Understanding Society User Support - Support #716

Weighting linked USoc-NPD dataset

02/07/2017 01:12 PM - Matt Barnes

Status:	Closed	**Start date:**	02/07/2017
Priority:	High	**Due date:**	
Assignee:	Matt Barnes	**% Done:**	100%
Category:	Weights	**Estimated time:**	0.00 hour
Target version:			

Description

I'm exploring the wellbeing of secondary school children with a Special Educational Need (SEN) using the linked Understanding Society (USoc) – National Pupil Database (NPD) dataset. The linked dataset contains NPD information (including SEN) matched to children who were part of USoc in wave 1 , living in England (NPD is only for English schools), whose parent gave consent to matching. Because the actual data matching took place a couple of years after consent was given, most of the NPD data is from a couple of years after USoc wave 1 (USoc wave 1 is from 2009/10, most of the NPD data is from 2012/13).

A child's SEN status can change year on year, so to ensure a child's SEN status is recorded from a time period near to their wellbeing data, I've decided to use wellbeing data from USoc wave 3 (2011/12). This is a year before most of the NPD data, which isn't ideal but it's the closest wave of USoc that contains the wellbeing data we need (some of which is only collected every other year in USoc).

So my analytical sample is children who were part of USoc wave 1, who completed the youth self-completion questionnaire in wave 3 (wellbeing data) and who have matched NPD data (SEN data). I need to create a weight that accounts for the various types of non-response that can happen to this sample. So if the starting group is [1] children living in England aged 8-13 in 2009/10 (USoc wave 1), [2] some non-respond to the initial survey (there's a cross-sectional weight for this on the survey), [3] some drop out of the USoc survey by 2011/12 (wave 3, the wave we're using the wellbeing data from) (there's a longitudinal weight for this on the survey), [4] of the remaining some don't complete the self-completion survey (there's a weight for this on the survey), [5] of the remaining some parents refused consent to matching (there's no weight for this), [6] of the remaining some didn't get matched to the NPD (most of data from which is from 2012/13)(there's no weight for this).

My approach has been to identify children up to and including stage [4] above, and create a weight that readjusts these back to the initial population [1]. This is created by multiplying the wave A-C longitudinal weight (c_psnenus_lw) by the wave C youth self-completion weight (c_ythscub_xw) i.e. new weight = c_psnenus_lw X c_ythscub_xw. I then need an additional component of the weight that adjusts for children whose USoc data did not get matched to the NPD. This is achieved by running a logistic regression model for the sample at stage [4] with the dependent variable indicating whether they were matched to the NPD or not. The logistic regression analysis uses the following predictor variables: age of child, sex of child, ethnicity of child, household income, family work status, highest educational qualification of parents, rurality, government office region. The logistic regression analysis is weighted by 'new weight' (above). The 'NPD weight' is calculated as the inverse of the predicted probabilities. This has been scaled so the mean is 1 and trimmed. The final weight is then 'new weight' X 'NPDweight'.

Any comments gratefully received! Thank you.

History

#1 - 02/08/2017 09:48 AM - Victoria Nolan

- Status changed from New to In Progress

- Assignee changed from Olena Kaminska to Matt Barnes

- % Done changed from 0 to 10

- Private changed from Yes to No

Dear Matt,

Many thanks for your enquiry, I have passed it on to our weighting team.

Best wishes, Victoria.

#2 - 02/08/2017 12:34 PM - Peter Lynn

- % Done changed from 10 to 80

Matt,

I suggest a different approach. Let me first describe the approach and then explain why.

Start with the sample of all relevant young people (8-13 at wave A, 10-15 at wave C) who were observed (enumerated) at both waves A and C. The base weight for analysing this sample would be c_psnenus_lw.

Then develop a log reg model based on this sample, in which the dep var is whether the person is available for your analysis (both wave C self-comp data and NPD data). If predicted values from this model are Pi, then your analysis weight is c_psnenus_lw/Pi.

I think this is the most parsimonious approach, as the first step deals simultaneously with your [1]-[3], while the model-based adjustment deals with your [4]-[6]. An alternative would be to model each of [4], [5] and [6] as dependent steps and make three multiplicative adjustments. This could be helpful if the predictors are very different, but would have the disadvantage of adding additional variance to your estimates, so i doubt that this would be worthwhile.

Note that you cannot multiply together provided weights in the way you have suggested, as each analysis weight includes all the previous stages (each is designed to be used on its own). You would be, for example, correcting twice for differences in selection probabilities, so you would end up greatly under-representing ethnic minorities and people in Northern Ireland.

Given the above, you might think that you could use c_ythscub_xw as a base weight, and then model for [5] and [6] based on respondents to the wave C youth self-completion, but this would also be incorrect as you are restricting your analysis to the wave A ("us") sample, whereas c_ythscub_xw is for analysis that includes also the BHPS sample (i.e. the "ub" sample). So, you would only be able to do this starting from a weight, c_ythscus_xw, which is not one that has been developed!

Also, your predictor variables look good, though i wonder about including also ethnicity of parent? It may be too strongly correlated with ethnicity of child though, so you should probably only include whichever of the two is a stronger predictor. You may also want to check this paper for other strong predictors of consent: Al Baghal, T. (2016) Obtaining data linkage consent for children: factors influencing outcomes and potential biases, International Jnl of Social Research Methodology 19: 623-643.

Hope that helps,

Peter

#3 - 02/08/2017 03:37 PM - Matt Barnes

Hi Peter

Thank you very much for your quick and helpful reply. What you suggest makes perfect sense. Yes I did wonder whether multiplying those two variables together made sense! And yes, I'll look at that paper as you suggest.

Thanks again.

Best wishes

Matt

#4 - 02/16/2017 09:27 AM - Victoria Nolan

- Status changed from In Progress to Closed

- % Done changed from 80 to 100