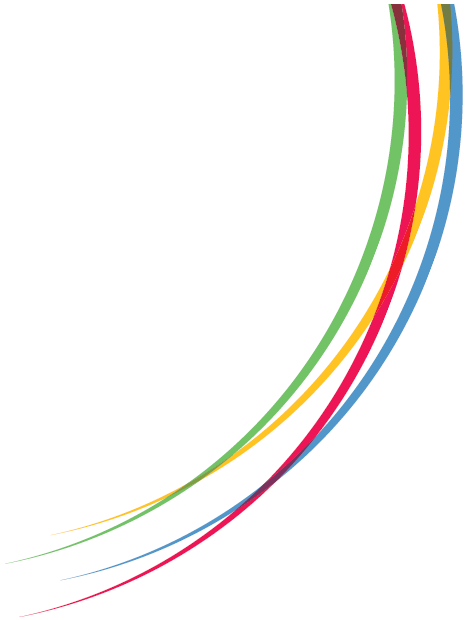
Understanding Society
THE UK HOUSEHOLD LONGITUDINAL STUDY



# Weighting and Sample Representation:

# Frequently Asked Questions

Olena Kaminska
Peter Lynn
Institute for Social and Economic Research
University of Essex
Colchester
Essex

V1.0
October 2019

University of Essex

ISER INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

UKRI Economic and Social Research Council

# Weighting and Sample Representation: Frequently Asked Questions

## 1. Which population does *Understanding Society* represent?

*Understanding Society* can be used in different ways, to represent several different populations. You can represent the cross-sectional population (those currently resident in the country) in any year since 1991 or the longitudinal population over a series of years (those continuously resident in the country over a period of time). You need to identify the appropriate data files and the appropriate weight to use, depending on the population you wish to represent. There are some important points to note:

From 1991 to 2000, the Study only covered Great Britain (England, Scotland and Wales). It was extended to Northern Ireland in 2001. Consequently, you can represent:

- the cross-sectional population of Great Britain in any year since 1991;
- the longitudinal population of Great Britain over any period of years since 1991;
- the cross-sectional population of the United Kingdom in any year since 2001;
- the longitudinal population of the United Kingdom over any period of years since 2001.

However, a much larger sample size is available from 2009-10 onwards, when data collection from the main *Understanding Society* samples (General Population Sample and Ethnic Minority Boost Sample) started, so longitudinal analysis starting at this point can be particularly valuable for the study of small subgroups or rare events.

Due to the sampling methods used, some recent immigrants are excluded from several of the possible reference populations. The only populations with no such under-coverage are Great Britain in 1991, Wales and Scotland in 1999, Northern Ireland in 2000, UK in 2009-10 and in 2014-15:

- The data collected between 1992 and 2008 in England exclude households consisting entirely of recent (since 1991) immigrants.
- In Wales and Scotland, data collected between 1992 and 1998 exclude households consisting entirely of immigrants since 1991 and data collected between 2000 and 2008 exclude households consisting entirely of immigrants since 1999.
- In all countries of the UK, data collected between 2010/11 and 2013/14 exclude households consisting entirely of immigrants since 2009/10, and data collected since 2015/16 exclude households consisting entirely of immigrants since 2014/15.

## 2. Can I represent a subpopulation?

Yes, you can represent any subpopulation of any of the populations described in the answer to Q1, provided it is defined by substantive variables. If you use appropriate analysis methods for the relevant population (see Q3), but restrict your analysis to members of the subpopulation, your results will be representative of the subpopulation.

Examples of subpopulations that you can represent:

- Residents of Northern Ireland
- Females in full time employment
- Babies born in the last 12 months
- Conservative voters in the 2017 election
- Males aged between 17 and 29 who hold a driving license

## 3. What do I need to do to represent a population or a subpopulation?

UKHLS is a probability survey with a complex design, but it is easy to take the design into account and obtain results that represent the population. For this you need to specify clustering, weight and stratification. In Stata use the `svyset` command. Example:

```
use [...]a_indall.dta
svyset a_psu [pweight = a_psnenus_xw], strata(a_strata)
svy: tabulate a_ethn_dv
svy: logistic a_single_dv a_dvage
```

## 4. Are sample sizes adequate to represent ethnic minorities or immigrants?

In 2009-10 (wave 1 of *Understanding Society*) data were collected for the first time from an ethnic minority boost sample which was designed to provide substantially boosted sample sizes for the following subgroups: Indian, Pakistani, Bangladeshi, Afro-Caribbean and black African.

From 2014-15 (wave 6) we further boosted the five ethnic minority groups listed above and also added a boost of immigrants (i.e. persons born outside of the UK). If you are interested in immigrants other than of the five ethnic groups listed above you may want to start your analysis from wave 6.

These subgroups are also asked additional questions, referred to as the “extra 5 minutes” questionnaire. These additional questions are also asked of a small random

subsample of the general population sample which can be used to compare findings for ethnic minority groups to the total population. For this use the weight *w*_ind5mus_*aa* (see Q6).

## 5. Why should I use weights in my analysis?

If you don't use weights your analysis will not correctly reflect the population structure, as some groups are over-represented in the sample by design, while some groups are more likely to respond than others. For example, we have over-sampled ethnic minorities and residents of Northern Ireland. If a statistic of interest differs between groups which are over or under represented in the sample, then unweighted estimates of that statistic will be biased. For example, if education is a stronger predictor of a certain health outcome amongst some ethnic minority groups than in the white British population then unweighted analysis will over-estimate the strength of this association in the population. An unweighted analysis does not correctly reflect the population structure. The weights correct for unequal selection probability, nonresponse at wave 1, sample attrition at subsequent waves, and include a slight correction for a sampling error. These corrections are important.

## 6. Which weight should I use for my analysis?

There are a number of weights reflecting the complex structure of the data. The weight name has the following structure: w_xxxyyzz_aa. To select a weight please answer the following questions:

1. **_aa** part: Is your analysis longitudinal or cross-sectional?

- Longitudinal _lw
- Cross-sectional _xw

2. **w_** part:

- if your analysis is cross-sectional – which wave you are using? e.g. wave 8: h_
- if your analysis is longitudinal – which is the last wave in your analysis? e.g. you are looking at wave 1-9: i_

| Wave: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Prefix: | A | b | c | d | e | f | g | h | i |

3. **xxxyy** part: Is your analysis household level or individual level?

- If it is household level: _hhden
- If it is individual level see below

4. **xxxyy** part: Is your analysis for all persons aged 0+, for youth (10-15) or for adults (16+)?

- 0+ population: _psnen
- Youth (10-15): _ythsc
- Adults (16+): see below

5. **xxxyy** part: you are studying adults aged 16+. Where does your data come from?

- Just one survey instrument (e.g. individual questionnaire): use the weight indicated on the appropriate row of the table below

  A combination of instruments: use the weights from the lowest level in the table below.

| Level of Analysis | Data source | _xxxyy |
|---|---|---|
| **5** | Household grid and/or household questionnaire | _psnen |
| **4** | Adult proxy and main interview | _indpx |
| **3** | Adult main interview only (no proxy) | _indin |
| **2** | Adult self-completion interview | _indsc |
| **2** | Extra 5 minutes interview | _ind5m |
| **2** | Youth questionnaire | _ythsc |
| **2** | Nurse visit | _indns |
| **1** | Blood sample | _indbd |

For example, if you are using information from the household grid and self-completion questionnaire, the levels are respectively 5 and 2 with 2 being lower – hence the weight will be for self-completion data (_indsc ). Similarly, if you are combining information from household grid, adult main interview and nurse visit, your lowest level is 2 so the weight will be _indns.

There will be situations when you combine information from different instruments at the same level: an example would be adult self-completion interview and nurse visit. In this situation we do not have an optimal weight for you and you could use either a suboptimal weight (see Q14) or you can create a weight adjustment tailored to your analysis (see Q15).

6. **zz**_ part: what is the timeline of your research?

- Starting at wave 6 (2014-15) onwards: ui_
- Starting between wave 2 (2010-11) and wave 5 (2013-14): ub_
- Starting at wave 1 (2009-10): us_
- Starting at any point between 2001 and 2008: 01_
- Starting at any point between 1991 and 2000: 91_

## 7. What happens if I don't use a weight?

You implicitly assume that sample members have equal probabilities of selection and of response. This is not true. See Q5.

In addition to biasing your estimates, unweighted analysis will tend to systematically underestimate standard errors. Consequently, confidence intervals will be downwardly biased (too narrow) and models will be over-fitted.

Taking account of selection probabilities and response probabilities using a method other than weighting is very challenging for *Understanding Society*, because of the complex nature of the sample design and the complexity of non-response patterns (multiple waves, instruments and dependencies in data collection).

## 8. Will it be sufficient to include a weight variable in my regression model as a control variable?

Simply using a weight as a control in a regression analysis is not sufficient to take into account the complexities of sample design and correct for nonresponse. This would only suffice if variations in selection and response propensities affected only the dependent variable directly, and not the relationship between dependent and predictor variables in the model. If relationships are affected, then interactions between the weight variable and each other predictor should also be included: this soon becomes unwieldy and statistically inefficient.

## 9. What happens if I don't correct for clustering?

Taking sample clustering into account is simple to do in most standard statistical software for most kinds of estimation (see Q3). However, if you do not do this, while your estimates are not affected, associated standard errors will tend to be under-estimated – sometimes considerably so – resulting in biased hypothesis tests and over-fitting of models.

## 10. What happens if I don't correct for stratified sampling?

Taking the stratified nature of the sample design into account is simple to do in most standard statistical software for most kinds of estimation (see Q3). However, if you do not do this, your estimates are not affected, but associated standard errors will tend to be slightly over-estimated. This makes your analysis slightly conservative, which is often acceptable.

## 11. Can I run analysis on a calendar year / month?

Yes, it is possible to run analysis relating to a calendar year or month with a few extra adjustments. The survey sample is designed such that each sample month (identified by the variable w_month) is a random representative (once weighted) sample of the population with some exceptions:

- Northern Ireland is only present in months 1-12 (first year of each wave)
- BHPS is only present in issue month 1-12 (first year of each wave)
- The IEMB sample is only present in issue month 13-24 (second year of each wave)

Because of this we recommend use of the us_lw weight in analysis, including for cross-sectional estimates. This weight correctly excludes BHPS and IEMB.

Please also note that if you use months 13-24 you are excluding Northern Ireland from your analysis. If you use months 1-12 Northern Ireland will be over-represented without an additional adjustment to the weight. Here is the Stata syntax for adjustment if you use month 1-12:

```
gen adj=1

replace adj=0.5 if w_country==4

gen weight=w_xxxyyus_lw*adj
```

We suggest that you use sample month / year (`w_month`) to identify the analysis sample rather than month / year of interview. For each sample month, interviews take place over 3-4 months, but the majority of interviews take place in the calendar month coinciding with the sample month. The interviews that come in later calendar months tend to be with sample members who are either hard to contact or reluctant to participate. Our weights are designed for each whole sample month to represent the population. If you omit the interviews from the calendar months following the sample months you will be excluding a category of respondents who tends to be very different to earlier respondents, so it is unlikely that your analysis sample will remain representative.

If you still want to define your analysis sample by month / year of interview (rather than sample month) there are two ways you can adjust for the late respondents:

- Create a tailored adjustment to our weight (see Q15)
- Use late respondents from other issue months with our weights (see below).

Let's say you are interested in studying December 2014. Your optimal option with the largest sample size will be to combine all interviews carried out in December of 2014 from the following samples:

- Wave 5 sample months 21, 22, 23 and 24
- Wave 6 sample months 9, 10, 11 and 12
- Create a new variable that equals `e_xxxxxus_zz` weight for the wave 5 interviews and `f_xxxxxus_zz` weight for wave 6. No Northern Ireland adjustment is needed. No extra nonresponse adjustment is needed as late respondents in the month 24 sample are compensated for by bringing in the late respondents from previous sample months. But you will need a scaling factor (see Q12*).*
- Use psu and strata variables from `xwave.dat` to take into account clustering and stratification.

Note if you want to study January 2014 for example, the information will come from 3 waves, because to compensate for missing of late respondents from wave 5, sample month 1, you will need to include January respondents from wave 4, sample months 22-24. The rest will follow the above example.

If you use respondents from calendar months / year just from one wave you will need an extra adjustment for Northern Ireland and potentially also for late respondents (if your period of interest includes sample months 1, 2 or 3).

## 12. Can I pool data from different waves for cross-sectional analysis?

Data from different waves can be 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).

However, the weight requires an additional adjustment. This is because each weight is scaled to a mean value of 1.0 within each wave, and therefore produces a different weighted sample size in each wave. As a result, cases from later waves will tend to be under-represented when pooling waves, unless the weight is adjusted. This matters because each monthly sample is not a random subset. For example, if we pool sample months 1 to 12 from wave 3 with sample months 13 to 24 from wave 2, the former will be under-represented (as the responding sample size is smaller at wave 3 than at wave 2)[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.) Stata syntax to do this re-scaling is shown in box 1 below.

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.

Note:

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

Do not create analyses bases that are not either

[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).

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.

**Box 1: 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
```

## 13. Should I worry that some members of my analysis sample have a weight of zero?

There are some sample members who have provided data from a combination of waves/instruments for which they have a weight of zero. This is not a mistake: these weights are zero for one of two reasons: there are zero weights by sample design, and there are zero weights as a result of fieldwork issuing rules.

Temporary Sample Members (TSMs) are not strictly part of the *Understanding Society* sample. Data are collected from these persons in order to provide contextual information regarding the sample members (OSMs). Consequently, all longitudinal weights will always be zero for all TSMs. For example, any TSMs who participated in waves 6, 7 and 8 will have a zero value of `h_indinui_lw`, even though this weight is designed for longitudinal analysis starting at wave 6.

Cross-sectional weights are not necessarily zero for TSMs, but there is a sub-set of TSMs for whom these weights too are always zero, as a consequence of the sample design. These people have zero design weights as well. They are 'TSMs from wave 1' selected through the EMB and IEMB boosts. These are non-eligible people (e.g. white British) who are co-residents with eligible people (e.g. ethnic minorities or immigrants) at the first wave of sample selection. Their zero weights are correct and are related to the sample design. The sample design was implemented in the most cost efficient way to meet the need for analysing ethnic minorities, recent immigrants and the whole of UK population. Avoiding these zero weights would have had a substantial cost implication but would have added very little to the precision of estimates.

There are also zero weights which result from a fieldwork issuing rules. If all nonresponding households that missed the previous wave were not issued to the fieldwork and were dropped you would not observe zero cross-sectional weights other than 'TSMs from wave 1' zero weights. This fieldwork practice would have resulted in higher attrition rate initially, so the sample size would have decreased much more rapidly. Instead, households are issued to the field even if they have missed the previous one or two waves, resulting in non-monotone sample attrition patterns but larger responding sample sizes.

Our longitudinal weights are developed for monotone attrition, i.e. requiring a response to a particular instrument in each of a number of consecutive waves. If you use some but not all combinations of waves (e.g. waves 1, 3 and 8) you can increase the number of respondents in your analysis via creating a tailored weight. See Q14 and Q15.

Our cross-sectional weights are derived from the longitudinal weights and will consequently be zero for persons in a household in which no person has a longitudinal weight, i.e. when the whole household missed at least one previous wave in the sequence to which that particular longitudinal weight refers. Again, it is possible to increase the sample size in your analysis by creating a tailored weight (see Q15).

Before you decide to create your own tailored weight please consider the following example. Let's say we are interested in estimating the proportion of mothers (natural/adoptive/step) of children under 16 years old in wave 8. In Table 1 you will find an unweighted estimate that uses all 39,289 people who have provided this information, and a weighted estimate that uses the 33,818 people who have a non-zero cross-sectional weight. The table also shows projections of the change in the estimate if we had smaller sample size. For this we use all people with non-zero weights and randomly select samples of 20, 30, 40, 50 and so on. It can be seen that the estimate stabilizes at around 1000 respondents in our sample, adding another 4000 brings the estimate to 13.3%, but adding a further 25,000 respondents does not change the estimate much. Looking at the standard error of these estimates, at n=1,000 it is 0.013, by 5,000 the standard error is 0.006, at 20,000 it is 0.003, and adding another 10,000, at 30,000 it is 0.002. If we project it further past 33,818 the standard error does not change (to 3 d.p.) with the further 5,471 people that could potentially be included in the analysis.

Table 1

| sample size | estimate | Std Error | CI low | CI high |
|---|---|---|---|---|
| 20 | 0.193 | 0.096 | -0.008 | 0.395 |
| 30 | 0.171 | 0.075 | 0.017 | 0.325 |
| 40 | 0.188 | 0.067 | 0.053 | 0.323 |
| 50 | 0.139 | 0.051 | 0.037 | 0.241 |
| 75 | 0.097 | 0.036 | 0.025 | 0.168 |
| 100 | 0.106 | 0.033 | 0.039 | 0.172 |
| 300 | 0.171 | 0.028 | 0.115 | 0.226 |
| 500 | 0.146 | 0.020 | 0.107 | 0.185 |
| 1000 | 0.138 | 0.013 | 0.113 | 0.164 |
| 5000 | 0.133 | 0.006 | 0.122 | 0.144 |
| 10000 | 0.133 | 0.004 | 0.125 | 0.140 |
| 15000 | 0.132 | 0.003 | 0.126 | 0.138 |
| 20000 | 0.131 | 0.003 | 0.126 | 0.136 |
| 30000 | 0.131 | 0.002 | 0.127 | 0.135 |
| 33818 | 0.132 | 0.002 | 0.128 | 0.136 |
| | | | | |
| unweighted | | | | |
| 39,289 | 0.150 | 0.002 | 0.146 | 0.153 |

There are two important points to add. First, don't use an unweighted estimate just because it has a higher number of respondents: here the unweighted estimate is 15%, but the weighted estimate even at n=1000 respondents (with not much change until n=33,818) is closer to 13.2%. If you create your own tailored weight to gain higher sample size, remember that you will limit the number of predictors used in your nonresponse correction – in other words you will choose quantity over quality omitting

much of the information about people known from in-between waves used in the current weights.

One more point to add: before you create your tailored weights run your model with our weights first and look at the results. If your results are already statistically significant – you can draw conclusion with this sample size already. If your p-value is 0.4, it is unlikely that adding 10-20% in sample size will make it small enough to reach the 0.05 level.

### 14. My analysis sample is a subset of one for which weights are provided. What weight should I use?

It depends whether the subset is defined by personal characteristics (e.g. socio-demographics or geography) or by having participated in a particular combination of survey waves or completed a particular combination of survey instruments.

In the former case, it is appropriate to use the provided weight, even though this was derived for the whole sample. Though not tailored specifically to your analysis sample, the provided weights should not only make the total sample representative of the total population but should also make any subset of the sample representative of the equivalent subset of the population. For example, sample members resident in Wales will represent the population of Wales; female sample member born between 1951 and 2000 will represent all women in the population born between 1951 and 2000, and so on.

In the latter case, you have a choice between three options:

a) Use the weight provided for the (smallest) hierarchically-superior (larger) sample;
b) Use the weight provided for the (largest) hierarchically-inferior (smaller) sample;
c) Derive your own weight, tailored to your analysis sample (see Q15).

The first two options are both sub-optimal, in different ways, but are simple to implement and the sub-optimality may be minimal. With option a), the weights will be correcting for a different nonresponse process to the one relevant to your analysis sample. For example, suppose your analysis sample consists of people who gave a full interview, including the self-completion component, at waves 1, 2 and 5. The smallest hierarchically-superior sample for which weights are provided is the set of people who gave a full interview, including the self-completion component, at waves 1 and 2, `b_indscus_lw`. This set of people consists of 59.1% of the wave 1 OSM individual respondents, whereas your analysis sample consists of 38.2% of the wave 1 OSM individual respondents, so the weight corrects for the correlates of the drop-off to 59.1% of the sample, but not the additional drop-off to 38.2%. If the correlates of nonresponse

differ for these two responding samples (the 59% and the 38%), using the sub-optimal weight may introduce a little bias.

With option b), weights will not be defined for all potential members of your analysis sample, but the weights will correct for the relevant nonresponse process. In the example, the largest hierarchically-inferior sample for which weights are provided is the set of people who gave a full interview, including the self-completion component, at all of waves 1 to 5, `e_indscus_lw`. This weight is only defined for 15,613 out of the 17,977 members of your potential analysis sample, i.e. 86.8%, so using this weight will cause a (very slight) loss of precision (but will not introduce additional bias).

Option c) is described in answer to Q15.

### 15. There isn't a weight for the combination of waves and instruments that defines my analysis sample: How do I derive my own?

If your analysis sample is a nonresponse-defined subset of a sample for which analysis weights have been provided, you can derive your own analysis weight.

***When?***

You may consider doing this if no analysis weight has been provided for the combination of waves and instruments that you wish to include in your analysis and if the solutions suggested in response to Q14 are not satisfactory. For example, suppose you wish to carry out longitudinal analysis of responses to questions that were included at waves 1, 4 and 7. Your analysis base is therefore sample members who completed an individual interview at each of those three waves (let's assume that your survey questions of interest were not all included in the proxy questionnaire and that you therefore cannot include proxy responses in your analysis).

One option would be to use the wave 7 longitudinal weight for the wave 1 sample, i.e. `g_indinus_lw`. However, this weight is only defined for sample members who gave a full personal interview at all seven waves, thus 18,510 persons have this weight, whereas 20,390 responded at waves 1, 4 and 7 (so, 1,880 of those who responded at waves 1, 4 and 7 must have failed to respond at one of waves 2, 3, 5 or 6). Using this weight for your analysis would therefore cause almost 10% of your potential analysis sample to be dropped from the analysis. This reduction in sample size will cause a modest reduction in the precision of your analysis (increase in standard errors). The effect will be rather small, and you may well be willing to accept this slight reduction in sample size, unless you are producing estimates for very small population subgroups. But if you want to be able to include all 20,390 respondents in your analysis, you will need to derive your own weight.

***How?***

First, identify the (smallest) hierarchically-superior sample for which weights have been provided. In this example case, it is the wave 1 responding sample. For this sample, the weight `a_indinus_xw` has been provided. This will serve as your “base weight”, to which you will make an adjustment tailored to your analysis sample.

Next, fit a conditional model (e.g. logit) of response to your wave-combination of interest. In the example case, the base for the model would be all wave 1 responding OSMs (i.e OSMs with a non-zero value of `a_indinus_xw`) and the dependent variable would be a 0/1 indicator of whether they also responded at both wave 4 and wave 7 (and removing from the base any known to have died or emigrated before wave 7). Predictor variables in the model can be anything relevant observed at wave 1. The model will give you a predicted probability for every wave 1 respondent of responding also at waves 4 and 7. Call this $P_i$.

Now, to make the adjustment to your base weight you simply multiply `a_indinus_xw` by by $1/P_i$ for all the cases in your analysis sample.