

#9 - 10/28/2024 02:06 PM - Nico Ochmann

Hi Olena,

Thanks so much for your kind help, I really appreciate it.
I will do the wave-level analysis as you recommend it, it seems doable.
I have one more final comment/question in the hope that I do not waste your time.
With regard to my second question, I understand your suggestion and the fact that I need to construct new weights due to population differences across LADs.
Thanks for pointing this out to me.
However, here is my question/comment. For LAD level analysis, why can't I just include a dummy variable for each LAD (assuming that population does not change much over the waves)?

Again, thanks so much for your time and effort so far.
Much appreciated.
Best,
Nico

#10 - 10/28/2024 02:36 PM - Olena Kaminska

Nico,

Imagine one LAD has 1000 interviews and all others have 1 (e.g. there are 100 others). The first LAD will drive all the values. Also, technically you should use full interactions with all of your LADs, and as there are many of them you may quickly run out of degrees of freedom. Having said that, I suggest that you create an imaginary data with the above specifications and test which method works best.

Hope this helps,
Olena

#11 - 10/28/2024 04:08 PM - Nico Ochmann

Thanks a lot Olena, I will give it some thought.

Have a good evening,

Best,
Nico

#12 - 10/30/2024 01:59 PM - Understanding Society User Support Team

- Status changed from New to Feedback

- % Done changed from 0 to 50

#13 - 02/10/2026 05:57 PM - Understanding Society User Support Team

- Status changed from Feedback to Resolved

- % Done changed from 50 to 100

- Private changed from Yes to No