EXAMPLE 6:
WORKING WITH WEIGHTS AND COMPLEX SURVEY DESIGN

EXAMPLE RESEARCH QUESTION(S):

- How does the average pay vary across different countries, sex and ethnic groups in the UK?
- How does remittance behaviour vary by socio-demographic characteristics?
- What is the 3-year average pay in the UK?

DESCRIPTION: We show how to use weights and sampling design information provided with the data to obtain appropriate population (mean, regression coefficient) estimates and their confidence intervals[1].

FILES: a_indresp, b_indresp, c_indresp

WAVES: 1, 2 and 3

R packages: `'tidyverse','naniar','haven','survey','zoo','summarytools'`

[1] For further information on why weights should be applied, see, e.g., Wooldridge, J., Haider, S., Solon, G. 2013. What are we weighting for? *NBER working paper* No. 18859.

6.1 OVERVIEW: WEIGHTS, PSU, STRATA

R, like most statistical software, assumes that the sample is a simple random sample and that each sample unit is selected with equal probability and independently of each other. However, most surveys, including Understanding Society, do not fall into this category.

- If sample units are not selected with equal probability by design or if not every selected sample unit responds to the survey, and those included in the sample are systematically different from those not in the sample then population estimates based on this sample will be biased. Applying weights will help produce unbiased estimates of population parameters.
- If the sample design is not a simple random sample then the sampling design features (such as whether this is a clustered and/or stratified sample) need to be considered to produce unbiased estimates of standard errors of the population estimates.

The `survey` package in R offers very convenient commands for producing unbiased estimates of population parameters. We will discuss this in detail in the analysis section.

Variables representing weights and the sample design are routinely available in the Understanding Society data (listed in Table 1). For a brief overview of the sample design see **Appendix A**; further detail is available in the study user guide on the study website, see https://www.understandingsociety.ac.uk/documentation/mainstage.

Table 1: Description of key survey design variables in Understanding Society

Variable	Description	Data file available in
*w*_psu*	Primary sampling unit	All files
*w*_strata*	Strata	All files
*w*_hhorig	Sample indicator	All files
*w*_lda	Low ethnic minority concentration area indicator	All files
*w*_sampst	Individual sample membership status to identify original sample members (OSMs), temporary sample members (TSMs) and permanent sample members (PSMs)	All files
*w*_month*	Monthly sample indicator	All files
*w*_ivfio	Individual interview outcome	All individual level files

*Note: The values of the primary sampling unit, strata and monthly sample indicator are set at the time of sample selection (as these are features of the sample design) and so although these variables, **w_psu w_strata**, have a wave prefix these don't change over time. These variables are also available in the cross-wave file, **xwavedat** without the wave prefix: **psu, strata**.

To choose the correct weights for your analysis, the Weights sections in the User Guide provides detailed tables of options when analysing:

- households or enumerated individuals
- adult respondents including proxy
- adult respondents excluding proxy
- the "Extra 5 minutes" sample
- adult respondents who completed the self-completion questionnaire
- youth respondents

There also is information on design weights (for advanced users).

Note that weights in Understanding Society assume the value of zero for anyone who does not qualify for that sample. Specifically, none of the longitudinal weights (except the household

weights) apply to TSMs. Weights designed for analyses of the general population sample (GPS) and of the ethnic and minority boost sample (EMBS) will be zero for the BHPS sample and vice versa. Individual response weights will be zero for all proxy respondents. For example, the self-completion longitudinal weights (**b_indsc*_lw**) will be zero for TSMs, proxy respondents and anyone who did not complete the self-completion questionnaire. Additionally, the self-completion weights designed for the BHPS sample (**b_indscbh_xw)** will be zero for GPS and EMBS members, while the self-completion weights designed for the GPS and EMBS (**b_indscus_xw)** will be zero for the BHPS sample members.

6.2 DATA PREPARATION

In this example, we will conduct a number of different analyses around three themes:

- producing population estimates of average monthly pay for the UK and separately for each of its four countries
- producing population-level estimates of coefficients of a model of remittance behaviour, i.e., to examine what socio-demographic characteristics are associated with whether a person sends money to his/her home/origin country.
- producing the population estimate of the 3-year average pay using longitudinal weights.

The first two types of analyses will focus on data from Wave 1 (**a_indresp)** only. We will draw on the imputed gross monthly pay variable (**a_paygu_dv)** and on whether or not respondents send or receive remittances (**a_remit5)**. Table 2 lists the substantive variables we will use in this example.

Table 2: Variables to be used in the analyses

Variable description	Variable name
Sex	a_sex_dv
Age	a_age_dv
De facto marital status	a_mastat_dv
Ethnic group	a_racel_dv
Region of residence	a_gor_dv
Educational qualification	a_hiqual_dv
Usual gross monthly pay	a_paygu_dv
Reasons for sending or giving money to people in another country (remittance)	
For repayment of loan	a_remit1
To support family members or friends	a_remit2
To support a local community	a_remit3
For personal investments or savings including property	a_remit4
No money sent/given	a_remit5[#]

[#] This is the relevant variable for our analysis as it is an indicator of remittance (i.e., whether any money was sent or given to anyone in another country)

Note that the remittances question is a multi-choice question, i.e., respondents can choose as many options as apply to them. Responses are stored in five variables and assume a value of 1 if the reason is mentioned or zero otherwise. **a_remit5** is exclusive.

> Many people make gifts or send money to people in another country. Did you send or give money to anyone in a country outside the UK in the past 12 months for any of the following reasons?
>
> 1 Repayment of a loan
> 2 Support for family members or friends

3 Support for a local community. Please do not include donations to large charities such as Oxfam or Save the Children
4 Personal investment or savings, including property
5 No money sent/given

We will also consider a number of standard socio-demographic and socio-economic characteristics; please look through the Online Documentation (as you did for Example 1) to identify such variables. Table 2 provides a suggested list of variables useful for this analysis. The other important variables you will need are those representing the weight, primary sampling unit and strata.

Start by setting up your script and loading all packages required for this exercise. Then read in the data file and keep only the variables you need:

```
a_indresp <- read_dta(
  file = paste0(inpath, "ukhls/a_indresp.dta"),
  col_select = c("a_hidp", "pidp", "a_mastat_dv", "a_strata", "a_psu",
"a_hhorig", "a_indinus_xw", "a_indpxus_xw", "a_ind5mus_xw", "a_ivfio",
"a_sampst","a_age_dv", "a_sex_dv", "a_hiqual_dv", "a_paygu_dv", "a_country",
"a_mastat_dv", "a_racel_dv", "a_remit5", "a_xtra5min_dv")
)
```

Next, examine the data frame and variables of interest, including their distribution. We have come across different commands for this purpose (e.g. `summary` and `table`. You can also type `view(dfSummary(a_indresp))` to see descriptive statistics for all variables in the viewer. The `descr()` , too, is useful to produce descriptive statistics quickly:

```
descr(a_indresp, headings = FALSE, stats = "common", transpose = TRUE)
names(a_indresp)
summary(a_indresp)
table(a_indresp$a_remit5)
```

We see that the remittance question **a_remit5** is missing for many respondents. Specifically, most respondents have a value of [-8], which means they were not asked this question. Why? The remittances module is part of the so-called "Extra 5 minutes" set of questions that are exclusively asked of the so-called "Extra 5 minutes" sample. For a more detailed description of who is in this analytical sample, see Appendix A. The flag variable **a_xtra5min_dv** identifies this sample. We can use it to check whether those with **a_xtra5min_dv==0** are the ones for whom **a_remit5** is inapplicable:

```
a_indresp %>% filter(a_xtra5min_dv == 0) %>%
      freq(a_indresp$a_remit5)
```

We can use **a_ivfio** (full respondent or proxy respondent) and **a_sampst** (sample membership status OSM, TSM, or PSM), and **a_hhorig** (sample origin) to check the applicability rules for the different weights. Create a smaller temporary data frame with just the relevant sampling variables and weights:

```
a_indw <- a_indresp %>%
      select(a_indinus_xw, a_indpxus_xw, a_ind5mus_xw, a_ivfio,
             a_hhorig, a_sampst)
```

and summarise the weights by interview outcome to confirm that the full adult response weights are zero for proxy respondents:

```
stby(data = a_indw, INDICES = a_indw$a_ivfio, FUN = descr,
```

```
stats = c("mean", "sd", "min", "max"), transpose = TRUE)
```

Task: Adapt the **`stby()`** *command above to confirm that*
a) the "Extra 5 minutes" weights are zero for the Northern Ireland sample
b) all wave 1 weights are zero for temporary sample members.
Can you explain why this is the case?

Which weights to use for our analyses?

With so many different weights provided it may appear daunting to choose the correct (or most suitable) one. The Weights section of the Understanding Society user guide provides useful guidance. Having looked at the tables provided on how to choose weights, we can conclude that we should use:

(i) **a_indinus_xw** for the analysis of monthly pay as this information is available for all adult respondents but not proxy respondents, and

(ii) **a_ind5mus_xw** for the analysis of remittance behaviour, as **a_remit5** is an 'Extra 5 minutes' question**.**

Next, let us examine some of the sample design features. Remember that the GPS-NI sample is a simple random sample, so there is just one strata! However, standard statistical software cannot estimate standard errors if there are single PSU strata. By design, this is the case for GPS-NI sample. However, you can think of every household in this sample as a separate PSU. Using that logic, each household in the UKHLS GPS-NI sample has been assigned a separate pseudo-PSU number to allow computations using standard statistical software. You can check that by comparing the mean and standard deviation of the **a_psu** and **a_strata** variables for the GPS-GB (**a_hhorig==1**) and GPS-NI (**a_hhorig==2**) samples.

```
tmp_df <- a_indresp %>% select(a_psu, a_strata, a_hhorig)
stby(data = tmp_df, INDICES = tmp_df$a_hhorig, FUN = descr, stats =
      c("mean", "sd", "min", "max"), transpose = TRUE)
```

6.3 NOTES ON USING WEIGHTS AND "SVYDESIGN" IN R

If this is a simple random sample but not all sample units are selected with equal probability, then only weights need to be specified. However, if the sample design is more complex, like that of Understanding Society, the **`svydesign()`** function needs to be used. The syntax is as follows: newdatafilename <-

svydesign(id=~*psu*, strata=~*strata*, weights=~*weight*, data=*datafilename*)

where *weight* is the variable representing probability weight, *psu* is the variable representing the primary sampling unit and *strata* is the variable representing the stratification variable, *datafilename* is the name of the original data file and *newdatafile* the name of the new data file created with these features. See manual for **survey package** for more explanation. In Understanding Society, the survey design variables are **w_psu** and **w_strata** so we have:

```
svy_indresp <- svydesign(id=~a_psu, strata=~a_strata,
               weights=~a_indinus_xw, data=a_indresp)
```

6.4 ANALYSIS: ESTIMATING AVERAGE MONTHLY PAY IN UK

Before computing population estimates, deal with missing values in the UKHLS[2], using the same for loop we have used previously to set these to system missing [NA]:

```
missval <- c(-9, -8, -7, -2, -1)
for (i in 1:5) {
  a_indresp <- a_indresp %>% mutate_all(., list(~na_if(.,
missval[i])))
}
```

To estimate unweighted mean of gross monthly pay and its standard error without correcting for complex survey design:

```
a_indresp %>% summarise(mean(a_paygu_dv, na.rm=T),
      se=sd(a_paygu_dv, na.rm=T)/sqrt(sum(!is.na(a_paygu_dv))))
```

To estimate weighted mean of gross monthly pay and its standard errors, after correcting for the complex survey design, first specify the survey design, then use the **`svymean()`** command:

```
svy_indresp <- svydesign(id=~a_psu, strata=~a_strata,
               weights=~a_indinus_xw, data=a_indresp)

svymean(~a_paygu_dv,svy_indresp,na.rm=TRUE)
```

Note that if those who are over or under-represented in the sample or those selected with higher or lower selection probabilities are different in terms of gross monthly pay then the weighted estimates will be different from un-weighted estimates.

6.5 SOLVING A POSSIBLE ISSUE: STRATA WITH A SINGLE PSU

When this happens, R cannot compute the standard errors. The programme offers different options to address this issue; specify them after loading the survey package：

```
options(survey.lonely.psu="adjust")
```

The default option is “fail”. The other options are “remove” and “certainty” which ignores the single-PSU strata, “adjust” where the single-PSU strata are centred at the sample grand mean.

6.6 ANALYSIS: ESTIMATING AVERAGE MONTHLY PAY ACROSS THE FOUR COUNTRIES OF THE UK

To examine the differences in mean pay for each country of the UK and to see what difference it makes to apply weights and to consider the survey design features we first compute the average pay in a country without considering the survey design:

```
a_indresp %>% filter(a_country == 1) %>%
      summarise(mean(a_paygu_dv, na.rm=T),
        se=sd(a_paygu_dv, na.rm=T)/sqrt(sum(!is.na(a_paygu_dv))))
```

We then subset the survey data frame svy_indresp for the population of interest by using **`subset()`**. Here, we subset the survey data for England (**a_country==1**), saving a new data frame (**df_e**) and then use this data frame in the **`svymean()`** command to compute the population estimate for England with the correct mean and standard error:

```
df_e <- subset(svy_indresp, a_country == 1)
```

[2] The most common values are [-9] missing, [-8] inapplicable, [-7] proxy respondents, [-2] refused, [-1] don’t know. There are additional missing values in Wave 6 to mark out questions that were only asked of the then new IEMB sample [-11] or not asked of that sample [-10], and a few more in the harmonised BHPS files.

```
svymean(~a_paygu_dv, design = df_e, na.rm=TRUE)
```

Task: Repeat this for the other three countries of the UK.

You may also want to test whether the observed differences in pay across the countries are statistically significant. For this, you can run a t-test. But first create a variable to indicate the two countries you want to compare:

```
a_indresp$engNI <- factor(
  ifelse(a_indresp$a_country %in% c(1),1,
         ifelse(a_indresp$a_country %in% c(4),2,NA)),
level=c(1,2),
label=c("England","Northern Ireland"))
```

Again, specify the survey design dataset,

```
svy_indresp <- svydesign(id=~a_psu, strata=~a_strata,
              weights=~a_indinus_xw, data=a_indresp)
```

Then run the t-test,

```
svyttest(a_paygu_dv~a_country, svy_indresp)
```

The result shows that these differences are statistically significant.

6.7 ANALYSIS: ESTIMATING DESIGN EFFECTS

A clustered sample generally leads to higher standard errors (of some estimated value) compared to a simple random sample of equal size. The opposite is generally the case for a stratified sample. As standard error is a measure of the precision of an estimate, it is good to know how much precision you gain or lose by using a particular sample design.

Two commonly used measures to assess are the design effects DEFF and DEFT. They compare the sample-to-sample variability from a given survey dataset with a hypothetical simple random sample (SRS) design of the same size. DEFF is the ratio of two variance estimates. The design-based variance is in the numerator; the hypothetical SRS variance is in the denominator. DEFT is the ratio of two standard-error estimates. The design-based standard error is in the numerator; the hypothetical SRS with-replacement standard error is in the denominator. If the given survey design is sampled with replacement, DEFT is the square root of DEFF. In R you can compute the DEFF by specifying the option **`deff=TRUE`** in the **`svymean()`** command:

```
svymean(~a_paygu_dv, indresp3, na.rm=TRUE, deff=TRUE)
```

To compute DEFT, type the reported value of DEFF in the **`sqrt()`** function (interactively).

6.8 ANALYSIS: HOW DOES REMITTANCE BEHAVIOUR VARY BY SOCIO-DEMOGRAPHIC CHARACTERISTICS?

Let's practice how to use weights in a multivariate analysis of how remittance behaviour varies by socio-demographic characteristics. Our outcomes variable is whether someone sent money or not, so create a 0-1 indicator variable (or dummy variable) that takes on the value 1 if a person sends money and 0 otherwise. We treat cases with missing values on **a_remit5** as NA:

```
remit <- subset(a_indresp, a_remit5>=0,)
table(remit$a_remit5)
remit$remit <-
      ifelse(remit$a_remit5==1, 0, ifelse(remit$a_remit5==0, 1, NA))
table(remit$remit, remit$a_remit5)
```

For illustration, we include the following socio-demographic characteristics in the model: age **(a_age_dv**), sex (**a_sex_dv**), education **(a_hiqual_dv)**, marital status (**a_mastat_dv)**, ethnic group (**a_racel_dv**), and UK country of residence **(a_country)**. For the categorical variables to be treated as such in the regression model, we define them as factor variables:

```
remit$a_sex_dv    <- factor(remit$a_sex_dv)
remit$a_hiqual_dv <- factor(remit$a_hiqual_dv)
remit$a_mastat_dv <- factor(remit$a_mastat_dv)
remit$a_racel_dv  <- factor(remit$a_racel_dv)
```

To estimate a logit model ignoring the complex survey design and differential response:

```
mylogit <-
      glm(remit ~ a_age_dv + a_sex_dv + a_hiqual_dv + a_country +
         a_mastat_dv + a_racel_dv, data = remit, family = "binomial")
summary(mylogit)
```

To account for complex survey design and differential response:

```
svy_remit <-
      svydesign(id = ~a_psu, strata = ~a_strata, weights =
      ~a_indinus_xw, data = remit)
weighted_logit <-
      svyglm(remit ~ a_age_dv + a_sex_dv + a_hiqual_dv +
                     a_country + a_mastat_dv + a_racel_dv,
            design= svy_remit, family= "binomial")
summary(weighted_logit)
```

You will get an odd warning message that you can ignore. For further details see: https://github.com/alan-turing-institute/PosteriorBootstrap/issues/16.

6.9 USING LONGITUDINAL WEIGHTS

In this section we will show you how to use longitudinal weights for producing estimates in longitudinal analysis. We will estimate the average gross monthly wage of employed individuals across 3 years, i.e., across Waves 1, 2 and 3. As this information is asked of adult respondents we will be using the files **a_indresp b_indresp** and **c_indresp.**

Which weights should we use? We will use information provided by adult respondents but excluding proxies. We will not be using any of their responses to the self-completion questionnaire. The purpose of longitudinal weights is to make the estimates representative of the core sample selected in the first wave and so the longitudinal weights for each wave is a product of the initial weight and all wave-on-wave non-response adjustment weights (for further details see the User Manual). So, the longitudinal weights are non-zero for respondents who have continually responded from the first wave. If you are conducting longitudinal

analyses using N waves of data then use the longitudinal weight from the N^{th} wave. Hence, the appropriate longitudinal weight for this exercise is **c_indinus_lw**.

Below is an example to illustrate longitudinal weights, psu and strata across three waves for individuals with different response patterns. We will organise the data in a long format as it is preferred for panel data.[3] Aspects of this data to note:

- psu and strata values remain the same across all three waves for everyone
- Person 1 responded in all 3 waves and so has a positive longitudinal weight in both Waves 2 and 3.
- Person 2 has missed the second wave and so the value of this weight for Wave 3 is zero
- Person 3 joined the household in the second wave and so has zero longitudinal weights
- Person 4 responded in the first two waves but not in the third wave and so has a longitudinal weight for Wave 2

pidp	wave	paygu_dv	indinus_lw	psu	strata
1	1	1000	.	67	14
1	2	1200	0.9934	67	14
1	3	1200	0.9092	67	14
2	1	500	.	2002	2555
2	3	600	0	2002	2555
3	2	1650	0	1555	1445
3	3	1700	0	1555	1445
4	1	2000	.	560	3003
4	2	2500	0.8745	560	3003

We would like to have the following data structure – i.e. we want a new longitudinal weight variable that assumes the value of **c_indinus_lw** for each person.

pidp	wave	paygu_dv	indinus_lw	psu	strata	c_indinus_lw
1	1	1000	.	67	14	0.9092
1	2	1200	0.9934	67	14	0.9092
1	3	1200	0.9092	67	14	0.9092
2	1	500	.	2002	2555	0
2	3	600	0	2002	2555	0
3	2	1650	0	1555	1445	0
3	3	1700	0	1555	1445	0
4	1	2000	.	560	3003	.
4	2	2500	0.8745	560	3003	.

Let's see how to go about it. First, create a long format file using **a_indresp b_indresp** and **c_indresp** (as shown in Example 3), keeping the variables: **pidp**, **w_paygu_dv**, **w_strata**, **w_psu, w_indinus_lw**. Even though the strata and psu variables are provided with each wave specific file with a wave prefix, as these are sample design variables, their values do not change

[3] For those of you who are curious, the data management tasks and analysis can also be undertaken in the wide format. We have included the steps in the R script accompanying this worksheet.

across the waves for any person. We use the method we used in exercise 3, with different variables.

```
var_list <- paste0("|._",
                   c("paygu_dv", "strata", "psu"),
                   collapse= "")
var_list <- paste0("pidp", var_list)

waves    <- 1:3
prefixes <- paste0(letters[waves], "_")

read_wave <- function(prefix, wave_no) {
  fn <- file.path(inpath, "ukhls", paste0(prefix, "indresp.dta"))

  read_dta(
    file       = fn,
    col_select = matches(var_list)
  ) %>%
    rename_with(~ str_remove(., prefix), starts_with(prefix)) %>%
    mutate(wave = wave_no, .before = 1)
}

long <- map2_dfr(prefixes, waves, read_wave)
```

As before, remove the temporary data frames for good housekeeping and recode the UKHLS missing values to system missing:

```
rm(prefixes, waves, var_list)

missval <- c(-9, -8, -7, -2, -1)
  long <- long %>%
    mutate(across(everything(), ~ replace(., . %in% missval, NA)))
```

The most efficient method to attach the longitudinal weight variable from Wave 3 (**c_indinus_lw**), to each person's previous wave observations, is to merge it by pidp into the long format file that you have just created:

```
    long <- read_dta(file = paste0(inpath, "ukhls/c_indresp.dta"),
col_select = c("pidp", "c_indinus_lw")) %>%
full_join(long, by = "pidp")
```

As **c_indinus_lw** is missing for some person's in some waves, you need to set the missing values to system missing:

```
long$c_indinus_lw[long$c_indinus_lw<0]<-NA
```

Done! The newly attached weight variable (**c_indinus_lw**) can now be used just like any weight variable. Next, create the 3-year rolling average wage (**avg_wage)**. The most elegant function to use here is `rollmean()`; using the align=right option makes sure only cases with valid observations on paygu_dv in the last 3 rows (specified by k=3) are included in the calculation:

```
long <- long %>% group_by(pidp) %>% mutate(avg_wage =
         rollmean(paygu_dv, k = 3, fill = NA, align = "right")) %>%
           ungroup()
```

You can check whether this has worked by browsing the data frame using `view()`. Next, filter on wave 3 cases, specify the survey design, and compute the unweighted and the population weighted mean:

```
long <- long %>% filter(wave ==3)
svy_long <- svydesign(id=~psu, strata=~strata,
```

```
                        weights=~c_indinus_lw, data=long)
long %>% summarise(mean(avg_wage, na.rm=T),
       se=sd(avg_wage, na.rm=T)/sqrt(sum(!is.na(avg_wage))))
svymean(~avg_wage, svy_long, na.rm=TRUE)
```

This completes this worked example. Don't forget to save your script and to produce the log file for future reference!

Appendix A

Understanding Society sample design

- General Population Sample (GPS) has two components: GPS-GB and GPS-NI
  - GPS-GB: A clustered and stratified sample drawn from Great Britain where each unit had an equal selection probability.
  - GPS-NI: A simple random sample from Northern Ireland where sampling units had approximately twice the selection probability as the units in GPS-GB.
- The Ethnic Minority Boost Sample (EMBS): A clustered, stratified sample drawn from high ethnic minority concentration areas in Great Britain. Households at selected addresses were screened in to include households where at least one person was from an ethnic minority group, or their parents or grandparents were.
- The British Household Panel Survey (BHPS) sample became part of the Understanding Society sample from the second wave of the Study.

"Extra 5 minutes" questions

Part of the sample, often referred to simply as the "Extra 5 minutes" sample, are asked some extra questions (approximately 5 minutes' worth) in addition to all the questions the rest of the sample are asked. These questions are generally those of particular relevance to ethnicity related research. For example, in Wave 1 this included questions on remittances, harassment, discrimination, and a detailed migration history.

The "Extra 5 minutes" sample comprises of

– OSMs in the EMBS
– OSMs selected to be part of the General Population Comparison Sample (GPCS). The GPCS consists of approximately 1000 households randomly selected from the GPS-GB (one of every 18 selected addresses in 40% of the selected PSUs). The achieved sample size was approximately 500 households in Wave 1.
– Ethnic minority OSMs in the GPS-GB living in low ethnic minority concentration areas. This status was frozen in Wave 1 and from Wave 2 onwards, all household members of these individuals were included in the "Extra 5 minutes" sample.

Note all TSMs co-resident with the "Extra 5 minutes" sample members are also asked the "Extra 5 minutes" questions.