Understanding Society User Support - Support #839

Pooling data from all waves, 1-6, using all subsamples of USoc (GPS, EMBS, BHPS, IEMB)

08/21/2017 11:54 AM - Nico Ochmann

Status:	Closed	**Start date:**	08/21/2017
Priority:	Normal	**% Done:**	100%
Assignee:	Nico Ochmann		
Category:	Weights		

Description

Dear Peter,

sorry to bother you again, but we had that exchange in support #758.
I have been implementing your suggestions as we discussed. I am contacting you now again because I would really appreciate you having a quick look again at our previous exchange and this current issue. With the weighting scheme, my results change quite a bit (point estimates and standard errors), so I really want to make sure that I am doing things right. I do think I do, but I rather double check.
So I am using all subsamples of all six waves and I generate a new weighting variable accordingly (newwgt). I use these observations across waves as if they were repeated cross sections. For various reasons, I cannot utilize the panel data structure. The time dimension of the pooled cross sections is not the wave variable, but the year variable, istrtdaty, start of the individual interview. Given this information and the discussion we had in support #758, I would highly appreciate your verification of me doing things correctly.
Thank you very much in advance!
Nico

History

#1 - 08/24/2017 04:54 PM - Peter Lynn

- Status changed from New to Feedback

- Assignee changed from Peter Lynn to Nico Ochmann

- Target version set to X M

- % Done changed from 0 to 30

- Private changed from Yes to No

Nico,

I think the time variable might cause you problems. The weight should be fine for a genuine cross-sectional analysis, but introducing a time dimension changes things a bit. Broadly speaking, the samples issued/interviewed in each calendar year 2011 to 2014 are equivalent, but years 2009, 2010 and 2015 are rather different due to the fact that the "year 1" and "year 2" samples have different components.

In 2009 there is only a "year 1" sample, with the result that Northern Ireland will be over-represented and Bangladeshis will be under-represented (and some other rather minor differences);
In 2010 there will be an over-representation of Scotland and Wales, as the wave 1 weights do not reflect the inclusion of the BHPS sample (as this joined the study only at wave 2);
In 2015 there is only a "year 2" sample, so no Northern Ireland at all, under-representation of Scotland and Wales (as they are boosted in the BHPS sample, which is entirely in year 1), and over-representation of Bangladeshis.

I wonder if you might do some kind of sensitivity analysis by restricting your analysis to 2011-14 and comparing with the unrestricted results? Or maybe restrict to England only, where the issues are more minor?

HTH,

Peter

#2 - 08/26/2017 02:41 PM - Nico Ochmann

Dear Peter,

now this reply helps a lot. I have decided to restrict my sample to England only. This should be the cleanest solution.
Two short follow up questions:
First, if I understand you correctly, I leave out the weights altogether when restricting to England.
Second, in order to restrict the sample to England, what is the best way of doing so, taking the hhorig variable and dropping the corresponding subsamples or taking the gor_dv variable and dropping the appropriate regions?

Thank you very much!

Best wishes.

Nico

#3 - 08/30/2017 09:59 AM - Stephanie Auty

- Assignee changed from Nico Ochmann to Peter Lynn

#4 - 09/04/2017 04:18 PM - Peter Lynn

- Assignee changed from Peter Lynn to Nico Ochmann

- % Done changed from 30 to 50

First, no! You should still use weights, even if restricting the analysis to England only. There are big differences in selection probabilities between ethnic groups and in response probabilities between various socio-demographic groups.

Second, hmm, I realise now that my suggestion is not so simple to implement! The weighting problem that you had will only go away if you restrict the sample to people in households that were **sampled** in England. This means selecting based on GOR variable from BHPS wave 1 in the case of BHPS sample, and from Understanding Society wave 1 in the case of GPS and EMBS samples.

A simpler approach would be to select based on current wave GOR. This will not remove the sample imbalance problem, but should reduce it to a minimum, so may be good enough.

#5 - 09/06/2017 09:24 AM - Nico Ochmann

Hello Peter,

what I do now, I take the HHORIG variable and drop subsamples 2,4,5, and 6 and then use the GOR variable to restrict to England.

sample origin	Freq.	Percent	Cum.
ukhls gb 2009-10	211,670	64.85	64.85
ukhls ni 2009-10	10,579	3.24	68.09
bhps gb 1991	32,486	9.95	78.04
bhps sco 1991	7,296	2.24	80.27
bhps wal 1991	8,712	2.67	82.94
bhps ni 1991	9,752	2.99	85.93
ukhls emboost 2009-10	37,411	11.46	97.39
ukhls iemb 2014-15	8,517	2.61	100.00

Finally, and this is my last question for you for now. Why do you not recommend using longitudinal weights given that I have that time dimension?

Best wishes.

Nico

#6 - 09/07/2017 02:35 PM - Stephanie Auty

- Assignee changed from Nico Ochmann to Peter Lynn

#7 - 09/15/2017 04:57 PM - Peter Lynn

- Assignee changed from Peter Lynn to Nico Ochmann

- % Done changed from 50 to 90

Nico,

The non-response pattern in your analysis data will be very different from that in a truly longitudinal (balanced panel) sample, so longitudinal weights would not be appropriate. The weight you have created correctly reflects your data structure (6 cross-sections). For example, many of the cases in your analysis will have longitudinal weights of zero.

Peter

#8 - 09/18/2017 11:07 AM - Stephanie Auty

Email received from user:

Peter,

thank you very much for your patience throughout.

I very much appreciated your help.

Have a nice weekend.

Nico

#9 - 10/02/2017 01:40 PM - Stephanie Auty

- Status changed from Feedback to Resolved

- % Done changed from 90 to 100

#10 - 10/16/2017 04:25 PM - Stephanie Auty

- Status changed from Resolved to Closed