Statistical inference using weights and survey design – a practical guide with UKDS surveys

Pierre Walthéry

27 April 2023

Introduction

This note aims at providing users of UKDS social surveys with guidelines for statistical inference using weights and survey design variables. It focuses on a limited number of practical procedures and does not discuss the theoretical underpinnings of survey design, sampling or estimation with weighted survey data. The content is based on technical documents by data producers such as the Office for National Statistics as well as the relevant statistical literature. Examples are currently drawn from the Labour Force and Family Expenditure surveys and will be gradually be expanded. A list of essential references is provided in the bibliography.

Key notions

Statistical inference is the process through which unknown characteristics (sometimes called *parameters*) of 'large' populations are estimated from 'small' samples that are drawn from them, usually at random. In part as a result of design decision, in part – and increasingly so – due to non response , real-life social surveys present some degree of bias, which needs to be compensated for by weighting estimates and accounting for the survey design when carrying out statistical inference. Indeed most UK social surveys rely on sampling techniques such as stratification or multi-stage clustering, sampling proportional to size in order to strike a compromise between non response, unequal probability of selection and/or improve representation of hard to reach groups and cost limitation. Although often referred to in the theoretical literature, Simple Random Sampling (SRS) is the exception rather than the rule.

Weights are a special type of numeric variables included in survey datasets. They have been designed in order to correct for bias in the representativeness of the sample. They are usually made of at least two components: *design weights* that account for issues of unequal probability of selection resulting from survey design constraints and decisions; *non-response weights*, correcting for lower propensity to take part to surveys among certain categories of respondents. Often social survey weight have been rescaled so that they can be used to gross up sample counts to population totals thus becoming *grossing weights*. In that sense the weights attached to observations are an indication of the number of units these observation 'represent' in the population. Computation rely on calibration algorithms that optimise the conditional distribution of the weighting variables given the sample size (for example the conditional distribution of age, by gender and economic status).

Sample design variables typically consist of variables identifying the strata and/or clusters, especially the Primary Sampling Units (PSU) used during the sampling process. Used in conjunction with weights, they enable researchers to produce more accurate estimates. Typically, clustered survey present lower precision than non clustered ones, and conversely statrified surveys are more precise than non stratified ones.

Point estimates (ie mean, median, proportion, counts) are single quantities or parameters that are used to describe the characteristics of a population: for example the mean/median age, the proportion of people in employment, the total number of men and women. They are the most commonly produced estimates in inferential statistics. Their value is usually directly affected by weights.

Uncertainty, precision, error confidence: point estimates are not very informative in that they do not provide a measure of their own uncertainty. Parameters such as standard errors or confidence intervals provides an indication of the degree of precision of point estimates. A large standard error or a wide confidence interval indicate imprecise estimates, whilst the reverse is true if they are small/narrow. Crucially our ability to estimate uncertainty is influenced by knowledge of the survey design, and likely to be biased in its absence.

1. The problem: infering things with survey data

It is usually considered that in order to produce ‘good’ statistical estimates from survey data, including as much of the survey design information as possible alongside non response and sampling weights is required. ‘Good’ estimates need to be at the same time ‘representative’ – reflect the characteristic of interest in the population, but also reasonably precise in order for the inference to be meaningful.

This can prove challenging to some researchers as the information required is not always available and may be costly and/or time-consuming to acquire, the statistical techniques are not always widely documented, and their implementation in statistical software may not always be clear. At the same time, ad hoc, sometimes naive inference from survey data that relies on the sole usage of weights is a common practice, with seemingly little consideration of the methodological implications involved, and sometimes wrong results.

Using weights and accounting for sample design when making inferences about a population raise complex issues for which in most cases no ideal solution exist, and trade-offs have to be made by researchers. Factors such as the sensitivity of the analysis, the size of the group studied and the statistical techniques used, the type of license of the dataset held, the capabilities of the statistical software, as well as the sample design and the kind of weights provided should be taken into consideration.

Let us consider these briefly:

- The variability (and therefore the degree of precision with which they can be estimated) of point estimates is contingent on survey weights and survey design. The extent to which the latter departs from simple random sampling affects both the representativeness and the precision of a characteristics measured in a sample. As a result knowledge of and ability to use survey design variables is important. Not having the information available may lead to under- or over estimation of the precision of estimates, and therefore biased conclusions.

- Reflecting on the **scope** of the intended analysis will help researchers decide how important it is to strive to use the most accurate estimation technique available or instead to settle for one that is 'good enough'. Not all analyses necessarily require the highest degree of precision. This imperative could best be seen as lying on a continuum ranging from 'playing with the data' to producing numbers that will be subject to public scrutiny, or when policy decisions will be made based upon them. The latter usage require such a degree of precision – for example when publishing official population numbers or writing a research article, other less so – for instance when exploring data or preparing examples for teaching. In the latter cases, users may simply need to get a rough idea of a population figure or the interval within which it may lie.

- **Point estimates vs confidence intervals** It can be tempting to consider that the main goal of statistical inference consists in identifying the representative point estimate of a quantity of interest such as the 'mean weight of adult males', the 'median poverty rate', or the value of some regression coefficient in a multivariate study. This is potentially misleading as point estimates can be at the same time representative and very imprecise. It could even be argued that focusing on a single value when conducting estimation of population characteristics wrongly entertains the idea that indeed a unique, true value exists for the population of interest. This can lead to users to disproportionately focus on weighting point estimates and less so on their precision thus also overlooking factors that impact it - chiefly survey design. It is in fact more useful to think of them as a single reality: the range of values we think a parameter of interest can take in the population, with a certain degree of confidence.

- Most UKDS data are available under **End User License**. This presents the advantage of enabling large numbers of users to access data with a minimal level of formalities to go through but comes with the significant trade off that often the survey design information is not included in the data, due to concerns about the risk of personal information disclosure. For a large number of studies including key ones such as the Labour Force Survey, only users granted access to the data via Secure Lab can use these variables. As applying to Secure Lab access is a lengthy process – it can take up to 6 months in total, it not practically feasible for all researchers, in particular those outside academia or large organisations. The consequence is that for some, there will inevitably be a ceiling in the level of precision of the estimates they will be able to compute.

- Common statistical software (ie R, SAS, SPSS, Stata) have functions for both performing analyses with casual weighting as well as specific procedures for survey design informed analysis, such as the R *Survey* package, the SPSS *Complex Survey* add-on and Stata's *svy:* set of commands. This is because weighting can be used in other contexts than statistical inference and can be potentially confusing, especially for non advanced users who might be tempted to simply use casual weighting without giving too much thought to the assumptions about survey design under which such estimates are produced (namely simple random sampling), as they are left implicit by the software. There are also cases where results can be plainly wrong: for instance the standard weighting command of of SPSS (without the Survey add-on) returns wrong values for standard errors and statistical test of casually weighted survey data, as they are calculated with population rather than sample totals.

- The **degree of complexity of the analysis** namely whether estimation involving small numbers of observations is required and/or whether parameters for subgroups of the population (also known as domains) need estimating. Here again this can be seen as lying on a continuum ranging from producing simple univariate descriptive estimates for the population as a whole to more complex estimation of small groups characteristics and/or multivariate analysis. The former is conceptually and practically more straightforward than the latter and in some cases, the estimate of interest may already exist. In others the data producers may have published design factors ie numbers allowing to adjust the precision of estimates produced without survey design variables.

2. What to do? Practical recommendations for statistical inference with survey data

Although the gold standard of parameter estimation with social surveys relies on using both weights and survey design variables, it is not always possible to go down that path for the reasons just highlighted. The section below provide a series of practical recommendations for robust estimation taking into account the research context in which it is taking place, starting with the degree of sensitivity of the analysis. **It is important that it is not recommended to carry out any statistical inference from social surveys without applying weights.**

2.1 Medium to high sensitivity analysis

These are cases when knowing about the uncertainty of our estimates is required, and therefore inference goes beyond simply getting an idea about the value of some point estimates. Most of the time survey researchers or data analysts are interested in or required to provide an indication of the degree of precision of their point estimate. This is usually achieved by computing confidence intervals and/or standard errors whose correct estimation depends on the information we hold about the survey design.

1. Two cases can arise: when the information about sample design is not available, and when it is. Sometimes survey design variables are included in datasets accessible under End User licence, more often than not, these can only be accessed in the Secure Data Lab, a process that can take up to 6 months. Information about how to apply for Secure Lab Access is available here on the UKDS website. When survey design characteristics such as strata, cluster, or primary sampling unit identifiers are not present, computation of estimates (produced with or without SDI commands) will wrongly rely on the assumption that the sample was obtained via random sampling. Depending on the design, this will lead to under- or over- estimate the variability and precision of estimates.

2. As a general rule, we recommended to use survey design informed (SDI) inference, that is to use survey design specific commands available in statistical software, irrespective of whether survey design information is available in the data. Some post processing of the result may be involved in the former case.

3. If the survey design information is present a typical workflow involves (examples are provided in Section 3):

- Finding out about the survey design and identify the relevant weights and survey design variables in the data documentation;
- Declaring the survey design using software-specific commands
- Producing the estimates of interest, using survey design specific estimation commands available
- Documenting the confidence interval for the estimate of interest or alternatively the point estimates *and* its standard error.
- If required, provide a brief discussion of the possible source of bias of the results (specifically under/over estimation of the uncertainty of the estimates)

4. If the survey design information is not available, in case of simple population characteristics, the estimate of interest may already have been published by the data producer, in which case they may be directly quoted instead of computed from data.

5. If not, then an alternative workflow could consist of;

- Finding out about the survey design in the data documentation and identify the weights variable ;

- Declaring the survey design as simple random sampling using software-specific commands

- Producing the estimates of interest, using survey design specific estimation commands available

- Research whether the data producer has published design factor that could be applied to the produced standards errors (for example design factors computed for the same population at another point in time). A design factor is a number by which to multiply standard errors estimated under the assumption of simple random sampling, that will adjust it for survey design characteristics.

- Documenting the resulting confidence interval for the estimate of interest or alternatively the point estimates *and* its standard error.

- If no design factors are available for the estimates of interest, state the likely bias they may suffer from ie over estimation in case of cluster sample, under estimation in case of stratified sample, usually available from the survey documentation. The wider the initial confidence interval (ie computed under SRS assumptions) the larger the likely bias. Or from another perspective, the smaller the (sub)sample, the larger the likely bias.

6. Computing SDI estimates for subpopulations (also known as ‘domains’) rather than for the population as a whole requires extra precautions. This is the case for example when we are interested in the mean age by employment status, or some other categories, or alternatively,

in analyses restricted to a subset of the population (for example those in employment). The key differences is that when computing domain estimates we are in fact producing estimates about an group of the population whose size we also need to estimate. This requires ensuring that the whole distribution of weights in the sample is taken into accoung, not just the weights values for the groups we are interested in. Failure to do so might result in computing incorrect values. SDI commands in statisticalk software are designed to tackle this potential issue.

2.2 Lower sensitivity analysis

We do not recommend using casual weighting for inferential analysis, but there are circumstances where this will be the only option open to users. There are also cases when users are not interested in knowing about the uncertainty of their estimates (ie their confidence interval, standard errors of point estimates, or conduct statistical testing), for example because they are simply learning or teaching basic statistical concepts or how to use software.

In such cases, it can be acceptable to compute point estimates by adding weights to commands that accepts them, without using survey design-informed functions. Most of these will provide a good idea of the values of the point estimate in the population. By default however, some statistical software will also provide an estimate of standard errors or confidence intervals, which is likely to be misleading as they 'silently' assume simple random sampling, and in some cases will carry out computation with population (ie grossed) totals, resulting in the wrong values.

3. Study-specific weighting and sample design information

3.1 British Social Attitudes Survey

The BSA is a three stage stratified random survey, with postcode sectors, addresses and individuals as the units selected at each stage. Primary sampling units were furthermore stratified according to geographies (sub regions), population density, and proportion of owner-occupiers. Sampling rate was proportional to the size of postcode sectors (ie number of addresses). Some issues of the BSA such as the 2017 include survey design information. The 2017 issue included information about Primary Smapling Units (`Spoint`), strata (`StratID`). Weights are called `WtFactor`.

3.2 Labour Force Survey

The LFS is a geographically stratified random survey. For the main part Primary sampling units are addresses within postcode sectors, drawn from the Small Users Postcode Address File (PAF). The small users PAF is limited to addresses which receive, fewer than 50 items of post per day. In a small number of cases a second stage sampling occurs where several households exist at a given address. A clustering effect is also present to the extent that units of observations are individuals withing households, and that some groups are clustered within these, typically ethnicity. LFS weights: - PWTxx – person level sampling weight; enables inferring population counts - IWTxx - Person-level sampling weight for income analysis (ie subsample of people in paid work) - PHHWTxx - Household-level sampling weight (for household-level analysis)

3.3 Family Resources Survey

The FRS is a stratified clustered random survey, with survey design differing slightly between countries of the UK. In great Britain, Primary sampling units are postcode sectors, drawn from the Small Users Postcode Address File (PAF). The small users PAF is limited to addresses which receive, fewer than 50 items of post per day. Before being selected, PSUs are stratified according to geography, proportion of household reference persons from higher social classes in the area, proportion of economically active respondents in the area, and proportion of economically active men who ware unemployed. In Northern Ireland, the sample is a systematic random sample of addresses.

FRS weights: GROSS4: person-level sampling weight; enables inferring population counts

4. Software notes

4.1 R

Being open source, R does not provide a centralised/unified sets of command to compute weighted estimates and accounting for sample design. The algorithms implementations of statistical theory may vary between packages, but are usually provided in the package documentations. In practice, computing weighted point estimates is straightforward. The *Hmisc* package provides a number of useful functions allowing to do so. The *Survey* package provides a comprehensive set of function for computing point estimates and reliability from survey data. It is recommended to use this for other usages than casual weighting.

4.2 SAS

TBC

4.3 SPSS

Standard editions of SPSS do not include support for survey design variables, and only limited use of sampling weights. When using grossing weights – ie weight that have been designed to enable computing population totals from sample data – as is the case for instance with the Labour Force and Family Resources surveys, measures of dispersion and standard errors will not be adequately computer. It is therefore not recommended to attempt using the base version of SPSS with survey data beyond estimating point estimates. Significance test, and standard errors will not reflect the correct values. USers willing to use SPSS with survey data will need to acquire the Premium Edition or the Complex Samples option of the software.

4.4 Stata

Stata provides comprehensive support for computing estimates from survey data. Users may either opt to add sampling weights to the standard estimation commands, or use survey-specific commands. The latter is recommended when knowledge of estimate precision is required. Stata provides a conceptual distinction between four types of weights: Frequency weights, Variance weights, Importance weights and Probability weights. These differences impact on the way standard errors are computed. In most cases, social survey weights from UKDS datasets should be treated as probability weights. A number of of basic estimation commands, such as *summarise* do not allow using probability weights. This is an explicit features of Stata, meant to nudge users of survey data to prioritise the survey commands rather than 'casual' weighting.

Using standalone weight specification (ie not using survey design functions). In Stata it consists in the weighting variable being specified between square brackets. Stata defines four kind of weights: frequency weights (`fweight`), analytical weights (`aweight`), importance weights (`iweight`) and probability weights (`pweight`). Technically speaking, only the latter (abbreviated as `pw` in most Stata commands) should be used with survey data. However, Stata does not allow using probability weights standalone with its main commands, for the reason highlighted above ie in order for users not overlook survey design issues in their data. Therefore, one has to specify instead the wrong frequency weights (`fw`) if one does not wish to use the survey design functions.

5. R and SPSS Examples

5.1 SDI Inference with design information using R

Example 1 Estimating the proportion of people interested in politics using the 2017 British Social Attitudes Survey

```r
rm(list=ls())
library(dplyr) ### Data manipulation functions
library(haven) ### Importing stata/SPSS files
library(Hmisc) ### Extra statistical functions
library(survey) ### Survey design functions

bsa17<-read_spss("Surveyskills/BSA/UKDA-8450-spss/spss/spss25/bsa2017_for_ukda.sav")
dim(bsa17)
```

```
## [1] 3988  580
```

Once this is done we can specify the survey design: using `Spoint` as Primary Sampling Unit, `StratID` as strata, and `WtFactor` as weights. R does this by creating a `svydesign` object, ie a SDI version of the data, which will be used for subsequent estimation.

```r
bsa17.s<-svydesign(ids=~Spoint, strata=~StratID, weights=~WtFactor,data=bsa17)
class(bsa17.s)
```

```
## [1] "survey.design2" "survey.design"
```

Mean age and its 95% confidence interval

We can now produce a first set of estimates using this information and compare them with those we would have got without accounting for the survey design. We will compute the average (ie mean) age of respondents in the sample. We will need to use `svymean()`

```r
svymean(~RAgeE,bsa17.s)
```

```
##         mean     SE
## RAgeE 48.313 0.4236
```

By default `svymean()` computes the standard error of the mean. We need to embed it within `confint()` in order to get a confidence interval.

```r
confint(svymean(~RAgeE,bsa17.s)) ### Just the confidence interval...
```

```
##          2.5 %  97.5 %
## RAgeE 47.48289 49.1433
```

```r
round(
  c(
    svymean(~RAgeE,bsa17.s),
    confint(svymean(~RAgeE,bsa17.s))
    ),
  1)### Estimate and CI, rounded
```

```
## RAgeE
##  48.3  47.5  49.1
```

Computing a proportion and its 95% confidence interval

We can now similarly compute the distribution of a categorical variable in the population by estimating proportions (or percentages), for instance, the proportion of people who declare that they are interested in politics. This is the `Politics` variable in the BSA. It has five categories ranging from 1 'A great deal' to 5- 'Not at all'. We could recode 1 and 2 - `quite a lot` into 'Significantly', but since we are only interested in estimating the confidence intervals, we will select the relevant values 'on the go'.

```r
attr(bsa17$Politics,"label")     ### Phrasing of the question
```

```
## [1] "How much interest do you have in politics?"
```

```r
attr(bsa17$Politics,"labels")     ### Value labels
```

```
## skip, version off route     Item not applicable       ... a great deal,
##                       -2                      -1                       1
##             quite a lot,                   some,          not very much,
##                        2                       3                       4
##         or, none at all?              Don`t know                 Refusal
##                        5                       8                       9
```

```r
table(as_factor(bsa17$Politics)) ### Sample distribution
```

```
##
```

```
## skip, version off route     Item not applicable      ... a great deal,
##                        0                       0                    739
##             quite a lot,                   some,         not very much,
##                      982                    1179                    708
##         or, none at all?              Don`t know                Refusal
##                      379                       1                      0
```

Note: Changes in a data frame are not automatically transferred into `svydesign` objects used for inferences. We therefore need to recreate it each time we create or recode a variable.

```
round(100*prop.table(svytable(~(Politics==1 | Politics==2),bsa17.s)),1)
```

```
## Politics == 1 | Politics == 2
## FALSE  TRUE
##    57    43
```

Let us now compute the confidence intervals for these proportions. Traditional statistical software compute these without giving us an idea of the underlying computations going on. Doing this in R requires more coding, but also a better understanding of what is actually estimated.

Confidence intervals for proportions of categorical variables are usually computed as a sequence of binomial/dichotomic estimations – ie one for each category. In R this needs to be specified explicitly via the `svyciprop()` and `I()` functions. The former actually computes the proportion and its confidence interval (by default 95%), whereas the latter allows us to define the category we are focusing on.

```
svyciprop(~I(Politics==1 | Politics==2),bsa17.s)
```

```
##                                          2.5% 97.5%
## I(Politics == 1 | Politics == 2) 0.430 0.411  0.45
```

```
round(100*
        c(prop.table(svytable(~(Politics==1 | Politics==2),bsa17.s))[2],
attr(svyciprop(~I(Politics==1 | Politics==2),bsa17.s),"ci")),1
)
```

```
##  TRUE  2.5% 97.5%
##  43.0  41.1  44.9
```

Computing domain estimates

Computing domain estimates, that is estimates for subgroups adds a layer of complexity to the above example. They key point is that as weights were designed using the whole of the sample, computing estimates, in particular confidence intervals or standard errors for part of the sample, therefore using a fraction of these weights may affect the estimates. Instad it is recommended to use commands that take into account the entire distribution of the weights.

In R, the command that does this is `svyby()`

For instance, if we would like to compute the mean age of BSA respondents by Government Office Regions, we need to specify:

- The outcome variable whose estimate we want to compute: ie RAgeE
- The grouping variable(s) GOR_ID
- The estimate function we are going to use here: svymean, the same as we used before
- And the type of type of variance estimation we would like to see displayed ie standard errors or confidence interval

```
# bsa17$gor.f<-as_factor(bsa17$GOR_ID)
# bsa17.s<-svydesign(ids=~Spoint, strata=~StratID, weights=~WtFactor,data=bsa17)

round(svyby(~RAgeE,by=~as_factor(GOR_ID),svymean,design=bsa17.s,vartype = "ci")[-1],1)
```

```
##                              RAgeE ci_l ci_u
## A North East                  46.1 43.6 48.6
## B North West                  49.6 47.3 52.0
## D Yorkshire and The Humber    48.0 45.2 50.8
## E East Midlands               48.6 45.9 51.3
## F West Midlands               48.1 45.0 51.2
## G East of England             49.0 46.0 52.0
## H London                      45.0 43.0 46.9
## J South East                  48.0 45.1 50.8
## K South West                  53.4 51.5 55.2
## L Wales                       49.1 45.1 53.1
## M Scotland                    47.3 44.7 50.0
```

Note: we used [-1] from the object created by svyby() in order to remove a column with alphanumeric values (the region names), so that we could round the results without getting an error.

Our inference seem to suggest that the population in London is among the youngest in the country, and that those in the South West are among the oldest – their respective 95% confidence intervals do not overlap. We should not feel so confident about differences between London and the South East for example, as the CIs partially overlap.

We can follow a similar approach with proportions: we just need to specify the category of the variable we are interested in as an outcome, for instance respondents who are significantly interested in politics, and replace svymean by svyciprop.

```
round(
     100*
     svyby(~I(Politics==1 | Politics==2),
           by=~as_factor(GOR_ID),
           svyciprop,
           design=bsa17.s,
           vartype = "ci")[-1],
           1)
```

```
##                              I(Politics == 1 | Politics == 2) ci_l ci_u
## A North East                                             33.4 26.6 40.9
## B North West                                             41.9 36.1 48.0
## D Yorkshire and The Humber                               35.6 29.1 42.6
## E East Midlands                                          36.9 32.9 41.1
## F West Midlands                                          36.3 31.5 41.5
## G East of England                                        47.2 41.4 53.1
## H London                                                 54.2 47.2 61.1
## J South East                                             44.6 38.7 50.8
## K South West                                             46.5 39.4 53.8
## L Wales                                                  38.6 27.7 50.7
## M Scotland                                               42.7 36.0 49.8
```

5.2 SDI Inference with design information using R

5.3 SDI Inference without design information using R

Example: count and proportion of the regional population of the UK using the LFS with End User License (EUL)

The EUL version of the LFS does not include sample design variables, just two weighting variables:

- `pwt22` for estimation with the whole sample
- `piwt22` for estimation using respondents currently in employment (typically used for earnings estimation)

```
svyset [pw=pwt22]
svy:tab uresmc, cell count percent format(%10.1g)
```

Syntax Using R

```
library(survey)
lfs.s←svydesign(ids=~1,weights=~pwt22,data=lfs) ### Assuming the dataset is stored as lf
svytable(lfs.s)
```

5.4 SDI Inference without design information using R

5.5 Point estimates using casual weighting

See here for the Labour for Survey and for the family resource survey - For the FRS: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/972808/Ch1_Methodology_and_Standard_Errors.xlsx - For the LFS: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesno9guidetocalculatingstandarderrorsforonssocialsurveys#annex-a-labour-force-survey-standard-errors-january-to-march-2015-united-kingdom

- use point estimates with their standard errors using the survey design commands available in most statistical software. These will assume that the data was collected under simple random sampling using provided weights then adjust them using the design factors published by the data providers (LFS). In some cases,

Example: count and proportion of the regional population of the UK using the LFS

Syntax Using Stata

Secure access data Produce estimates using the sample design variables (ie clusters and/or strata) and the specialist survey functions of existing statistical software.

#Estimating quantities about subgroups in the data

Using EUL data In the case of the LFS, standard errors would likely to be overestimated, and therefore the estimates would be conservative, whereas in the case of the FRS, these would be underestimated. The seriousness of these would increase with smaller subgroups, therefore users should try and avoid working with groups that are too small. When estimating quantities for domain, it is also recommended to use functions that explicitly take into the grouping rather than only working with the subpopulation of interest. Not doing so could lead to incorrect weighting of estimates. In case of simple domain estimates, it might still be possible to rely on estimates published by the data producer.

In the case of the LFS

In the case of the FRS

Using secure lab data Point and reliability estimates may be computed using the survey-based estimation commands.

Sample design and multivariate analysis ie regression

TBC – same message as previous section+ issue of controlling for vs using weights.

6. References & further information

UKDS (2019) Weights in social surveys: an introduction: https://www.youtube.com/watch?v=Vllr4olp3N4&t=39s

Goldsmiths(2020) W7: Using survey weights in R https://www.youtube.com/watch?v=brxx81U6N1o

Datacamp (2020) R Tutorial: What are survey weights? https://www.youtube.com/watch?v=8iMV7ei61IM Note: basic, partially available, complex survey design in R

DWP (2014) Uncertainty in Family Resources Survey-based analysis. Guidance on estimating uncertainty in Family Resources Survey-based analysis. https://www.gov.uk/government/publications/uncertainty-in-family-resources-survey-based-analysis

UKDS (2018) Data Skills Modules: Applying weights to survey data https://www.youtube.com/watch?v=TIad5__WP8g Note: point and click howto in SPSS

Curran (2016) Complex Survey Designs and Weighting Using Stata: Part 1-3, https://www.youtube.com/watch?v=oOpJdC_oeKY

ONS (2022) Family Resources Survey, 2020/21 Methodology and Standard Error Tables https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1065513/Ch1_Methodology_and_Standard_Errors.ods

https://www.ibm.com/support/pages/inconsistency-output-when-using-weighting-procedure