Understanding Society
Working Paper Series

No. 2024 – 01

January 2024

Weighting and Population Estimation using Understanding Society: Some Frequently Asked Questions

Olena Kaminska and Peter Lynn

Institute for Social and Economic Research, University of Essex

Understanding Society
THE UK HOUSEHOLD LONGITUDINAL STUDY

UKRI Economic and Social Research Council

Non-technical summary

If a user analyses data without taking account of the complex sample design and discusses estimates without their confidence intervals, the underlying assumption is that the data comes from the census with 100% response rate, meaning that we have full information on everyone in the population. And if data is reported with confidence interval, but without an account of its complex sample design, there are two underlying assumptions: 1) the data has been collected through a simple random sample (equivalent to a lottery selection); and 2) everyone invited has responded (or at least respondents are identical to nonrespondents on all values in a model). Neither of the scenarios apply to Understanding Society, the UK Household Longitudinal Study (UKHLS): the dataset has a complex sample design and some nonresponse, all of which need to be taken into account when analysing the data.

In order to use the dataset correctly, and make population inferences from it, users do not need to understand all the complex details of the sample design, or even sample structure. But there are four important points that a user needs to identify prior to analysis: 1) to which population should inferences be made (whether cross-sectional or longitudinal; households, people or events; specific subgroups); 2) clustering variable (usually, W_psu); 3) stratification variable (W_strata); and 4) a weight relevant to the analysis.

The population comes from a research question, but defining it clearly before opening the dataset may help the user to structure the data correctly. The clustering variable reflects the way the data was collected, where postcode sectors (clusters in UKHLS) were selected first and households were selected within these postcode sectors. As households and individuals in the same postcode sector may be similar to each other this needs to be taken into account when analysing the dataset. Stratification improves the statistical efficiency of a sample and provides a perfect sample distribution across the strata. Taking stratification into account reflects the narrower confidence intervals of the design with stratification. Weighting accounts for unequal selection probabilities (in UKHLS related to a boost in Scotland, Wales and larger one in NI, and additional boosts of ethnic minorities and recent immigrants), and corrects for nonresponse and attrition (known to be related to many variables collected in UKHLS). To reflect the complex data structure there are multiple weights provided with the dataset, so selecting the correct weight is another step to complete before finally starting analysis.

For newcomers to UKHLS reading through the first half of the questions may be useful, reading them in order. For the experienced UKHLS users looking for a specific question may be most suitable.

Weighting and Population Estimation using Understanding Society: Some Frequently Asked Questions

Olena Kaminska and Peter Lynn

Institute for Social and Economic Research, University of Essex

Abstract: *Understanding Society* has a complex sample design, and selection due to nonresponse, both of which need to be taken into account when the data are analysed. This paper provides support for users on how to take the sample design into account, why this is important, and how to identify, or derive, a weight suitable for the analysis at hand. We provide answers to many frequently asked questions collected over the years since *Understanding Society* data was first released in 2011. They start with basic concepts, and progress to more complex situations and their solutions. It will be most useful to *Understanding Society* users but will also have a wider service as a successful example of supporting data users after data release.

Keywords: complex sample design, weights, stratification, clustering

Acknowledgments: We would like to thank the many UKHLS users, who over years asked all the provided and many more questions in their search of the correct use of data and inference from it.

Understanding Society is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by the National Centre for Social Research (NatCen) and Verian. The research data are distributed by the UK Data Service.

Data Citation: University of Essex, Institute for Social and Economic Research. (2023). Understanding Society: Waves 1-13, 2009-2022 and Harmonised BHPS: Waves 1-18, 1991-2009. [data collection]. 18th Edition. UK Data Service. SN: 6614, http://doi.org/10.5255/UKDA-SN-6614-19.

Corresponding author: Olena Kaminska, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Essex, CO4 3SQ, olena@essex.ac.uk

Weighting and Population Estimation using Understanding Society: Some 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 2001, UK in 2009-10, in 2014-15 and in 2022-2023:

- 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 between 2009/10 and 2013/2014, data collected between 2015/16 and 2021-2022 exclude households consisting entirely of immigrants between 2014/15 and 2021/2022, and data collected since 2023/2024 exclude households consisting entirely of immigrants since 2022/2023.

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
- 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, stratification and a weight. 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 or you can create a weight adjustment tailored to your analysis (see Q17).

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

- Starting at wave 14 (2022-23) onwards: ug_
- Starting between wave 6 (2014-15) and wave 13 (2021-2022): 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. How to deal with one PSU per stratum?

UKHLS has a very detailed stratification sampling method. This improves precision of the sample and for a user to reflect this gain in precision in their analysis, the detailed stratification should be taken into account. This may however create an issue in analysis if there is one or more stratum in the dataset that contains just one PSU. This situation occurs over time due to attrition, but is also much more likely in analysis of small population subgroups (as the subgroup may not be represented in all PSUs evenly). To help Stata run the syntax while taking into account detailed stratification, we advice to always use 'singleunit(scaled)' option in svyset syntax:

```
svyset psu [pw=weight], strata(strata) singleunit(scaled)
```

12. Analysis of small subpopulations

When working with small subpopulation, e.g. a sample size of 300 or lower, consider using higher p-values, e.g. 0.1, one-sided hypothesis, and follow this advice from Stata:

> svy commands can produce proper estimates for subpopulations by using the subpop() option. Using an if restriction with svy or standard commands can yield incorrect standard-error estimates for subpopulations. Often an if restriction will yield the same standard error as subpop(); most other times, the two standard errors will be slightly different; but sometimes— usually for thinly sampled subpopulations—the standard errors can be appreciably different. Hence, the svy command with the subpop() option should be used to obtain estimates for thinly sampled subpopulations. See [SVY] Subpopulation estimation for more information.

from https://www.stata.com/manuals/svy.pdf page 103.

13. 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 Q17)
- 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 Q16*).*
- 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).

14. 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.

[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.

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
```

15. What is important when pooling data?

Aside from the need to combine a multiple of all 24 months of sample (see Q14 in this document), there are three most important points to keep in mind when you pool data:

1. Always take into account clustering within PSUs with UKHLS data. Taking into account clustering within a person (in case you have multiple entries per person) is optional and could be used in addition to clustering within PSUs. This implies that you don't need to use

multilevel models while pooling – you could use the standard svy command if this suits your purpose.

2. When pooling information from multiple waves, especially BHPS waves and UKHLS waves, you need to apply additional scaling to the weights in order for each wave/year to contribute a similar proportion to the analysis. See Q16 in this document for how to implement it.
3. Define your population carefully. Values need to vary at the lowest level of analysis. For example you can study events across time and pool individual questions from across waves, but you can't pool individuals over time and study year of birth as it is constant over time.

Pooling can be cross-sectional or longitudinal. Theoretically, you will be combining 'separate samples of events / states' each of which will represent a population of these over a particular period of time, and will have the corresponding weight.

If all of your information comes from the concurrent point in time (and the same wave) for each event / state that you study, you are pooling cross-sectional information. For this create a new weight variable new_weight, and give it a value of the cross-sectional weight for the wave from which the observation comes (e.g. new_weight=a_indinus_xw if wave==1; new_weight=b_indinub_xw if wave==2 etc.)

Alternatively, you may be interested in what happens before and / or after a particular event therefore using longitudinal data for each event / state. In this situation you need to choose a longitudinal weight from the last wave in your analysis for each event / state. For example, you may be interested in mother's experience one year before and one year after giving birth. You therefore may use a combination of t-1 wave, t wave (where birth occurs) and t+1 wave for each observed birth. In such situation your weight for each birth and information around it comes from the last wave of analysis, which in this situation is t+1.

16. Scaling weights

In pooled analysis and sometimes in other types of analysis you may need to apply an additional scaling to our weights. Our weights have a mean of 1 in each wave, which means that if combined in a pooled analysis the waves with smaller sample size will have a smaller contribution in your analysis. This includes BHPS waves and later waves (as sample size decreases with attrition). Ideally, when combining events / states over 30 years (for example) you want each year to have the same importance. To ensure this follow this example to calculate an additional scaling for your weights.

For example, you are looking at job quality and therefore are pooling information from wave 2, 4, 6 & 8 as these are the waves when the questions are asked. Here is how to create a scaled weight for this analysis.

```
ge weightscaled=0
replace weightscaled=b_indpxub_xw if wave==2

ge ind=1
sum ind [aw=b_indpxub_xw] if wave==2
gen bwtdtot=r(sum_w)
sum ind [aw=d_indpxub_xw] if wave==4
gen dwtdtot=r(sum_w)
sum ind [aw=f_indpxub_xw] if wave==6
```

```
gen fwtdtot=r(sum_w)
sum ind [aw=h_indpxub_xw] if wave==8
gen hwtdtot=r(sum_w)

replace weightscaled=d_indpxub_xw*(bwtdtot/dwtdtot) if wave==4
replace weightscaled=f_indpxub_xw*(bwtdtot/fwtdtot) if wave==6
replace weightscaled=h_indpxub_xw*(bwtdtot/hwtdtot) if wave==8
```

You can double check by looking at the sum of ind with weightscaled for each wave – it should be the same.

```
sum ind [aw=weightscaled] if wave==2
sum ind [aw=weightscaled] if wave==4
sum ind [aw=weightscaled] if wave==6
sum ind [aw=weightscaled] if wave==8
```

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

If there is no analysis weight provided for your subsample, you can derive your own analysis weight. 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. For example, suppose you wish to carry out analysis of youth and parental income at the time. This would require a response by youth as well as full interview response by their parents. Another example may be a use of biomarkers collected at wave 2/3 together with responses to Covid questionnaire. Analysing 5-year olds pooled from across all the waves in a longitudinal context, with observations in previous years and later years, and maybe connecting this to parental information or sibling information will also require a tailored weight.

We have created an online course to help you create a tailored weight on your own. Access it here:

https://www.understandingsociety.ac.uk/help/training/creating-tailored-weights/

A suboptimal alternative to creating your own tailored weight is to use one of our weights that is not perfectly suitable for your analysis. This weight will correct perfectly for unequal selection probabilities, and for a large part but not all of nonresponse. In such situation you have a choice between two options: a) use the weight provided for the (smallest) hierarchically-superior (larger) sample; or b) use the weight provided for the (largest) hierarchically-inferior (smaller) sample. In the above example where parental information is combined with youth information, you could use enumeration weight (hierarchically-superior), or youth weight or parent's weight (hierarchically-inferior).