Pooling data from different waves of *Understanding Society* for cross-sectional analysis

Peter Lynn & Olena Kaminska 27-01-2016

Data from different waves can be easily combined for cross-sectional analysis, provided that each of the 24 monthly samples is included in the analysis base an equal number of times.

For example, for analysis relating to a calendar year the wave *n* year 1 sample can be combined with the wave *n-1* year 2 sample.

A similar approach can be used for any other 12-month period. For example, for a financial year (April to March), months 4 to 15 from wave *n* can be combined with months 16 to 24 from wave *n-1* and months 1-3 from wave *n+1*. And equivalently for any other period that is a multiple of 12-months.

All variables involved in the analysis must be pooled from the respective waves. This includes the weight variable. We strongly recommend that a non-zero value of the weight variable is used to define the analysis base (see example below).

Box 1: Example syntax for pooled analysis for cross-sectional estimation relating to calendar year 2011

```
* Use year 2 of wave 2 and year 1 of wave 3
* Objective is to estimate distribution of jbstat

use "\\....\b_indresp.dta", clear
merge 1:1 pidp using "\\....\c_indresp.dta"

ge jbstat2011=0
replace jbstat2011=b_jbstat if b_month>=13 & b_month<=24
replace jbstat2011=c_jbstat if c_month>=1 & c_month<=12

ge weight2011=0
replace weight2011=b_indpxub_xw if b_month>=13 & b_month<=24
replace weight2011=c_indpxub_xw if c_month>=1 & c_month<=12

ge psu2011=0
replace psu2011=b_psu if b_month>=13 & b_month<=24
replace psu2011=c_psu if c_month>=1 & c_month<=12

ge strata2011=0
replace strata2011=b_strata if b_month>=13 & b_month<=24
replace strata2011=c_strata if c_month>=1 & c_month<=12

svyset psu2011 [pw=weight2011], strata(strata2011) singleunit(centered)
svy: proportion jbstat2011 if weight2011>0
```

Starting from the next release (wave 6) we hope that this will be all you will have to do.

However, at present there is an additional wrinkle. The weights provided currently are not designed for pooling as they are scaled to a mean value of 1.0 within each wave, and therefore produce different weighted sample sizes in each wave. As a result, cases from later waves will be under-represented. This matters because each monthly sample is not a random subset. In the example of box 1, sample months 1 to 12 will be under-represented (as we take their data from wave 3, rather than wave 2 for months 13 to 24)[1]. To overcome this, we should scale the weights for these cases to give the same weighted total that this sample had at wave 2. (Or we could equivalently scale the weights for the months 13 to 24 sample to equal their weighted total from wave 3.) Thus, the syntax becomes that in box 2.

This rescaling becomes even more important when pooling data from more than one 12-month period (e.g. two calendar years). In that case, in addition to the imbalance between the 24 monthly samples, the relative contribution to the estimate (weighted sample size) will also tend to be less for the later year(s) unless rescaling is done, such that each year contributes equally to the estimate. This is achieved by scaling all of the weights to the relevant weighted totals from one common wave.

Box 2: Example syntax for pooled analysis for cross-sectional estimation relating to calendar year 2011, with weight re-scaling

```
use "\\....\b_indresp.dta", clear
merge 1:1 pidp using "\\....\c_indresp.dta"

ge jbstat2011=0
replace jbstat2011=b_jbstat if b_month>=13 & b_month<=24
replace jbstat2011=c_jbstat if c_month>=1 & c_month<=12

ge weight2011=0
replace weight2011=b_indpxub_xw if b_month>=13 & b_month<=24
ge ind=1
sum ind [aw=b_indpxub_xw] if b_month>=1 & b_month<=12
gen bwtdtot=r(sum_w)
sum ind [aw=c_indpxub_xw] if c_month>=1 & c_month<=12
gen cwtdtot=r(sum_w)
replace weight2011=c_indpxub_xw*(bwtdtot/cwtdtot) if c_month>=1 & c_month<=12

ge psu2011=0
replace psu2011=b_psu if b_month>=13 & b_month<=24
replace psu2011=c_psu if c_month>=1 & c_month<=12

ge strata2011=0
replace strata2011=b_strata if b_month>=13 & b_month<=24
replace strata2011=c_strata if c_month>=1 & c_month<=12

svyset psu2011 [pw=weight2011], strata(strata2011) singleunit(centered)
svy: proportion jbstat2011 if weight2011>0
```

[1] As a result, Northern Ireland will be under-represented (as the Northern Ireland sample is entirely in year 1), Bangladeshis and, to a lesser extent, Indians and Pakistanis, will be over-represented (as these groups were boosted more in year 2 than in year 1) and recent immigrants will be over-represented (as these are largely missing from the BHPS sample, which is entirely in year 1).

Note:

DO NOT use ONLY the year 1 sample, or ONLY the year 2 sample.

Do not create analyses bases that are not either

a) a multiple of 12 complete months of data collection (and therefore a multiple of all 24 months of sample), or
b) a multiple of whole waves of data collection (and therefore a multiple of all 24 months of sample)

The analysis sample is only representative when all 24 monthly samples are combined in equal measure.