Analysing Complex Surveys

June 2024

Dr Roxanne Connelly

Welcome

Workshop Aims

By the end of this workshop you will:

- Understand the nature of complex surveys;
- Understand the limitations of analysing complex surveys as if they were simple random samples;
- Understand how to adjust analyses to take into account the design of complex surveys.

Timetable

1400 - 1445	Talk 1
1445 - 1515	Lab Activities
1515 - 1600	Talk 2
1600 - 1630	Lab Activities
1630 - 1700	Final Advice and Questions

Analysing Complex Surveys

June 2024

Dr Roxanne Connelly

Talk 1

A Census

A study of an entire or complete population

n = all

In the UK the decennial census began in 1801 and has been repeated every 10 years*

*Except 1941 due to WWII. The 2021 Scottish Census was also delayed by one year due to the COVID-19 Pandemic.

Non-Probability Samples

- Non-probability sampling methods do not use a formal method of selecting cases.
- With non-probability samples you do not know the probability of an individual in the population being selected into the sample.
- Non-probability samples are probably not representative of the population of interest (but might be informative).
- We can not make inference from the sample to the wider population using non-probability samples.

Probability Samples

- Probability samples are designed to be representative of the population of interest.
- Each individual in the population has a known probability of being included in the sample.
- Cases may have unequal probabilities of being included in the sample (but that is ok, as long as we know what that probability is!).

What are Complex Samples?

Probability Samples

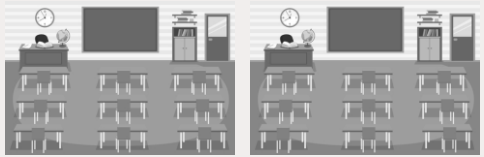
- In order to generalise from a sample to a population you need a probability sample.
- In a probability sample, each individual unit in the population has a non-zero and known probability of selection.
- This is not the case with non-probability samples.

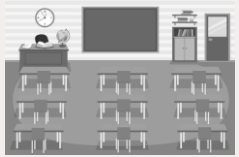
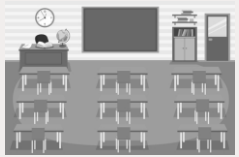
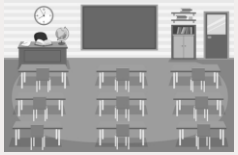
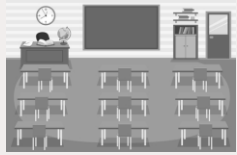
Treiman, D. J. (2014). Quantitative data analysis: Doing social research to test ideas. John Wiley & Sons. Chapter 9: Sample Design and Survey Estimation.

Sturgis, P. (2004). Analysing complex survey data: Clustering, stratification and weights. Social research UPDATE, 43. https://eprints.soton.ac.uk/80188/1/fulltext.pdf

Simple Random Samples

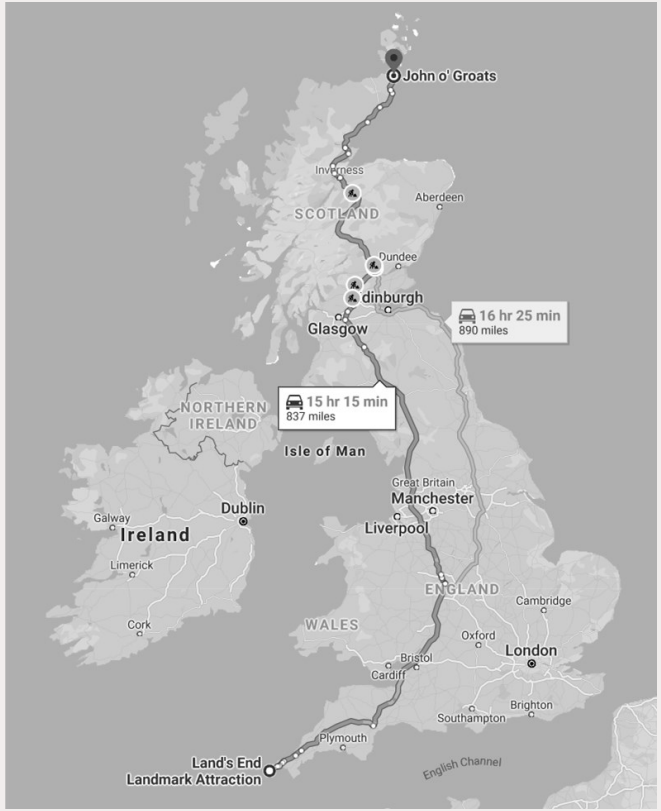
Multistage Probability Samples



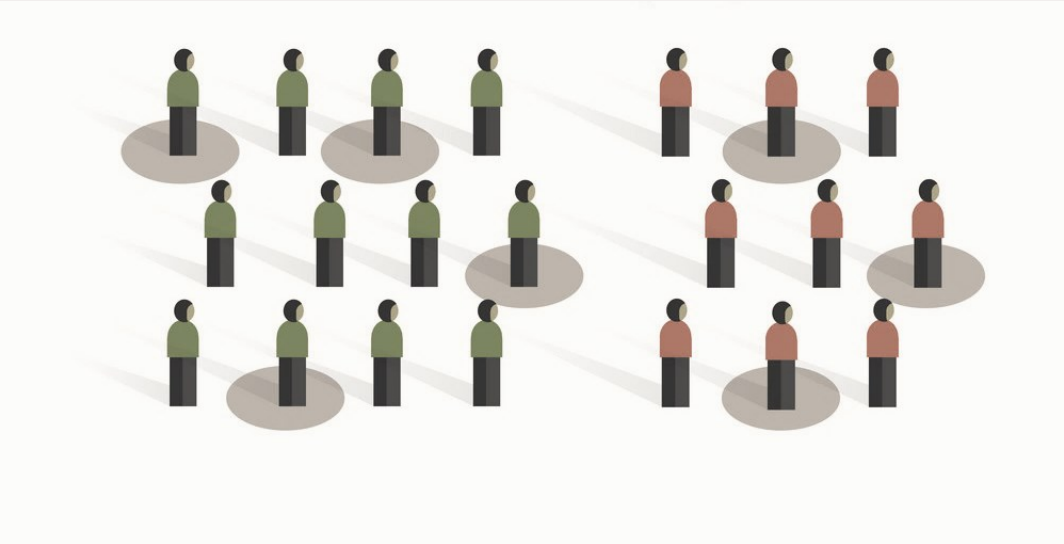








Clustered Samples



Stratified Samples



Complex Samples

As a result of the data collection process, probability samples which deviate from a simple random sample are called complex samples.

- Complex samples can be cost effective.
- Complex samples can ensure sufficient representation of small sub-populations.
- Complex samples can allow researchers to access difficult to access sampling frames.

The Quarter Century Bairns

Imagine you had to design a new nationally representative birth cohort study for Scotland.

Your target population is babies born in Scotland in 2025. You have a substantial budget but the Scottish Government would like to keep the costs contained.

What is your sampling frame?

What is your sampling strategy?

Millennium Cohort Study

- The Millennium Cohort Study (MCS) is a longitudinal birth cohort study which follows the lives of babies born in the UK at the turn of the new millennium.
- The population was stratified by the four UK countries, and each country had two strata (disadvantaged areas and not disadvantaged areas).
- England also had an additional strata for areas with a high proportion of ethnic minority group members.
- The primary sampling unit was the electoral ward.

For a detailed introduction please see: Plewis, I., Calderwood, L., Hawkes, D., Hughes, G., & Joshi, H. (2007). Millennium Cohort Study: technical report on sampling. Centre for Longitudinal Studies, University College London.

Understanding Society (The UK Household Longitudinal Study)

- Understanding Society is one of the largest panel studies in the world.
- It began in 2009 and collects new data annually from sample of approximately 100,000 individuals in 40,000 households in the UK.
- Understanding Society subsumed the British Household Panel Study which started in 1991.
- It includes a newly selected general population sample, an ethnic minority boost sample, a general population comparison sample and an innovation panel sample.
- Each of these elements involve multistage sampling where a sample of addresses were first selected, followed by households and individuals.

For a detailed introduction please see: Lynn, P. (2009). Sample design for understanding society. Understanding Society Working Paper Series No.2009-01. University of Essex.

What are the implications for us as data analysts?

Disproportionate Representation

Certain groups in the population can be disproportionately represented in the data.

Homogeneity

The variance observed within the sampled groups, such as the strata or PSUs, is usually less than the variance between sampled groups.

The homogeneity of observations within the sampled groups can violate the independence assumption which underlies inferential statistics.

Sampling Error

Sampling error is the difference between a survey estimate and the population value due to the random selection of individuals or households to include in the sample.

There are several measures of sampling error which you should be familiar with including confidence intervals, standard errors and p-values.

How do we adjust our analyses?


SPSS®


R


STATA®

What can be done?

Weights

Weights assign a value to each case in your sample to indicate how much ‘weight’ it should have during an analysis.

The weight will be variable in your dataset.

A Simple Example

ID Number	Age	Weight
1	10	2
2	30	1

The unweighted mean age of these two cases is:

$$(10 + 30) / 2 = 20$$

The weighted mean age of these two cases is:

$$(10 \times 2) + (1 \times 30) / (2 + 1) = 50 / 3 = 16.67$$

A Simple Example

	Males	Females
Target Population	50%	50%
Target number of interviews	200	200
Achieved number of survey interviews	300	100
Ratio of achieved to target interviews	300/200 = 1.5	100/200 = 0.5
Weight	1/1.5 = 0.67	1/0.5 = 2

Design Weights

Design weights are used to compensate for over- or under-sampling of specific cases or for disproportionate stratification.

$$Design\ Weight = \frac{1}{P}$$

Non-Response Weights

Non-response weights are used to compensate for the fact that persons with certain characteristics are not as likely to respond to the survey.

Survey Weights

Sample weights are the inverse of the selection probability of a particular group.

You’re not just ‘Weighting’

For most complex surveys you also need to adjust for clustering and stratification in the survey design.

Design Based Approaches vs Model Based Approaches

Model-Based Approaches

In a model-based approach, the statistical analysis incorporates the clustering and stratification in the sample into the analysis.

A multilevel model would be a model-based approach.

Design-Based Approaches

A design-based approach is a single-level model where the complex features of the sample are addressed in the analysis.

There are multiple design-based approaches:

- Taylor Series Linearization
- Replication Methods (Balanced Repeated Replication and Jackknife Methods).

Kish L and MR Frankel (1974), "Inference from Complex Samples," Journal of the Royal Statistical Society B(36), 1-37.

The Design Based Approach

The Design Based Approach

The built in complex surveys commands (svy) are designed for use with complex survey data in Stata, and the ‘survey’ package is designed for use with complex survey data in R.

Typical survey design characteristics include sampling weights, clustered sampling, and stratification.

Many analytical techniques can be undertaken in Stata or R once the characteristics of the data have been declared, and the results will be suitably adjusted to represent the complexity of the survey’s design.

Primary Sampling Unit

The primary sampling unit (PSU) is the first unit that is sampled.

Strata

Strata are a grouping of individuals who share a common characteristic of interest in the survey design.

Probability (Sampling) Weight

Probability (sampling) weights indicate weighted sampling. An individual's 'p-weight' is equal to the inverse probability of being sampled, or equivalently the number in the population represented.

Finite Population Correction (FPC)

The Finite Population Correction Factor (FPC) is used when you sample more than 5% of a finite population.

```
svyset psu [pweight=weight], strata(strata) fpc(fpc)
```

Primary Sampling Unit Variable

Weight Variable

Strata Variable

Finite Population Correction Variable

Stata Survey Data Reference Manual (Release 18):
https://www.stata.com/manuals/svy.pdf

```
design_object <- svydesign(id=~psu, weights=~weight,
                           strata=~strata, fpc=~fpc,
                           nest=TRUE, data=dataname)
```

Lumley, T. (2024) Package ‘survey’: https://cran.r-project.org/web/packages/survey/survey.pdf

Lumley, T. (2011). Complex surveys: a guide to analysis using R. John Wiley & Sons.

```
svyset psu [pweight=weight], strata(strata) fpc(fpc)
```

svydescribe

svy: tab x
svy: mean x
svy: proportion x

svy: regress y x x
svy: logit y x x x

```
svyset psu [pweight=weight], strata(strata) fpc(fpc)

Summary(design_object)

svymean(~x, design_object)

svytable(~x, design = design_object)

model1 <- svyglm(y ~ x + x + x, family=gaussian,
                 design=design_object)

model2 <- (svyglm(y ~ x + x + x, family=quasibinomial,
                  design= design_object))
```

```
. tab country
```

country	Freq.	Percent	Cum.
England	3,659	52.17	52.17
Wales	1,138	16.23	68.40
Scotland	1,242	17.71	86.11
NI	974	13.89	100.00
Total	7,013	100.00	

Country of Interview	column
England	81.57
Wales	5.073
Scotland	9.309
Northern	4.046
Total	100

country	count	obs
England	6,613.64	3,659.00
Wales	312.20	1,138.00
Scotland	670.91	1,242.00
NI	194.00	974.00
Total	7,790.76	7,013.00

Key: count = weighted count
obs = number of observations

```
. mean rpaygu

Mean estimation                   Number of obs    =         7,013
```

	Mean	Std. Err.	[95% Conf.	Interval]
rpaygu	1746.429	15.83329	1715.391	1777.467

```
. svy:mean rpaygu
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =     121          Number of obs    =       7,013
Number of PSUs   =   1,033          Population size = 7,790.7586
                                    Design df        =         912
```

	Linearized			
	Mean	Std. Err.	[95% Conf.	Interval]
rpaygu	1811.896	24.98695	1762.858	1860.935

	SRS Mean (SE)	Adjusted Mean (SE)
Take Home Pay	1746.43 (15.83)	1811.90 (24.99)

Linear Regression (Monthly Pay)

	SRS Coef. (SE)	
Age	13.97*** (1.13)	
Female	Ref.	
Male	804.92*** (29.77)	
England	Ref.	
Wales	-280.82*** (42.18)	
Scotland	-40.41 (40.81)	
Northern Ireland	-216.79*** (44.83)	
Constant	901.56*** (50.70)	
n	7013	

Linear Regression (Monthly Pay)

	SRS Coef. (SE)	**Adjusted** Coef. (SE)
Age	13.97*** (1.13)	12.45*** (1.52)
Female	Ref.	Ref.
Male	804.92*** (29.77)	874.27*** (43.66)
England	Ref.	Ref.
Wales	-280.82*** (42.18)	252.21*** (49.76)
Scotland	-40.41 (40.81)	-19.74 (51.13)
Northern Ireland	-216.79*** (44.83)	-194.13*** (47.79)
Constant	901.56*** (50.70)	895.21*** (63.27)
n	7013	7013

Analysing Complex Surveys

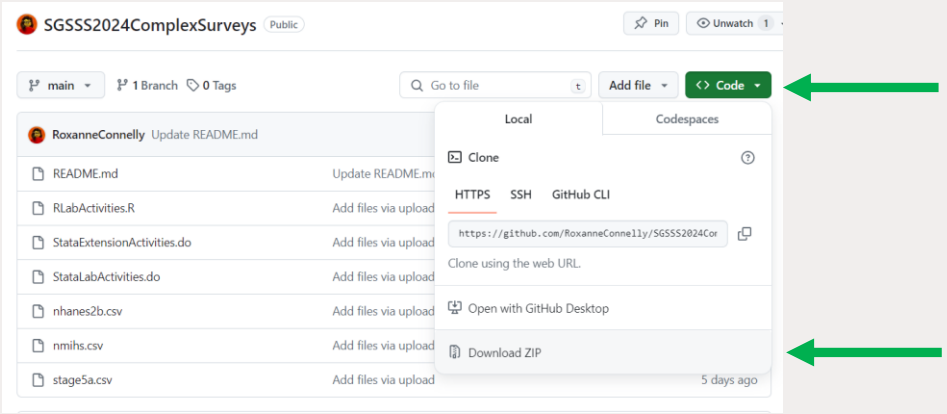
June 2024

Dr Roxanne Connelly

Computer Lab Activity 1

Computer Lab Activity 1

Materials are available here:
https://github.com/RoxanneConnelly/SGSSS2024ComplexSurveys



File: **StataLabActivities.do** or **RLabActivities.R**

Work up to ‘Subpopulation Analysis’

Timetable

1400 - 1445	Talk 1
1445 - 1515	Lab Activities
1515 - 1600	Talk 2
1600 - 1630	Lab Activities
1630 - 1700	Final Advice and Questions

Analysing Complex Surveys

June 2024

Dr Roxanne Connelly

Talk 2

More Complex Issues

Subpopulation Analysis

Subpopulation estimation involves computing point estimates and variance estimates for part of the population (e.g. Women Only, Scotland Only).

This should not be done by taking a subset of the data and carrying out the analysis. This approach would give the correct point estimates, but incorrect standard errors.

To undertake a subpopulation analysis correctly we need to use the selected cases only in the calculation of the point estimate, but all cases for the calculation of standard errors.

Cochran, W (1977). Sampling techniques. John Wiley & Sons.

Subpopulation Analysis

```
svyset [pw=finwgt], strata(stratan)

svy, subpop(highbp): mean birthwgt

svy, subpop(highbp): regress birthwgt age ib1.marital childsex, allbaselevels
```

Stata Corp. Subpopulation Estimation:
https://www.stata.com/manuals13/svysubpopulationestimation.pdf

Subpopulation Analysis

```
nmihs_design <- svydesign(id=~idnum, weights=~finwgt, strata=~stratan, nest=TRUE,
                          data=nmihs)

nmihs_design_high <- subset(nmihs_design, highbp == "hi BP")

svymean(~birthwgt, nmihs_design_high, na.rm = TRUE)
```

Lumley, T. (2024). Estimates in Subpopulations: https://cran.r-project.org/web/packages/survey/vignettes/domain.pdf

Things That Won’t Work

```
. logit diabetes weight age ib2.sex ib2.region, allbaselevels

Iteration 0:  Log likelihood = -1999.7591
Iteration 1:  Log likelihood = -1826.1046
Iteration 2:  Log likelihood =  -1791.471
Iteration 3:  Log likelihood = -1790.8651
Iteration 4:  Log likelihood = -1790.8647
Iteration 5:  Log likelihood = -1790.8647

Logistic regression                             Number of obs = 10,349
                                                LR chi2(6)    = 417.79
                                                Prob > chi2   = 0.0000
Log likelihood = -1790.8647                     Pseudo R2     = 0.1045
```

diabetes	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
weight	.0250883	.0029574	8.48	0.000	.019292	.0308847
age	.0609553	.003856	15.81	0.000	.0533976	.0685131
sex						
Male	-.3883137	.0982487	-3.95	0.000	-.5808777	-.1957497
Female	0	(base)				
region						
NE	.0077536	.1411996	0.05	0.956	-.2689926	.2844998
MW	0	(base)				
S	.1871163	.1252017	1.49	0.135	-.0582745	.4325071
W	-.0995248	.1351211	-0.74	0.461	-.3643574	.1653078
_cons	-8.032417	.346323	-23.19	0.000	-8.711198	-7.353636

```
. svy: logit diabetes weight age ib2.sex ib2.region, allbaselevels
(running logit on estimation sample)

Survey: Logistic regression

Number of strata = 31                           Number of obs    =      10,349
Number of PSUs   = 62                           Population size  = 117,131,111
                                                Design df        =          31
                                                F(6, 26)         =       52.98
                                                Prob > F         =      0.0000
```

diabetes	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
weight	.023655	.0032501	7.28	0.000	.0170263	.0302837
age	.0592227	.0044211	13.40	0.000	.0502059	.0682395
sex						
Male	-.5130463	.1410131	-3.64	0.001	-.8006445	-.2254482
Female	0	(base)				
region						
NE	-.1389128	.132068	-1.05	0.301	-.4082672	.1304417
MW	0	(base)				
S	.144001	.144784	0.99	0.328	-.151288	.4392899
W	-.0952058	.1270315	-0.75	0.459	-.3542882	.1638766
_cons	-7.747746	.3924014	-19.74	0.000	-8.548054	-6.947438

Maximum Likelihood Estimation

Standard maximum likelihood estimation is not possible with complex sample designs because the assumption of independent observations is violated by the stratification and cluster sampling inherent to complex samples.

Binder, D.A. 1981. “On the Variances of Asymptotically Normal Estimators for Complex Surveys.” Survey Methodology 7: 157–170.

Binder, D.A. 1983. “On the Variances of Asymptotically Normal Estimators from Complex Surveys.” International Statistical Review 51: 279–292. Doi: 10.2307/1402588.

Skinner, C. J. (1992). Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes. Computational statistics & data analysis, 13(4), 395-405.

Things You Can't Do

$$x^2 = -2 \text{ x } (LL_0 - LL_M)$$

$$\text{McFadden's Pseudo } R^2 = 1 - (LL_M / LL_0)$$

$$BIC = -2 \text{ x } LL_M + \ln(n) \text{ x df}$$

$$AIC = -2 \text{ x } LL_M + 2 \text{ x df}$$

What do we do?

We estimate the model fit statistics from the naïve model (i.e. we use independently and identically distributed based tests).

Design AIC and Design BIC

Although there is no straightforward likelihood function for survey data, it is possible to construct an analogue of the likelihood-ratio test based on the pseudo-likelihood that has many of the same properties.

- Lumley and Scott built on this to develop dAIC and dBIC
- Implemented in R: `survey`

Lumley, T., & Scott, A. (2017). Fitting regression models to survey data. Statistical Science, 265-278.

Lumley, T., & Scott, A. (2015). AIC and BIC for modelling with complex survey data. Journal of Survey Statistics and Methodology 3 1–18.

Design Pseudo-R^2

Lumley (2017) has also developed design-based versions of pseudo-R^2 measures (Nagelkerke and Cox–Snell pseudo-R^2).

- Implemented in R: `survey`

Lumley, T. (2017). Pseudo-R^2 statistics under complex sampling. Australian and New Zealand Journal of Statistics, 59(2) 187-194.

Issues You Will See In Real Data

Lonely Stratum

In instances where there is one Primary Sampling Unit in a Strata we are unable to calculate standard errors.

```
. svyset psu [pweight = c_indinub_xw], strata(strata)

Sampling weights: c_indinub_xw
             VCE: linearized
     Single unit: missing
        Strata 1: strata
 Sampling unit 1: psu
           FPC 1: <zero>
```

```
.
.                svy: regress c_paynl ib1.sex c_dvage ib1.hiqual_dv, allbaselevels
(running regress on estimation sample)

Survey: Linear regression

Number of strata = 1,834                        Number of obs    =     21,663
Number of PSUs   = 5,166                        Population size = 19,054.806
                                                Design df        =      3,332
                                                F(0, 3332)       =          .
                                                Prob > F         =          .
                                                R-squared        =     0.0443

------------------------------------------------------------------------------
             c_paynl | Coefficient  Linearized
                     |              std. err.      t    P>|t|     [95% conf. interval]
---------------------+--------------------------------------------------------
                 sex |
              1. male |          0  (base)
            2. female |  -997.7342          .        .       .            .           .
                      |
             c_dvage |   36.2206          .        .       .            .           .
                      |
            hiqual_dv |
            1. Degree |          0  (base)
2. Other higher degree |  -962.4378          .        .       .            .           .
       3. A-level etc |  -1304.211          .        .       .            .           .
          4. GCSE etc |  -1703.196          .        .       .            .           .
5. Other qualification |  -2193.398          .        .       .            .           .
   9. No qualification |  -2687.309          .        .       .            .           .
                      |
                 _cons |   2035.223          .        .       .            .           .
------------------------------------------------------------------------------

Note: 40 strata omitted because they contain no population members.
Note: Missing standard errors because of stratum with single sampling unit.
```

```
. svyset psu [pweight = c_indinub_xw], strata(strata) singleunit(scaled)

Sampling weights: c_indinub_xw
             VCE: linearized
     Single unit: scaled
        Strata 1: strata
 Sampling unit 1: psu
           FPC 1: <zero>
```

The ‘singleunit()’ specifies the method to handle strata with one sampling unit.

missing – leave missing values for standard errors (default)

certainty – Strata with single units do not contribute to the contribution to the standard error

scaled – a scaled version of certainty, the scaling factor comes from the average of the variances from the strata with multiple sampling units.

centered – strata with single units are centered at grand mean

```
. svy: regress c_paynl ib1.sex c_dvage ib1.hiqual_dv, allbaselevels
(running regress on estimation sample)

Survey: Linear regression

Number of strata = 1,834                    Number of obs    =     21,663
Number of PSUs   = 5,166                    Population size = 19,054.806
                                            Design df        =      3,332
                                            F(7, 3326)       =      74.73
                                            Prob > F         =     0.0000
                                            R-squared        =     0.0443
```

c_paynl	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
sex						
1. male	0	(base)				
2. female	-997.7342	76.2321	-13.09	0.000	-1147.201	-848.2677
c_dvage	36.2206	3.207816	11.29	0.000	29.93111	42.51008
hiqual_dv						
1. Degree	0	(base)				
2. Other higher degree	-962.4378	137.4565	-7.00	0.000	-1231.945	-692.9301
3. A-level etc	-1304.211	131.0315	-9.95	0.000	-1561.121	-1047.3
4. GCSE etc	-1703.196	112.6084	-15.12	0.000	-1923.985	-1482.408
5. Other qualification	-2193.398	140.9287	-15.56	0.000	-2469.714	-1917.082
9. No qualification	-2687.309	124.6013	-21.57	0.000	-2931.612	-2443.006
_cons	2035.223	146.6145	13.88	0.000	1747.76	2322.687

```
Note: 40 strata omitted because they contain no population members.
Note: Variance scaled to handle strata with a single sampling unit.
```

Lonely Stratum

In instances where there is one Primary Sampling Unit in a Strata we are unable to calculate standard errors.

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

Stata Survey Data Reference Manual (Release 18):
https://www.stata.com/manuals/svy.pdf

Lonely Stratum

In instances where there is one Primary Sampling Unit in a Strata we are unable to calculate standard errors.

This issue is dealt with in R using the ‘lonely.psu’ options as part of the ‘survey’ package.

The default is ‘fail’ which gives an error message, other options are ‘ignore’, ‘adjust’, and ‘average’.

Lumley, T. (2024) Package ‘survey’: https://cran.r-project.org/web/packages/survey/survey.pdf

Which Weight?

Weights assign a value to each case to indicate how much ‘weight’ it should have during data analysis.

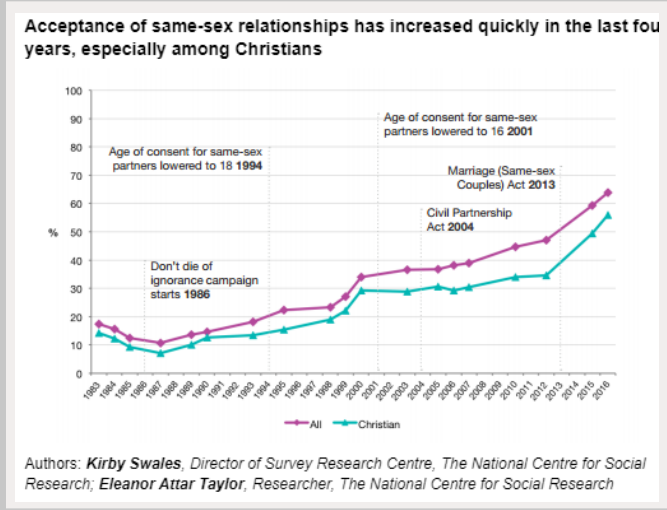
Most large social survey datasets contain more than one weight.

The details of the survey weights should be included in the data documentation.

British Social Attitudes Survey Weights

There are two weights provided with the 2018 British Social Attitudes Survey dataset.

- A weight for analysing variables from the CAPI interview (WtFactor).
- A weight for analysing variables from the self-completion questionnaire (WtFactorSC).



UKHLS Weights

There are many weights provided in the UKHLS datasets.

For example, separate sets of weights are provided for:

- The combined GPS and EMBS (from Wave 1)
- The former BHPS sample (from Wave 2)
- The combined GPS, EMBS and BHPS (from Wave 2)
- The combined GPS, EMBS, BHPS and IEMBS (from Wave 6)

Table: Naming convention for Understanding Society weights w_XxxYyZz_aa

Wave letter	Who are you studying?	Which questions(naire)?	Which sample/timeline?	Analysing one wave or across waves?
w_ (a to i)	**Xxx** (Hhold or individual)	**Yy** (instrument)	**Zz** (samples cover different waves)	**_aa** (cross-sectional/longitudinal)
a_ **b_** **c_** **d_** **e_** **f_** **g_**	**hhd:** household **psn:** persons 0+ **ind:** persons 16+ **yth:** persons 10-15	**en:** enumeration **in:** interview **px:** interview or proxy **5m:** "extra 5 minutes" **sc:** self-completion **ns:** nurse visit **bd:** blood	**us:** GPS & EMB (W1>) **bh:** BHPS (W2>) **ub:** GPS, EMB & BHPS (W2-W5) **ui:** GPS, EMB, BHPS & IEMB (W6>) **91:** BHPS original sample (91> excl. N.I.) **01:** BHPS original sample + boosts	**_xw:** cross-sectional analysis weight **_lw:** longitudinal weight **_xd:** x-sectional design weight **_li:** longitudinal inclusion weight

https://www.understandingsociety.ac.uk/documentation/mainstage/user-guides/main-survey-user-guide/selecting-the-correct-weight-for-your-analysis

There isn’t a weight for my analysis!

By carefully considering the potential available weights you should be able to choose a weight that is appropriate, even if it potentially suboptimal in some respect.

You could potentially derive your own weights, but I would warn against this unless you are an advanced analyst.

Other Weighty Issues

Weights can have a value of zero (on purpose).

You should also check for missing values in your weights.

More Complex Analyses

More Complex Analyses

It is possible to estimate most standard regression models whilst adjusting for complex survey designs. It is also possible to take complex survey designs into account with many more analysis techniques (e.g. panel data analysis, structural equation modelling, latent class analysis).

You may come across analyses that have not been incorporated into the survey packages of your software.

It is possible to adjust for complex survey designs when running multilevel models, but you will need to specify weights at each level of your model. These additional weights will generally not be available in multipurpose omnibus data sets.

More Complex Analyses

What can you do?

Present appropriately adjusted analyses where you are able.

Undertake sensitivity analysis of alternative analysis options.

Present a naïve analysis and be fully transparent about what you have done (and consider the limitations of this).

Analysing Complex Surveys

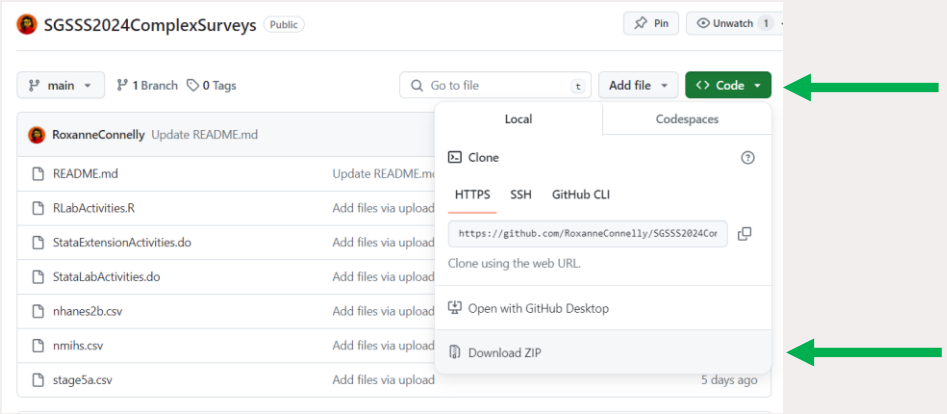
June 2024

Dr Roxanne Connelly

Computer Lab Activity 2

Computer Lab Activity 2

Materials are available here:
https://github.com/RoxanneConnelly/SGSSS2024ComplexSurveys



File: **StataLabActivities.do** or **RLabActivities.R**

Additional extension activities with ‘real data’ are available in Stata:
StataExtensionActivities.do

Timetable

1400 - 1445	Talk 1
1445 - 1515	Lab Activities
1515 - 1600	Talk 2
1600 - 1630	Lab Activities
1630 - 1700	Final Advice and Questions

Analysing Complex Surveys

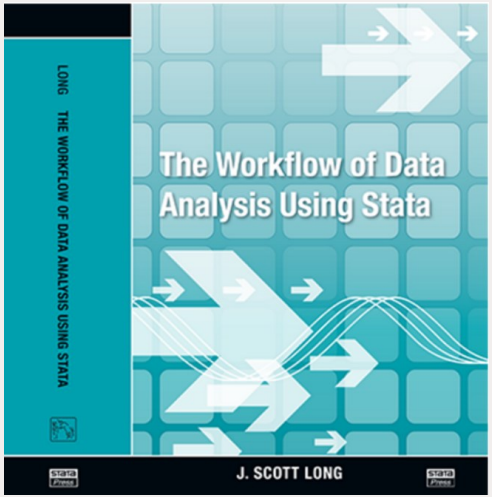
June 2024

Dr Roxanne Connelly

Final Advice

Final Advice: Work on Your Workflow

"The workflow involves the entire process of data analysis including planning and documenting your work, cleaning data and creating variables, producing and replicating statistical analyses, presenting findings, and archiving your work." (Long 2009: 1)



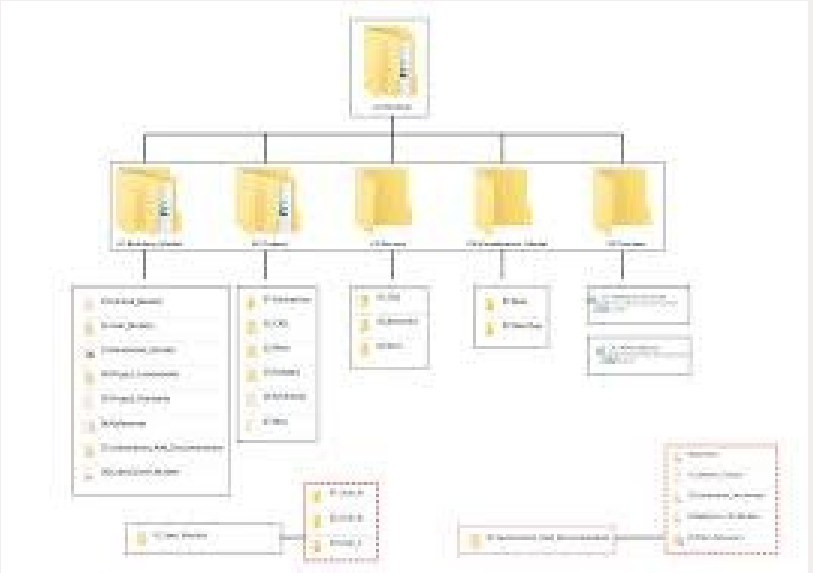
Long, J. S. (2009). The workflow of data analysis using Stata. Stata Press.

Gayle, V. (2020). NCRM Online Teaching Resource: The Data Analysis Workflow.

File Structures

A consistent filing structure is of great benefit to us as data analysts:

- It’s easier to find and save files;
- It’s easier for collaboration;
- It’s easier to pick up projects.



File Naming

report_final.docx

report_final2.docx

report_finalthisone.docx

statsreport_20201122_v12_rc.docx

statsreport_20201124_v13_rc.docx

statsreport_20201125_v14_dw.docx

statsreport_20201127_v15_rc.docx

Final Advice: Cite your Data

Useful blog post covering why you should cite your data, and how to cite your data:
https://blog.ukdataservice.ac.uk/spotlight-on-citethedata-make-the-data-count/

Final Advice: Share your Code

WORLD VIEW · 24 MAY 2018

Before reproducibility must come preproducibility

Instead of arguing about whether results hold up, let's push to provide enough information for others to repeat the experiments, says Philip Stark.

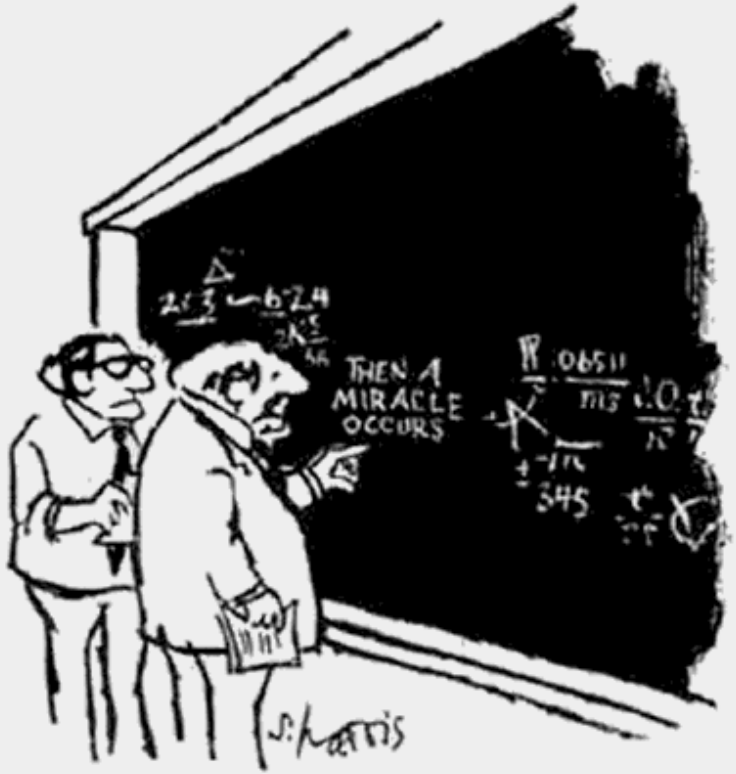
Philip B. Stark

Stark, P. B. (2018). Before reproducibility must come preproducibility. Nature, 557(7706), 613-614.

“A published paper should be considered as an advertisement of the research.”

(Claerbout 1994)

Claerbout, J. (1994). Seventeen years of super computing and other problems in seismology [Paper presentation]. National Research Council Meeting on High Performance Computing in Seismology. http://sepwww.stanford.edu/sep/jon/nrc.html



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

Final Advice: Date Your Data

Final Advice: Presenting Results

General advice: automate the production of your tables

Stata Resources:

- Stata Reporting Manual: https://www.stata.com/manuals/rpt.pdf
- NCRM Producing Automated Outputs (in Stata): https://www.ncrm.ac.uk/resources/online/all/?id=20788
- Introduction to –etable- and –collect- commands: https://osf.io/6h7gm/

R Resources:

- stargazer: beautiful LATEX, HTML and ASCII tables from R statistical output (Marek Hlavac): https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf
- Publication-ready APA tables (Rémi Thériault): https://cran.r-project.org/web/packages/rempsyc/vignettes/table.html
- NCRM Producing Automated Outputs (using R): https://www.ncrm.ac.uk/resources/online/all/?id=20832

Final Advice: Presenting Results

Present your adjusted and unadjusted results.

Consider sharing sensitivity analyses as an online supplement.

Tell your reader that your analysis has been adjusted to take into account the sample design.

Table 3: Multinomial Logistic Regression of Subjective Health (Base: Poor).

	Fair Log Odds (SE)	Average Log Odds (SE)	Good Log Odds (SE)	Excellent Log Odds (SE)
Sex				
Male	Ref.	Ref.	Ref.	Ref.
Female	0.38*** (.09)	0.20 (.11)	0.16 (.11)	-0.02 (.12)
Ethnicity				
Non-Black	Ref.	Ref.	Ref.	Ref.
Black	-0.45* (.19)	-0.89*** (.18)	-1.64*** (.19)	-1.88*** (.22)
Age (Years)	-0.02*** (.00)	-0.05*** (.00)	-0.07*** (.00)	-0.08*** (.00)
Constant	1.92*** (.21)	4.31*** (.22)	5.42*** (.24)	5.88*** (.24)
McFadden's Pseudo R^2	0.06			
n	10335			

Note: Model is adjusted for the complex sample design (*** p<.001, ** p<.01, * p<.05).
McFadden's Pseudo R2 is estimated from the non-adjusted model.

Should all my analyses be adjusted?

There may be reasons to present unadjusted analyses when presenting descriptive statistics.

Table 1. Descriptive Statistics

	n	%
Sex		
Male	911	54.39%
Female	764	45.61%
Drink Alcohol (At least once in last month)		
Yes	637	38.03%
No	1,038	61.97%
	Mean	**SD**
Age (Years)	13.67	0.94
Family Dinner (# days per week)	4.69	2.35
n		1675

Note: Percent, mean and standard deviation are adjusted for the complex sample design.

Table 1: Descriptive Statistics

	Unadjusted	Unadjusted	Adjusted
	n	%	%
Sex			
1. Male	4915	47.49%	47.95%
2. Female	5434	52.51%	52.05%
Age group			
1. 20–29	2320	22.42%	28.05%
2. 30–39	1621	15.66%	20.42%
3. 40–49	1271	12.28%	16.83%
4. 50–59	1291	12.47%	16.72%
5. 60–69	2860	27.64%	13.35%
6. 70+	986	9.53%	4.62%
Rural			
0. Urban	6547	63.26%	68.26%
1. Rural	3802	36.74%	31.74%
		Mean (SD)	Mean (SD)
Weight (kg)		71.90 (0.15)	71.90 (0.17)
Height (cm)		167.65 (0.09)	168.46 (0.15)
n	10349		

Data Source: nhanes2b.
Percentages, mean and standard deviation are adjusted for sample design.

Final Advice: A Warning!

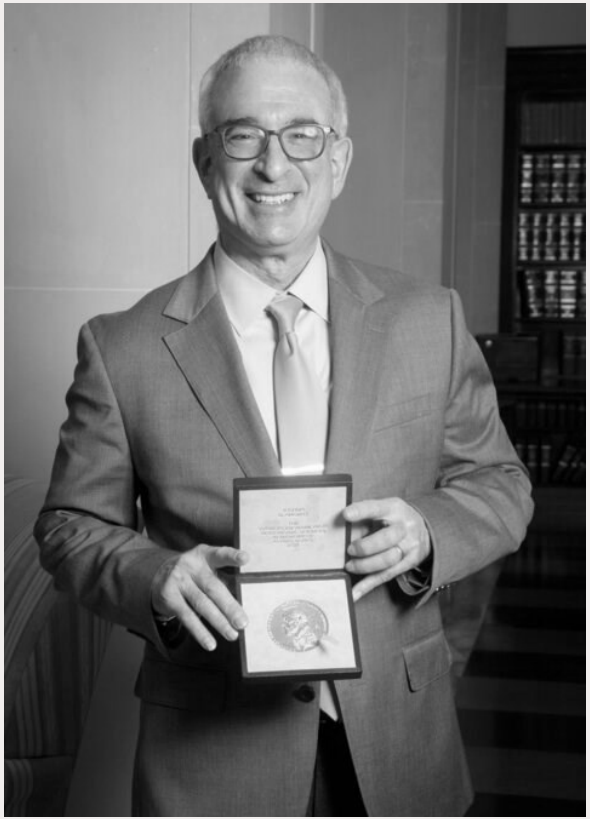
‘NEVER conduct unweighted analysis if you aim to generalise your results to the UK population’

(Understanding Society – UK Household Longitudinal Study: Wave 1, 2009-2010 User Manual – 24 October 2011)

Final Advice: Another Warning!

You cannot assume a priori that complex samples will have no impact on your results.

Finally



Finally

“Few things are as confusing to applied researchers as the role of sample weights. Even now, 20 years post-Ph.D., we read the section of the Stata manual on weighting with some dismay.”

(Angrist and Pischke, 2008, p. 92)

Angrist, J. D., & Pischke, J. S. (2008). Mostly harmless econometrics: An empiricist's companion. Princeton university press.

Analysing Complex Surveys

June 2024

Dr Roxanne Connelly

Questions

Analysing Complex Surveys

June 2024

Dr Roxanne Connelly

roxanne.connelly@ed.ac.uk