Understanding Society User Support - Support #448

weights

11/16/2015 12:17 AM - Vernon Hedge

Status:	Closed	**Start date:**	11/15/2015
Priority:	Normal	**% Done:**	100%
Assignee:	Olena Kaminska		
Category:	Weights		

Description

I am looking at data exclusively at Wave C Understanding Society c_indresp.sav. I am planning to employ model based inference which may (as needs be) incorporate weight, strata and PSU into the model.

I am having difficulty finding out how the weights were computed. I was hoping to use include the variables by which the weights were calculated within the model and specify PSU as level 2 random effects. I just cannot seem to find how the weights were calculated from Understanding Society documentation.

All the variables are from the c_indresp file. The 12 are listed here as name, “label”, [position number in variable view of c_indresp.sav]

c_sex_cr “sex (corrected)” [2292],
c_age_cr “age (corrected)” [2294],
c_birthy “year of birth” [2771],
c_big5c_dv “Conscientiousness” [2896],
c_big5o_dv “Openness” [2899],
c_hiqual_dv “Highest qualification” [2904],
c_gwri_dv “Cognitive ability: Immediate word recall: Number of correct items” [2915],
c_cgvfc_dv “Cognitive ability: Verbal fluency: Count of correct answers” [2932],
c_cgna_dv “Cognitive ability: Numeric ability: Count of items answered correctly”[2935],
c_jbnssec8_dv “Current job: Eight Class NS-SEC” [2947],

I am also having difficulty identifying which weight variable would be most appropriate to my analysis according to the w_xxxyyzz_aa scheme (p67 of the User Manual).

I can fill in this much c_indyyzz_xw – i.e., I know I am dealing with wave c only (so c_ and xw) and only with adult (16+) respondents (so ind).

I have identified 4 weight variables relevant to a cross-sectional design in the c_indresp file,

1. c_indpxub_xw “combined cross-sectional adult main or proxy interview weight” [3002],
2. c_indinub_xw “combined cross-sectional adult main interview weight” [3003],
3. c_indscub_xw “combined cross-sectional adult self-completion interview weight” [3004],
4. c_ind5mus_xw “cross-sectional extra 5 minute interview person weight” [3005].

The yy component must be either px, in, sc, or 5m. I think I can exclude 5m, as none of the variables on my list is on the list on Table 25 (p56) of the User Manual. Likewise, viewing Table 24 (p53), I think sc can be excluded.

As for the zz component it is tempting to just use "us" (for “understanding society”?). The user guide advises me that the "us" designation refers to “GPS [General Population Sample] and EMB samples” – is this what is meant by “Mainstage”?

Looking at the “Levels of Analysis” in Table 28 (p62), I think I can exclude level 4 “Adult or youth self-completion”. I cannot, however seem to find information on whether the c_indresp variables I am using are level 3 “Adult proxy and main interview” or level 2 “Adult main interview only (no proxy)”. Using the Understanding Society website to search each variable name they all return “Mainstage Variable”. I cannot tell from this which level of 1 to 4 is the most appropriate to select a weighting variable.

So the two problems I have are 1) identifying which variables were used to calculate survey weights and 2) identifying which ”xw“ survey weight variable is most appropriate to my analysis.

I would be enormously grateful for any clarification.

History

#1 - 11/16/2015 09:13 AM - Redmine Admin

- Category set to Weights

- Target version set to M3

#2 - 11/16/2015 09:13 AM - Redmine Admin

- Assignee set to Olena Kaminska

#3 - 11/16/2015 12:29 PM - Olena Kaminska

Dear Vernon,

Thank you for your question. Your thinking of selecting the weight for your analysis are all correct. Additional information on which instrument the questions were asked can be found either in the questionnaire:
file:///C:/Users/olena/Downloads/Wave_3_Questionnaire_Consultation.v02.pdf
or https://www.understandingsociety.ac.uk/documentation/mainstage/dataset-documentation
For example I notice that you use derived variable c_big5c_dv. On the documentation it says that it is calculated using C_SCPTRT5C1 C_SCPTRT5C2 C_SCPTRT5C3
(see https://www.understandingsociety.ac.uk/documentation/mainstage/dataset-documentation/wave/3/datafile/c_indresp/variable/c_big5c_dv).
In the table it has -7 for proxy which means it is not asked for proxy interviews.
Looking at the questionnaire for C_SCPTRT5C1 we find it is in self-completion part and also has a text next to it 'Mode is face-to-face and has agreed to self-completion', which means the question is asked in self-completion mode.
Thus, in your situation you need self-completion cross-sectional weight.
'us' weight will be fine, but I recommend 'ub' weight which combined BHPS, GPS and EMB and therefore gives you highest sample size.
The weight therefore will be c_indscub_xw.

The weight c_indscub_xw is calculated using c_psnenub_xw, which in turn is calculated using c_psnenub_lw, which in turn is calculated using b_psnenub_lw, which in turn is calculated using a_psnenus_xw and BHPS wave 18 longitudinal weight. All of the technical details are described in the technical part of the weight description in the user guide. Thus the weights are calculated at many stages using over 200 variables from different waves as well as from census and other geographical information which was linked to our dataset at postcode and other levels at wave 1 - such detailed information is not usually available to users without special permission.
Also, note that if you want to run analysis without weights please make sure to at least use design weights to account for unequal selection probabilities.

Please let me know if you have any further questions,
Olena

#4 - 11/16/2015 03:19 PM - Redmine Admin

- % Done changed from 0 to 50

#5 - 11/17/2015 12:53 AM - Vernon Hedge

Thank you so much, Olena, for that very detailed and considerate reply!

It is my first attempt at negotiating survey data and your explanation and links have been very educational. I had been struggling over the meaning of the proxy -7 designation, in general, for a while.

I am now very clear on the appropriate selection of c_indscub_xw for a weight variable. Thank you.

The information you give about how the weight variable c_indscub_xw was calculated tells me I am probably unable to take a model-based approach. I had hoped to fit a 2-level regression model where PSU (c_psu) was to be the level 2, random effects. And I hoped to have included, as explanatory variables, the variables which c_indscub_xw was computed from. I was hoping that they might have been straightforward ones like "age" or "gender" and, perhaps, the rural/urban indicator (c_urban_dv). A little optimistic, perhaps, as the first 2 were variables I needed to include anyway, for substantive/theoretical reasons. Wow, my optimism has brought me up a long way short (as usual).

Concerning the alternative route, and your clear and strong suggestion that I "use design weights to account for unequal selection probabilities". In this context does "design weights" mean the reciprocal of the relevant weight variable (i.e, 1/c_indiscub_x)? If I am to include this as an explanatory variable in the regression equation, is it wise to keep PSU (c_psu) as the level 2, random effects, or would the household identifier (c_hidp) be the wiser choice for the level 2, random effects?

If I am not (because now I am unable) hoping to generalise my findings to the UK population of 2011, would using the design weights, in this manner, make any estimates in my model more unbiased for the sample? From a variable that has ~13% valid observations within a relevant subsample, and following a series of exclusions given my study criteria, my sample size currently stands at ~3.2% (~600 respondents) of the initial subgroup of interest . I imagine there are all sorts of hidden biases that have resulted from these factors. I guess I don't want to make any claim that is too bold, considering all this, but I want to be clear about what using design weights would allow me to claim.

I, again, would be very grateful for your thoughts.

Thank you,

Vernon

#6 - 11/17/2015 12:57 AM - Vernon Hedge

In paragraph 5 "(i.e, 1/c_indscub_x)" should read "(i.e., 1/c_indscub_xw). Apologies.

#7 - 11/17/2015 11:22 AM - Olena Kaminska

Vernon,

We suggest where possible to use full information on design, including psu, strata and weights. It is possible to run random effects model with weights in Stata or MlWin. If you can do this, use c_indscub_xw and psu as your clusters.

Design weight is part of c_indscub_xw or any other main weight that we release. Basically each weight consists of correction for selection probabilities (design weight) and for nonresponse. We provide design weight for advanced users who may want to correct for nonresponse within the model or develop model-specific nonresponse correction.

From my limited research (not yet published) including either weight in the model as a control variable or using variables from weighting model as control variables does not help to decrease nonresponse error: the effect is unpredictible - sometimes this decreases the bias but sometimes it increases it. I am looking for references where this is suggested and will appreciate if you can share it with me. Intuitively this makes sense as in the model one uses only information from respondents while in nonresponse correction one uses information on nonrespondents as well. Simply put it is not easy (at least without strong assumptions) to correct for nonresponse without any information on nonrespondents.

I can't comment on your sample size much, except that 600 is a very good size to find important effects. If you think that some information is missing on item level where it should not have been consider correcting for missingness, possibly through imputation.

Hope this helps,
Olena

#8 - 11/17/2015 02:33 PM - Vernon Hedge

Thanks again, Olena,

With a fast encroaching penultimate deadline for a thesis, I guess it is too late for me to set up and learn how to run MLwiN on my Apple computer. Though I will after submission to see if I can get a weighted models for my final submission. I only have R and SPSS. I may use the survey package in R, but I understand it has some limitations that may mean I can't fit the models I want to. Anyway.

To be clear, you are saying that psu, and psu alone, would be the clusters?

From my explorations of this area, I believe that using variables from the weighting model as control variables was accomplished in this paper (Hanandita & Tampubolon, 2015), which should have citations concerning the validity of this method.

There is also an interesting related method concerning the decision to use survey weight in an imputation model. It comes from Penn State Methodology Centre. Survey weight is (OLS) regressed on the design variables. If the resulting R^2 values are high, this is used to infer that most of the important design variables have been included and, therefore, weights are not needed for the imputation model.

They go on to suggest a nice trick for limiting selection bias - by no less than 90%! - in the imputation model (they cite Rosenbaum & Rubin, 1983).

Apologies if this is not completely helpful or relevant, or it is information you already have. I am new to all this.

Anyway, thanks again, Olena!

Vernon

Hanandita & Tampubolon (2015) http://link.springer.com/article/10.1007/s11136-015-1152-y

Penn State, The Methodology Centre, https://methodology.psu.edu/eresources/ask/sp06

Rosenbaum & Rubin (1983) http://biomet.oxfordjournals.org/content/70/1/41.full.pdf+html

#9 - 11/18/2015 10:18 AM - Olena Kaminska

Vernon,

Thank you for your citations. And yes, you should use psu alone as clusters. This is because you are using only one wave - and these are the only clusters you have.
If you have multiple observations per person theoretically you would use psu as level 3, individual as level 2 and observation as level 1 (in those software that allows more than 2 levels).

Best,
Olena

#10 - 11/18/2015 12:41 PM - Vernon Hedge

And thank you, again, Olena.

Vernon

#11 - 11/20/2015 09:23 AM - Redmine Admin

- Status changed from New to Closed

- % Done changed from 50 to 100