Additional questions for ISER

Dan Brown[*]

May 2016

1. **Performing weighted t-tests when using only a small sub-sample of observations from the UKHLS:**

    - I am considering a very specific sub-sample of individuals of interest from the UKHLS dataset (a total of approximately 1,300 individuals)
    - I want to perform some t-tests to compare the mean values of a range of variables between two sub-groups within this sub-sample.
    - I can perform unweighted t-tests, but I would like to perform weighted t-tests to take into account sampling weights which reflect the sample design of the UKHLS. I would like to do this with the svy: mean command in Stata, having first svyset the data as instructed in the user guide (with my appropriate choice of weight).
    - The problem is that in my sub-sample of individuals of interest, the majority of strata include only a single primary sampling unit (and often these primary sampling units contain only a single individual), and so Stata cannot compute standard errors when using svy commands (it returns the message: 'missing standard errors because of stratum with single sampling unit').
    - I.e. once I have restricted the dataset to the small number of individuals of interest, there are very few individuals (and primary sampling units) in each strata.
    - Do you know how I should use svy commands in Stata if I have restricted my sample to a small sub-sample of interest where many strata contain only a single primary sampling unit? Most importantly, how I might be able to perform weighted t-tests that use the sampling weights UKHLS provides?

[*]Department of Economics, University of Oxford. Email address: daniel.brown@lincoln.ox.ac.uk.

2. **Understanding strata in the sample design:**

    - I have read the User Guide, which explains the sample design. However, there appears to be a difference between the number of strata described in the user guide, and the number of strata that exists in the dataset, which I am sure is because I have misunderstood something:
    - The User Guide says that for the GPS sample there are 108 sub-strata for England, Scotland and Wales (12 strata, each divided into 3 based on proportion of households in non-manual labour, which are in turn divided into 3 based on population density). Northern Ireland is treated as its own strata.
    - For the EMB sample, there are 4 strata.
    - So as far as I can see there are 12 strata for E/S/W, 1 for Northern Ireland, and 4 for the EMB sample. These are divided into sub-strata, but the total number of sub-strata is still presumably a little over 100 (I can't find the number of sub-strata listed for the NI or EMB samples).
    - So what I don't understand is why there are 1,776 'strata' in the dataset (I am looking here at wave 'a'). This sounds like it must be a different definition of strata to that described in the user guide - so what do these strata (those separate numbers under the variable a_strata) correspond to?

3. **Longitudinal weight choice with an unbalanced panel.**

    - Suppose that I want to undertake longitudinal analysis and include individuals who later leave the sample, such that I have an unbalanced panel.
    - I understand that the longitudinal weights calculated are only appropriate for balanced panel datasets, but suppose I have found some way of dealing with attrition from the sample in my model.
    - Am I correct to use the appropriate longitudinal weight from the first wave in my analysis?
    - So if I am using the GPS + EMB samples from wave 1 onwards in a longitudinal analysis (and the lowest level of 'question' I am using comes from the adult main interview (no proxies)), then I want to use the b_indinus_lw weight for my analysis. The indinus is the appropriate choice according to the user guide, and this weight is only available from wave 'b' onwards. So I then apply this weight to every observation over time for each individual, and this weight

should account for differential probabilities of selection into the sample at the start of the sample period, even though it does not account for subsequent attrition of some members of the sample over time?

4. **Weights of zero in 'b_indinus_lw':**

   - There are many individuals for whom the b_indinus_lw weight takes a value of zero.
   - As far as I can work out, the only reason an individual has a longitudinal weight of zero from waves b onwards is if: a) The individual is a TSM - i.e. the individual was a member of an EMB household interviewed in wave 'a', but was not of the targeted ethnic group. All the while they remain in the same household as the EMB sample member, they are still interviewed (and so still appear in the dataset in later waves), but the weights calculated in the UKHLS are not intended to cover these individuals as they are not really meant to be part of the sample, and so they receive a weight of zero. Or b) The individual was not present in one or more of the waves after the first wave.
   - Is that understanding correct?