# 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?

**DESCRIPTION:** In this example 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.

**FILES:** In this example we will use `a_indresp` data file

**WAVES:** In this example we will use information from the **first wave** only

## OVERVIEW

[*All Tables and Appendices referred to in text are at the end of this worksheet*]

SAS, 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 selected with unequal 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. In such cases weights are to be used to produce unbiased estimates of population parameters.
- If the sample design is not a simple random sample as it is in the case of Understanding Society, then these 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.

Variables representing weights and sample design are available in UKHLS data. SAS's `proc survey` suite of commands is very convenient for producing unbiased estimates of population parameters. We will discuss this in detail in the analysis section.

During the lecture we have discussed the sample design of Understanding Society and the weights provided with the data. For a quick recap of the sample design and the key sample design variables including primary sampling unit (PSU) and strata see Table 1 and Appendix A. For further details see the Understanding Society User Manual (https://www.understandingsociety.ac.uk/documentation/mainstage). To choose the correct analytical weights for your analysis see the following tables in the User Manual,

- Table 24 if analysing households or enumerated individuals
- Table 25 if analysing adult respondents including proxy
- Table 26 if analysing adult respondents excluding proxy
- Table 27 if analysing the extra 5 minute sample
- Table 28 if analysing adult respondents who completed the self-completion questionnaire
- Table 29 if analysing youth respondents

For information on design weights see Table 31 (for advanced users)

## DATA PREPARATION

As always, clear any data in memory which has been stored from a previous session.

```
proc datasets lib=work kill nolist memtype=data;
quit;
```

Set a working directory where an output, log and macro file can be stored.

```
/* writing all ouput to a log file */
filename myoutput "SPECIFY WHERE YOU WOULD LIKE THIS BE.txt";
proc printto print=myoutput new;
run;
/* writing to a log file */
filename mylog "SPECIFY WHERE YOU WOULD LIKE THIS TO BE.log";
proc printto log=mylog new;
run;
/* defining a global macro where files can be permanently
stored */
libname ukhls "SPECIFY DIRECTORY WHERE YOU WOULD LIKE THIS TO
BE";
options nofmterr;
```

Finally import the data:

```
proc import datafile= "SPECIFY DIRECTORY WHERE DATA IS HELD"
out=indrespa dbms=sav replace;
run;
```

*Examining the data*

The first step of this exercise is to examine the data by looking at the data, variables of interest and their distribution.

This is a good opportunity to see that you understand why some variables have missing values. As you are aware UKHLS data is provided without any system missing (.) but all missing values are replaced by negative values representing the reason for missing information.

As we outlined at the beginning we will be discussing two analyses: one about monthly pay and the other about remittance behaviour. See Table 2 for a complete list of variables useful for this analysis. The imputed gross monthly pay variable is `a_paygu_dv`:

```
proc freq data=indrespa;
where a_paygu_dv<0;
tables a_paygu_dv ;
run;
```

Whether a person sends remittance or not is a binary variable and so we will need to use a logit or probit model to estimate how remittance behaviour varies by socio-demographic characteristics. In Understanding Society, the question that asks about remittances is,

> 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

The responses were recorded in **a_remit1, a_remit2, a_remit3, a_remit4, a_remit5.** So, the variable that records whether any remittance was made is **a_remit5:**

```
* Check the remittances receipt variable;
proc freq data=indrespa order=internal;
table a_remit5;
run;
```

We also see that the remittance question was not asked (inapplicable) of a lot of respondents. Why? As this question is part of the extra 5 minute questions, it should only have been asked of the "Extra 5 minute" sample (see Appendix A).

Let us create a variable to identify this sample, and name it, a_xtra5min_dv.

```
data indrespa;
set work.indrespa;
a_xtra5min_dv=0;
if a_emboost=1 then a_xtra5min_dv=1;
if a_gpcomp=1 then a_xtra5min_dv=1;
if a_lda=1 and a_racel>4 and a_racl<=97 then a_xtra5min_dv=1;
if a_ivfio=2 then a_xtra5min_dv=-7;
     run;


proc freq data=indrespa;
tables a_xtra5min_dv;
     run;


proc freq data = indrespa;
    table a_xtra5min_dv*a_remit5/missing nocol norow
nopercent;
     run;
```

As we see from the above table this question *was* only asked of those who received the extra 5 minutes questions.

We can also check whether the distribution of weights is as we would have expected.

```
proc means data=indrespa;
class a_ivfio;
var a_indinus_xw a_indpxus_xw a_ind5mus_xw a_indscus_xw;
run;
```

*Which weights to use for our analyses?*

By looking at tables 24-39 in the Understanding Society User Manual we can conclude that (i) for the analysis of monthly pay, which is available for all adult respondent but not proxy respondents, we should use `a_indinus_xw` and (ii) for the analysis of remittance behaviour, which is an extra five minutes question, we should use `a_ind5mus_xw`.

Next, let us examine some of the sample design features. Remember that the GPS-NI sample is a simple random sample. But SAS cannot estimate standard errors if there are single PSU strata. By design this is the case for GPS-NI sample. You can think of every household in this sample as a PSU. Using that logic, each household in the GPS-NI sample has been assigned a separate pseudo-PSU number to allow computations using SAS. This is not the case for the other countries. You can check that by looking at the mean and standard deviation of the psu and strata variables for each UK country.

*Create a variable that identifies the four countries of the UK, name it `a_country`. Remember to attach a value label.*

```
data indrespa;
set work.indrespa;
a_country=1;
if a_gor_dv=10 then a_country=2;
if a_gor_dv=11 then a_country=3;
if a_gor_dv=12 then a_country=4;
     run;

proc format;
value a_country 1="England" 2="Wales" 3="Scotland" 4="NI";
     run;

data indrespa;
set work.indrespa;
format a_country a_country.;
run;
```

## ANALYSIS

Before we start with analysis we should recode the wage variable for all those respondents who provided a response which is not going to be informative for our particular piece of analysis.

```
data indrespa;
set work.indrespa;
if a_paygu_dv>=-9 & a_paygu_dv<=-1 then a_paygu_dv=.;
run;
```

*Estimating average gross monthly pay in the UK*

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

```
proc means data=indrespa;
var a_paygu_dv;
run;
```

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

```
proc means data=indrespa;
var a_paygu_dv;
weight a_indinus_xw;
run;
```

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.

To estimate weighted mean of gross monthly pay and its standard errors, after correcting for the complex survey design:

```
proc surveymeans data=indrespa; /*sas automatically drops
strata with single psu's*/
strata a_strata;
cluster a_psu;
var a_paygu_dv;
weight a_indinus_xw;
run;
```

Standard errors are estimated because SAS drops strata with single PSU's. This is not a problem of the sample design but could happen with any data based on a clustered and stratified design. In this case this happens because the analysis uses non-missing values of

pay, which results in a sample such that some of the observations belong to strata with a single PSU. If we were analysing a different variable this problem may not arise.

*Estimating average gross monthly pay across different regions of UK*

In this sub-section we will estimate mean pay in the four countries of UK and check if these are different.

```
proc sort data=indrespa;
by a_country;
run;
```

```
proc surveymeans data=indrespa;
strata a_strata;
cluster a_psu;
domain a_country;
var a_paygu_dv;
weight a_indinus_xw;
run;
```

We will next test differences in pay across the different countries.

```
data indrespa;
set work.indrespa;
england=0;  if a_country=1 then england=1;
wales=0;    if a_country=2 then wales=1;
scotland=0; if a_country=3 then scotland=1;
ni=0;       if a_country=4 then ni=1;
run;

proc surveyreg data=indrespa;
strata a_strata;
cluster a_psu;
model a_paygu_dv=wales scotland ni;
weight a_indinus_xw;
run;
```

The result shows that these differences are statistically significant.

*Estimating design and misspecification 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. One way to measure this is by using the **design effect (deff).** It is the ratio of the variance of a statistic based on the actual sample design to the variance of this statistic had the sample design been

a SRS (simple random sample) of the same size. In other words, it indicates by how much the variance is inflated or deflated due to the sampling design. **deft** is the square root of **deff**, i.e., it is the ratio of the two standard errors.

The following SAS code (written by Trent D. Buskirk, Ph.D) allows us to estimate the design effect:

```
* Compute SE under complex survey design (SE1);
proc surveymeans data=indrespa mean nomcar;
strata a_strata;
cluster a_psu;
weight a_indinus_xw;
var a_paygu_dv;
ods output Statistics=temp1 (rename=Stderr=SE1
rename=mean=MEAN keep=Stderr keep=mean);
run;
* Compute SE under SRS (SE2);
proc univariate data=indrespa vardef=WGT;
var a_paygu_dv;
weight a_indinus_xw;
ods output moments=temp2;
run;
* Keep SE under SRS and sample size (n);
proc transpose data=temp2 (keep=nvalue1)
prefix=Stat
Out=temp3 (rename=stat1=n rename=stat3=SE2 keep=stat1
keep=stat3);
run;
* Compute DEFF;
data temp4;
merge temp1 temp3;
DEFF=(SE1/SE2)**2*(n-1);
run;
proc print data=temp4;
var MEAN SE1 SE2 n DEFF;
run;
```

*How does remittance behaviour vary by socio-demographic characteristics?*

In this section we will use weights in multivariate analysis. To illustrate this we will use a specific research question: **How does remittance behaviour vary by socio-demographic characteristics?** The different socio-demographic characteristics that we want to control for in this model are: **age, gender, education, marital status, ethnic group, and UK country of residence.** You may want to add other variables such as number of own children, household income, years since arrival to the UK as these could also influence remittance behaviour.

To analyse whether someone sent money or not we will need to create a variable that takes on the value 1 if a person sends remittance and 0 otherwise,

```
data indrespa;
set work.indrespa;
remit=.;
if a_remit5=0 then remit=1;
if a_remit5=1 then remit=0;
run;
```

Check whether the variable is coded correctly:

```
proc freq data=indrespa;
tables a_remit5*remit/missing norow nocol nopercent;
run;
```

Clean the explanatory variables you wish to use in your model.

```
data indrespa;
set work.indrespa;
* recode missings;
if a_hiqual_dv>=-9 & a_hiqual_dv<=-1 then a_hiqual_dv=.;
if a_mastat_dv <0 then a_mastat_dv=.;
* dummy for white majority group;
whitemajority=.;
if a_racel=1 then whitemajority=1;
else whitemajority=0;
* dummy variables for married or cohabiting;
mar_coh=0;
if a_mastat_dv=2 then mar_coh=1;
if a_mastat_dv=3 then mar_coh=1;
if a_mastat_dv=10 then mar_coh=1;
if a_mastat_dv>=-9 & a_mastat_dv<=-1 then mar_coh=.;
run;
```

*How does remittance behaviour vary by socio-demographic characteristics? (Continued)*

Using factor variables:

*Unweighted estimates, without accounting for complex survey design*

```
proc logistic data=indrespa descending;
```

```
class a_hiqual_dv (param=ref ref="Degree") a_sex (param=ref
ref="female") a_country (param=ref ref="England")
      whitemajority (param=ref ref=first) mar_coh (param=ref
ref=first);
model remit=  a_dvage a_hiqual_dv a_sex a_country mar_coh
whitemajority;
run;
```

*Weighted estimates, without accounting for complex survey design*

```
proc logistic data=indrespa descending;
class a_hiqual_dv (param=ref ref="Degree") a_sex (param=ref
ref="female") a_country (param=ref ref="England")
      whitemajority (param=ref ref=first) mar_coh (param=ref
ref=first);
model remit=  a_dvage a_hiqual_dv a_sex a_country mar_coh
whitemajority;
weight a_ind5mus_xw;
run;
```

*Weighted estimates, accounting for complex survey design*

```
proc surveylogistic data=indrespa;
strata a_strata;
cluster a_psu;
class a_hiqual_dv (param=ref ref="Degree") a_sex (param=ref
ref="female") a_country (param=ref ref="England")
      whitemajority (param=ref ref=first) mar_coh (param=ref
ref=first);
model remit(event='1')= a_dvage a_hiqual_dv a_sex a_country
mar_coh whitemajority;
weight a_ind5mus_xw;
run;
```

Note while using factor variables directly on the existing categorical variables is quite convenient you may find it more useful to convert these variables into fewer categories that are more sensible. For example, some categories may just have a few cases. As in the above analysis, we found that some categories did not include any cases and were dropped from the analysis. Also, we may only be interested in knowing whether a person is living with a spouse or partner, so we should convert the marital status variable into a 0-1 indicator variable which takes on a value of one if the person is married, in a civil partnership or in living with someone as a cohabiting couple.

Finally, clean your SAS working directory:

```
/* Clean SAS work directory */
proc datasets lib=work nolist kill memtype=all;
run;

/* stop writing to external log and output files and simply
write to the SAS windows */
```

```
proc printto;
run;
```

## References

Wooldridge, J., Haider, S., Solon, G. 2013. What are we weighting for? *NBER working paper* No. 18859

**Table 1: Description of key survey design variables in UKHLS**

| 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 |
| a_month | Monthly sample indicator | All files |
| a_ivfio | Individual interview outcome | All individual level files |

**Table 2: Variables to be used in the analyses**

| Variable description | Variable name |
|---|---|
| Sex | `a_sex` |
| Age | `a_dvage` |
| De facto marital status | `a_mastat_dv` |
| Ethnic group | `a_racel` |
| 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)

## 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 Five Minute questions

Part of the sample, often referred to simply as the Extra Five Minute Sample, are asked some extra questions (approximately five minutes worth) in addition to all the questions that 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, detailed migration history.

The Extra Five Minute Sample

- Ethnic Minority Boost sample OSMs
- General Population Comparison (GPC) sample OSMs. The GPC consists of approximately 1000 households randomly selected from the General Population Sample (one of every 18 selected addresses in 40% of the selected PSUs). The achieved sample size was approximately 500 households.
- Ethnic minority OSMs in the GP sample 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 five minute sample.

Note all TSMs co-resident with the Extra Five Minute sample members are also asked the Extra Five Minute questions