Understanding Society User Support - Support #987

Weighting of sub-sample

06/26/2018 11:51 PM - Ante B

Status:	Resolved	**Start date:**	06/26/2018
Priority:	Normal	**% Done:**	100%
Assignee:	Olena Kaminska		
Category:	Weights		

Description

Dear Sir or Madam,

I would like to compare the means of several variables of a sub-sample (e.g. income, education) after data cleansing with those of the initial sample to test for representativeness of the sub-sample. If all variables are from the same wave (i.e. wave 4 of the UKHLS), cross-sectional weights can be applied. However, the sub-sample contains two variables that were not surveyed in wave 4, so they were carried forward from wave 1 and 3. Should in this case the variables for the comparison be weighted with the longitudinal weights of the last wave (i.e. wave 4) or should cross-sectional weights be used (i.e. cross-sectional weights from wave 1 and 3 for the two carried-forward variables and for the remaining variables, cross-sectional weights from wave 4)? The variables are from household level questionnaires and self-completion interviews, so that the lowest level of hierarchy is 1, which would suggest to use d_indscus_lw if longitudinal weights are appropriate? Do you agree?

Thank you for your help.

Best regards
Ante

History

#1 - 06/27/2018 12:37 PM - Stephanie Auty

- Category set to Weights

- 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,
Stephanie Auty - Understanding Society User Support Officer

#2 - 06/27/2018 03:23 PM - Olena Kaminska

Ante,
Yes, you should use the longitudinal weight for wave 4, and the choice of d_indscus_lw is correct.
Olena

#3 - 06/27/2018 04:01 PM - Stephanie Auty

- Status changed from New to Feedback

- Assignee changed from Olena Kaminska to Ante B

- % Done changed from 0 to 70

#4 - 06/27/2018 04:18 PM - Ante B

Dear Olena,

Thank you very much for your fast clarification.

Best regards,
Ante

#5 - 07/04/2018 11:21 AM - Ante B

Dear Olena,

I have a follow-up question.
If I choose d_indscus_lw to compare the sub-sample with the original sample (population), only matched IDs between Wave 4 and 1 (or 3) will be

accounted for to calculate the population's figures for Wave 1 variables (which are not available in Wave 4). For the unmatched IDs from Wave 1, there is no d_indscus_lw weight so that these are excluded. The population size can thus be considerably reduced. Does d_indscus_lw adjust for these unmatched IDs, or is the use of Wave 1 weights (i.e. a_indscus_xw) then nevertheless the right approach to most accurately describe the population since it will cover all IDs from Wave 1 (the matched and unmatched ones)?

Thank you for your help.

Best regards,
Ante

#6 - 07/04/2018 12:35 PM - Stephanie Auty

- Assignee changed from Ante B to Olena Kaminska

#7 - 07/10/2018 11:51 AM - Olena Kaminska

Ante,

Your problem is not weights but as far as I understand your suggested analysis violates iid assumption - because the subgroups is part of the population. It would be easier to compare the subgroup to everyone else (population excluding the subgroup). You could do this in many ways one of which would be to use data from w1-w4 with the longitudinal w4 weight.

I hope this helps,
Olena

#8 - 07/10/2018 02:26 PM - Ante B

Dear Olena,

Thanks for your comment.

I agree that the subpopulation should be compared with population excluding the subpopulation.

There are indeed many different ways of which weight to use for which wave. The difficulty is to pick the combination that most accurately describes the population: is the accuracy of the population figures distorted more if

1) only matched IDs are used (between wave 1 and 4) with consistently using one weight for all variables (d_indscus_lw) or
2) all IDs are used with inconsistent choice of weights (for w1 variables a_indscus_xw and for w4 variables d_indscus_lw).

Does for this matter exist a clear criteria according to which one can decide what the better approach is?

Best regards,
Ante

#9 - 07/10/2018 03:33 PM - Olena Kaminska

Ante,

Yes, the easiest way is a) - the weight will automatically select the correct IDs - just by using the weight in your analysis you will take care of 'the selection'. b) is not impossible, but you really have to have theoretical reasons to use it, and the statistical software will not do the comparison automatically for you - so you will have to do some calculation by hand which I don't recommend.

Hope this helps,
Olena

#10 - 07/10/2018 04:33 PM - Ante B

Dear Olena,

I appreciate your patience in this matter. Thank you very much for your help.

Best regards,
Ante

#11 - 08/14/2018 05:44 PM - Stephanie Auty

- Status changed from Feedback to Resolved

- % Done changed from 70 to 100