Analysing Large-Scale and Complex Social Survey Data: A Brief But Practical Introduction for Non-Academic Researchers

Vernon Gayle  
University of Edinburgh  
vernon.gayle@ed.ac.uk  
@profbigvern

Roxanne Connelly  
University of Warwick  
R.Connelly@warwick.ac.uk  
@ConnellyRoxanne

An output from the project ‘Have Socio-Economic Inequalities in Childhood Cognitive Test Scores Changed? A Secondary Analysis of Three British Birth Cohorts’ Funded by the ESRC Secondary Data Analysis Initiative  
https://tinyurl.com/h3nj59b.
Part 1

Welcome
Complex Social Surveys

You might never conduct a complex survey…

You might never analyse a complex survey dataset…

But you may need to read results from the analysis of complex social survey and therefore wish to understand more
Large Facilities Capital Fund
Complex Social Surveys

- Large in scale \((n = \text{big})\)
- Nationally representative
- Support analyses of subpopulations
- Multipurpose (omnibus)
- Many data files
- A large number of measures \((k = \text{big})\)
- Longitudinal data has repeated contacts \((t_1, t_2, t_3)\)
Complex Social Surveys Good Points

- Multipurpose (omnibus)
- Nationally representative
- Support analyses of subpopulations
- Support analysis of social change / social stability
Complex Social Surveys Less Good Points

- Designs are complex
- Selection and sampling strategies complex
- Dealing with many files
- Matching individual records over time
- Linking records (spouses, households, siblings, regional info)
- Making sense of measures over time (e.g. qualifications, neighbourhoods)
Prelude

Gayle and Lambert (2006)

Over the decades a virtual industry producing critiques of survey analyses from various standpoints has developed. We suggest that arguments for and against the analysis of data from social surveys, have at times resembled a caricature not dissimilar to the Shakespearean feud between the Montagues and the Capulets.
Prelude

Gayle and Lambert (2006)

Goldthorpe (2000) notes that critics of survey based social science research ritually characterise it as static and this is simply to ignore the rapid development of survey related work.
Goldthorpe (2000) notes that critics of survey based social science research ritually characterise it as static and this is simply to ignore the rapid development of survey related work.

Far from remaining static in their practices, survey researchers in social science have been reflexive and have attempted to respond to various critiques and specific shortcomings that have been highlighted.

These responses have usually involved improving statistical techniques of variable analysis, and improving the underlying data quality.

http://www.restore.ac.uk/Longitudinal/wp/lda_2006_3.pdf
Part 2

Revision (terminology etc.)
The Social Survey

Think of this as a data collection approach rather than a single technique

A common set of

‘systematic’ measures (variables) are collected on a set of cases
Cases

Entities such as:

Individuals, households and families
Cases

Entities such as:

- Individuals, households and families
- Pupils, graduates, ageing citizens
- Organisations, schools, firms, farms (or even police stations)
- Neighbourhoods, counties, states, regions, countries
Variables

Systematic measures
Variables

Systematic measures

Outcome measures (Dependent Variables, DV, Y variable)
- life satisfaction, age at death
Variables

Systematic measures

Outcome measures (Dependent Variables, DV, Y variable)
- life satisfaction, age at death

Explanatory measures (Independent Variables, IV, X variable)
- gender, occupation, ethnicity
Variables

*Be mindful...*

In one analysis *level of education* might be an outcome

In another analysis *level* of education might be an explanatory variable
Outcome

Explanatory variable
The Variable by Case Matrix

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Gender</th>
<th>Ethnicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t0caseid</td>
<td>t0source</td>
<td>t0sex</td>
<td>t0house</td>
</tr>
<tr>
<td>----------</td>
<td>----------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>5</td>
<td>301402654</td>
<td>male</td>
<td>rented</td>
</tr>
<tr>
<td>6</td>
<td>301402656</td>
<td>female</td>
<td>rented</td>
</tr>
<tr>
<td>7</td>
<td>301402660</td>
<td>female</td>
<td>owned</td>
</tr>
<tr>
<td>8</td>
<td>301402661</td>
<td>female</td>
<td>owned</td>
</tr>
<tr>
<td>9</td>
<td>301402666</td>
<td>male</td>
<td>rented</td>
</tr>
<tr>
<td>10</td>
<td>301402668</td>
<td>female</td>
<td>rented</td>
</tr>
<tr>
<td>11</td>
<td>302400202</td>
<td>female</td>
<td>rented</td>
</tr>
<tr>
<td>12</td>
<td>302400203</td>
<td>male</td>
<td>rented</td>
</tr>
<tr>
<td>13</td>
<td>302400205</td>
<td>male</td>
<td>owned</td>
</tr>
<tr>
<td>14</td>
<td>302400206</td>
<td>female</td>
<td>other</td>
</tr>
<tr>
<td>15</td>
<td>302401101</td>
<td>female</td>
<td>owned</td>
</tr>
</tbody>
</table>
Part 3

UK SURVEYS
• Major UK surveys
• Government surveys
• Opinion polls
• Historical data
• Qualitative data
Birth Cohort Studies

• Longitudinal i.e. repeated contacts with same individuals

• Principally concerned with charting the development of the ‘group’ from a certain point in time

• Data collected by parental interview early in life

• Data collected from the cohort member later (e.g. childhood onwards)
Lifelong Health & Ageing: LHA
National Survey of Health and Development: NSHD

The MRC Unit for Lifelong Health and Ageing at UCL (LHA) is the home of the MRC National Survey of Health and Development (NSHD).

The NSHD, the oldest of the British birth cohort studies, is unique in having data from birth on the health and social circumstances of a representative sample (N=5362) of men and women born in England, Scotland or Wales in March 1946.
Surveys

Since your parents were first interviewed when they were born, there have been ten further surveys of NCDS study members. In 2018, we will start the 11th survey at age 60.

In this section

**Age 55 Survey**

The Age 55 Survey took place between 2013 and 2014. This was the first time we asked you to complete a survey via the internet. We asked you about your relationships and work, your expectations for retirement, and your caring responsibilities. More than 9,000 of you took part with two thirds doing so online.

**Age 50 Survey**

The Age 50 Survey took place in 2008. Age 50 marked a big milestone in your lives, and this survey aimed to find out about what your lives were like at this age. We even asked you to write a little about how you imagined your life to be at age 60. Just under 10,000 of you took part.

**Age 46 Survey**

The Age 46 survey took place in 2004 and was conducted via telephone. We asked you about your lives in middle age, including your relationships, work, use of computers, smoking and drinking habits, and your experience of crime. Just under 10,000 of you took part.

**Biomedical Survey (Age 44/45)**

The biomedical survey took place between 2002 and 2003 when you were 44/45 and aimed to explore the factors associated with health in early middle age. A set of measures were taken and many of you gave blood, and saliva for analysis. Over 9,000 of you took part.

**Age 42 Survey**

The Age 42 survey took place in 2000. This was the first survey that was conducted by an interviewer with a computer rather than by pen and paper. Over 11,000 of you took part.
Welcome to the 1970 British Cohort Study

Principal Investigator: Prof Alice Sullivan

The 1970 British Cohort Study (BCS70) follows the lives of more than 17,000 people born in England, Scotland and Wales in a single week of 1970. Over the course of cohort members lives, the BCS70 has collected information on health, physical, educational and social development, and economic circumstances among other factors.

The BCS70 is managed by CLS and funded by the Economic and Social Research Council.

You can find out more about the 1970 British Cohort Study by viewing an interview with its Principal Investigator, Alice Sullivan below.

A conversation with Alice Sullivan, P...
Welcome to the Millennium Cohort Study

Principal Investigator: Prof Emla Fitzsimons

The Millennium Cohort Study (MCS) is a multi-disciplinary research project following the lives of around 19,000 children born in the UK in 2000-01. It is the most recent of Britain's world-renowned national longitudinal birth cohort studies. The study has been tracking the Millennium children through their early childhood years and plans to follow them into adulthood. It collects information on the children's siblings and parents. MCS's field of enquiry covers such diverse topics as parenting; childcare; school choice; child behaviour and cognitive development; child and parental health; parents' employment and education; income and poverty; housing, neighbourhood and residential mobility; and social capital and ethnicity.

The study is core funded by the Economic and Social Research Council (ESRC) and a consortium of Government departments.

You can find out more about the Millennium Cohort Study by watching an interview with its Principal Investigator, Emla Fitzsimons below.
Longitudinal studies track the same people throughout their lives, helping us understand how life in the UK is changing. CLOSER is enabling researchers worldwide to get the most out of these valuable studies.
Household Panels

- Longitudinal i.e. repeated contacts with same individuals
- The ‘panel’ of respondents are followed through time
- Surveyed (e.g. interviewed) regularly (e.g. each year)
The British Household Panel Survey began in 1991 and is a multi-purpose study whose unique value resides in the fact that:

- It follows the same representative sample of individuals - the panel - over a period of years;
- It is household-based, interviewing every adult member of sampled households;
- It contains sufficient cases for meaningful analysis of certain groups such as the elderly or lone parent families.

The wave 1 panel consists of some 5,500 households and 10,300 individuals drawn from 250 areas of Great Britain. Additional samples of 5,500 households in each of Scotland and Wales were added to the main sample in 1999, and in 2001 a sample of 2,000 households was added in Northern Ireland, making the panel suitable for UK-wide research.

- BHPS wave 1 data and documentation are available from the UK Data Archive.
- The Wave 1 BHPS online documentation, which includes pdf versions of the latest Volume A and Questionnaires, is also available.

As part of wave 10 BHPS participants were asked in the UK Data Archive to join the Understanding Society.

Almost 6,700 of the 8,000 BHPS participants invited to join did so. First interviews with BHPS participants in Understanding Society were carried out in wave 2 of the study in 2010-2011.

Find out more about Understanding Society.
BHPS Structure

• Longitudinal study (repeated contacts with same individuals)

• Began in 1991

• Multipurpose (or omnibus) survey

• Data suitable for investigating a wide range of research questions

• Core set of questions & rotating modules of questions
BHPS Data Collection

- Interviews
- Annual (usually in the autumn)
- Wave A 1991 through to Wave R 2008
- All members of sampled households aged sixteen plus
- Forty-five minute interview
- Short household questionnaire (one member)
BHPS Topics

- housing composition
- housing conditions
- residential mobility
- education and training
- health and the use of health services
- labour market behaviour
- socio-economic values
- income from employment, benefits and pensions
BHPS Topics

• New questions were sometime introduced and they often followed on from changing policies and changing research agendas

• Variable component has included questions on wealth and assets, additional health measures, ageing in retirement, quality of life, children and parenting, views on crime and neighbourhood and social networks

• Buck et al. (1994) showcase the wide range of topics that can be analysed using BHPS data
About the study

Understanding Society is an innovative world-leading study about 21st century UK life and how it is changing. It captures important information about people’s social and economic circumstances, attitudes, behaviours and health. The study is longitudinal in its design and of high quality.

Largest household study of its kind

We interview the same people in the same households each year to build a longitudinal picture of how their lives are changing over time.

Represents the diversity of the UK

Our participants live in Scotland, Wales, Northern Ireland and England and all of Britain’s ethnic and immigrant groups are fully represented in the study.

Issues that affect all our lives

Our questionnaires cover a wide variety of themes such as family, education, finance, employment, health and wellbeing. Read about the carefully-chosen questions we ask.
Data and documentation
User guides, fieldwork documents, questionnaires, technical reports

Data collected from the survey’s thousands of participants is securely stored by the UK Data Service, from where researchers can access it online. All of the associated documentation is available here.

Understanding Society – a design overview, published in the Longitudinal and Life Course Studies 2012 Volume 3 Issue 1, also provides a helpful introduction.

Main survey
Documentation relating to all households interviewed for the survey
- User guide
- Dataset documentation
- Questionnaires
- Technical reports
- Fieldwork documents
- Long term content plan
- Quality profile

Health, Biomarkers and Genetics
Documentation relating to Health, Biomarkers and Genetics data
- User guide
- Dataset documentation
- Questionnaires
- Fieldwork documents

See Key Features for an overview

Innovation panel
Documentation for households in the Innovation Panel
- User guide
- Dataset documentation
- Questionnaires
- Technical reports
- Fieldwork documents

See Key Features for an overview

Getting started
Access the documentation and data

Help and support
See also training courses and FAQs

Data releases
Information about latest releases of data and updates

Citation
Bibliographic citations and acknowledgements
Understanding Society –
The UK Household Longitudinal Study (UKHLS)

• Largest household panel survey in the world

• UKHLS shares common design and data collection characteristics with the BHPS

• Began in January 2009

• Approximately 40,000 UK households

• The UKHLS is available from the UK Data Archive (Study Number 6614)
Understanding Society –
The UK Household Longitudinal Study (UKHLS)

Approximately

- General population survey 26,000 households
- Ethnic minorities boost sample 4,000 households
- BHPS 8,400 households
- Innovation Panel 1,500 household
Youth self-completion questionnaire (for youths aged 10-15 years)

Individual questionnaire (one per eligible adult, aged 16 and above)
Part 4

Study Designs
A Census

A survey of an ‘entire’ or complete population \( (n=all) \)

In the UK the Decennial Census
A Census

A survey of an ‘entire’ or complete population \((n=\text{all})\)

In the UK the Decennial Census

Began in 1801 every 10 years (except 1941)

1911 Boycott - Emily Wilding Davison was sustained by meat lozenges and lime juice as she hid in a broom cupboard in the Houses of Parliament

http://www.theguardian.com/lifeandstyle/2011/apr/01/suffragettes-census-1911-boycott
27 March 2011

Why the census matters

The census is the official count of every person and household in Scotland. It is held every 10 years and helps to plan our future public services.

Please fill in this questionnaire on, or around, Sunday 27 March. Please include everyone at this address. It shouldn’t take long and you can fill it in online.

As a householder, you have a legal duty to fill in this questionnaire. If you don’t, or if you supply false information, you may be fined up to £1,000.

Your personal information is protected by law and we will keep it confidential for 100 years.

Thank you for helping to shape Scotland’s future.

Duncan Macniven
Registrar General for Scotland

Need help?

- www.scotlandscensus.gov.uk
- Helpline 0300 123 1702
- Textphone 18001 0300 123 1703

Please fill in this questionnaire:

- online at www.scotlandscensus.gov.uk

Enter the Internet Questionnaire Access Code:

You can fill in this questionnaire online in English or Gaelic.

Or

Fill in this paper version and post it back using the pre-paid envelope provided.

Declaration

I have filled in this questionnaire fully and accurately, as far as I know.

Signature(s): ____________________________

Date: ____________________________
A ‘Sample’ Survey

Collects information on a ‘sample’ or sub-set of the population

Early examples include…

Charles Booth Survey into life and labour of the people in London (1886-1903)
Seebohm Rowntree Study of York (1899, 1935, 1951)
Selecting Samples
Non-Probability Samples

• DO NOT have a formal statistically informed method of selecting cases

• Are NOT statistically representative of the population of study (but might be sociologically informative)
Non-Probability Samples

The Great British Class Survey was carried out online. It was widely publicized by the BBC, and completed by 298,571 participants. The GBCS was a non-representative, non-random-sample survey.

The Burke and Hare analogy...
Probability Samples

Are designed to be statistically representative of the population of study

First used by Arthur Lyon Bowley in New Survey of London Life and Labour
Probability Samples

Each case in the population has a ‘probability’ of being included in the sample.
Sampling Frame = list of population

Sampling Fraction = ratio of sample to population
Systematic Samples
Simple Random Sample

Each case has an ‘equal and non-zero’ chance of inclusion
Exercise 1

Study Designs
A Systematic Sample

1. Buttons
2. Cinderella
3. Robin Hood
4. Sleeping Beauty
5. Prince Charming
6. Snow White
7. Aladdin
8. Little Red Riding Hood
9. Peter Pan
10. Goldilocks

Sampling Fractions 1/2
A Systematic Sample

1. Buttons
2. Cinderella
3. Robin Hood
4. Sleeping Beauty
5. Prince Charming
6. Snow White
7. Aladdin
8. Little Red Riding Hood
9. Peter Pan
10. Goldilocks

Sampling Fractions 1/2
A Simple Random Sample

1. Buttons
2. Cinderella
3. Robin Hood
4. Sleeping Beauty
5. Prince Charming
6. Snow White
7. Aladdin
8. Little Red Riding Hood
9. Peter Pan
10. Goldilocks
END OF EXERCISE
Part 5

Selecting Samples
Simple Random Samples

Each case has an ‘equal and non-zero’ chance of inclusion

A high degree of representativeness…
Simple Random Samples

• Each case has an ‘equal and non-zero’ chance of inclusion

• A high degree of representativeness…

• Random selection greatly reduces selection biases

• Inference from the ‘sample’ data can be drawn about the population
Not a simple random sample

Each case does not have an ‘equal and non-zero’ chance of inclusion
Simple Random Sample

Each case has an ‘equal and non-zero’ chance of inclusion
Stratified Samples

Stratified Random Sampling
Stratified Random Sample

- The stratified random sample is usually an improvement on the simple random sample.

- It better reflects the underlying pattern of the population being studied.
Stratified Random Sample

In my PhD I stratified the sampling frame by Gender, Faculty, Age Category and then drew my sample(s)

I wanted to get a highly representative sample that reflected the student population of my university
Clustered Samples
Cluster(ed) Samples

Cases are chosen from 'some' clusters

Each case no longer has an 'equal and non-zero' chance of inclusion

Practical fieldwork (e.g. interview) and cost benefits
Cluster(ed) Samples

Cases are chosen from ‘some’ clusters

Access negotiations for fewer organisations
Stratified Cluster Samples
Multistage Cluster

1. Region
2. Educational Authorities
3. Primary Schools
4. Parents
Part 6

The Design of Some Major Surveys
1958 British birth cohort (National Child Development Study)
All births that took place during March 3–9, 1958

1970 British Birth Cohort (BCS70)
All births that took place during April 5–11, 1970
BHPS Sample Design

• A stratified design drawn from the Postcode Address File (PAF)

• All residents present at the selected addresses at the first wave of the survey were designated as panel members

• These same individuals were re-interviewed each successive year
Following BHPS Members

• If sample members split away from the original household to form new households they were still followed

• All adult members of their new households were interviewed

• This is a very important feature of the BHPS design because it led to more wide-ranging data collection
BHPS Samples

- The original BHPS survey contained 10,264 respondents.

  - Wave A (1991)
    - 4,833 men and 5,431 women
    - 5,511 households

  - Wave R (2008) 14,419 respondents living in 8,144 households

- By the time of the formal end of the BHPS, 43,272 adults had been interviewed at some time during the eighteen waves of the survey.
Extension Samples
The British Household Panel Survey

- Welsh extension from 1999 (Wave 9) (n=1500 households)
- Scottish extension from 1999 (Wave 9) (n=1500 households)

- Northern Ireland extension from 2001 (Wave 11) (n=1900 households)
  - Why was the Northern Ireland sample not clustered?

- Former European Community Household Panel survey low-income sub-sample from 1997 to 2001 (Waves 7 to 11) (n=1,000 households)
Understanding Society –
The UK Household Longitudinal Study (UKHLS)

- UKHLS shares common design and data collection characteristics with the BHPS
- Began in January 2009
- 40,000 UK households
- Largest household panel survey in the world
- Hobcraft and Sacker (2011) provide an overview of the origins and development of the UKHLS
- The UKHLS is available from the UK Data Archive (Study Number 6614)
Understanding Society –
The UK Household Longitudinal Study (UKHLS)

1. General population survey
2. Ethnic minorities boost sample
3. BHPS
4. Innovation Panel
Understanding Society –
The UK Household Longitudinal Study (UKHLS)

- Wave A (a.sex)

- 50,994 adult interviews
- 23,202 males (46%)
- 27,792 females (46%)
Understanding Society – The UK Household Longitudinal Study (UKHLS)

Country of Residence Wave A (a_country)

<table>
<thead>
<tr>
<th>Country</th>
<th>Count (Percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>England</td>
<td>42,972 (84%)</td>
</tr>
<tr>
<td>Wales</td>
<td>2,378 (5%)</td>
</tr>
<tr>
<td>Scotland</td>
<td>3,556 (7%)</td>
</tr>
<tr>
<td>Northern Ireland</td>
<td>2,088 (4%)</td>
</tr>
</tbody>
</table>
Understanding Society –
The UK Household Longitudinal Study (UKHLS)

1. General population survey
2. Ethnic minorities boost sample
3. BHPS
4. Innovation Panel
• The general population survey is a stratified clustered equal probability sample of residential addresses drawn to a uniform design throughout all of the UK (including the north of Scotland)

• The Northern Irish sample is not clustered however

• Within Great Britain the Primary Sampling Units (PSUs) are postal sectors

• They are stratified by the nine official regions of England, plus Scotland and Wales, by population density and by ethnic minority density

• This resulted in the systematic selection of 2,640 postal sectors with probability proportional to size (number of addresses)
• Within each sampled sector, 18 addresses were selected systematically, resulting in an equal probability sample of 47,520 addresses in Great Britain.

• In Northern Ireland 2,400 addresses were selected systematically from the Land and Property Services Agency list of domestic properties.
Response Rates

• General population survey 57.2% (60% target)

• Ethnic minority boost 56.9% (55% target)

• These rates are common for omnibus surveys in the UK (Buck and McFall, 2012)

• Burton et al. (2011) provide in-depth information on UKHLS response rates
Types of UKHLS Respondents

- Original Sample Members (OSM)
  *Present in households at the start of the study*
  *They are always followed*

- Temporary Sample Members (TSM)
  *Join a households with an OSM*
  *Only surveyed if living with an OSM*
UKHLS Following Rules (Example)

• When an adult female OSM moves in with her male partner she will also be interviewed and will become a TSM

• A TSM is not followed when they no longer live in a household with an OSM

• If a male TSM fathers a child with an OSM they become a Permanent Sample Member (PSM) and remain potentially eligible for interview for the life of survey
UKHLS Following Rules

The rules that guide following UKHLS respondents are designed to mimic the demographic processes by which the population is reproduced.

This includes births, deaths, partnership formations and dissolutions and emigration.

The study’s ‘following rules’ provide an organic selection method which over time represent the evolving pattern of families and households in the UK.
Part 7

Working with Complex Surveys

Using Software
You can’t analyse complex surveys without
Don’t try to use Excel
A beautiful solution.

No matter what your field, put Stata to work for you.

What can Stata do for you?
IBM SPSS Software

Predict with confidence what will happen next so you can make smarter decisions for your organization

Featured products

Free SPSS trials

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and Mac OS. To download R, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- R version 3.3.2 (Another Canary) has been released on Monday 2017-03-06.
- useR! 2017 (July 4–7 in Brisbane) has opened registration and more at http://useR2017.brisbane/
- Tomas Kolka has joined the R core team.
- The R Foundation welcomes five new ordinary members: Jennifer Bryan, Dianne Cook, Julie Josse, Tomas Kolka, and Raita Sibbing.

The R Journal Volume 11 is available.

- The useR! 2017 conference will take place in Brisbane, July 4-7, 2017
- R version 3.2.5 (Very, Very Secure Diet) has been released on 2016-04-14. This is a rebadging of the quickfix release 3.2.4 revised.
- Notice XQuartz users (Mac OS X) A security issue has been detected with the Sparkle update mechanism used by XQuartz. Avoid updating your rserver channels.
- The R logo is available for download in high-resolution PNG or SVG formats.
- useR! 2016, June taken place at Stanford University, CA, USA, June 27 - June 30, 2016
- The R Journal Volume 7/2 is available.
- R version 3.2.3 (Wooden Christmas Trees) has been released on 2015-12-10
- R version 3.1.2 (Smooth Sidewalk) has been released on 2015-03-06.
Survey data analysts are free to use whatever software they prefer…
Why Stata

• A wide (and growing) user community

• Excellent documentation and help files

• A helpful forum (www.statalist.org)

• Many good and accessible training resources
  • Stata http://www.stata.com/training/
  • UCLA http://stats.idre.ucla.edu/stata/
  • Timberlake http://www.timberlake.co.uk/training
  • Princeton http://data.princeton.edu/stata/
  • My old site www.longitudinal.stir.ac.uk
Stata Offers a Unified Environment

• Data Enabling (i.e. preparing complex survey data for analyses)

• Exploratory Data Analysis

• Comprehensive data analysis
  • Models in the regression family (glm)
  • Selection models
  • Multilevel models
  • Event history models
  • Longitudinal models (e.g. panel models)
Stata Offers a Unified Environment

• APPROPRIATE analysis of surveys with complex designs and selection strategies!

• The production of presentation and publication ready outputs (i.e. output suitable for reports and journal articles)
A leading US social scientist and survey data analyst Donald J. Treiman makes the following comments on Stata

‘for many years, SPSS was the package of choice among sociologists probably because it was written by and for sociologists…Although it is still widely used by social scientists in Europe and Asia it has lost its market share in leading U.S. research universities… [Stata] has gotten better and better over time, so that by now it can happily be used as a general-purpose package. Stata is powerful and fast which makes it viable to carry out analysis on a PC… Overall, Stata is a very good choice for our kind of work’ (Treiman, 2009 p.66).
• The UK Data Service provides most large-scale social surveys in SPSS and Stata format

• SPSS is far more restricted in the range of statistical models that it can estimate

• SPSS currently has few options for estimating models that are suited to longitudinal data

• Stata is able to offer more comprehensive facilities to analyse survey datasets with complex designs and selection strategies

• This is a clear benefit for social scientists working with contemporary datasets such as the UK Household Longitudinal Study (Understanding Society) and the UK Millennium Cohort Study
Part 8
Working with Complex Surveys
The Workflow
The Workflow: A Practical Guide to Producing Accurate, Efficient, Transparent and Reproducible Social Survey Data Analysis

Vernon Gayle, Paul Lambert

http://eprints.ncrm.ac.uk/4000/
J. Scott Long has posted a really good pdf version of a talk on the workflow

http://www.ihrp.uic.edu/files/Workflow%20Slides%20JSLong%20110410.pdf
The best habit that you can get into is to get into good habits!
Estimating the time taken to enable complex survey data for analysis
Part 9

Working with Complex Surveys

Enabling Data
About the study

Understanding Society is an innovative world-leading study about 21st century UK life and how it is changing. It captures important information about people’s social and economic circumstances, attitudes, behaviours and health. The study is longitudinal in its design and of high quality.

Largest household study of its kind
We interview the same people in the same households each year to build a longitudinal picture of how their lives are changing over time.

Represents the diversity of the UK
Our participants live in Scotland, Wales, Northern Ireland and England and all of Britain’s ethnic and immigrant groups are fully represented in the study.

Issues that affect all our lives
Our questionnaires cover a wide variety of themes such as family, education, finance, employment, health and wellbeing. Read about the carefully-chosen questions we ask.
<table>
<thead>
<tr>
<th>Datafile</th>
<th>Description</th>
<th>Waves</th>
</tr>
</thead>
<tbody>
<tr>
<td>adopt</td>
<td>Adopted and step-child information</td>
<td>1, 6</td>
</tr>
<tr>
<td>callrec</td>
<td>Call records</td>
<td>1, 2, 3, 4, 5, 6</td>
</tr>
<tr>
<td>child</td>
<td>Childcare, child development, consents and school information for each child</td>
<td>1, 2, 3, 4, 5, 6</td>
</tr>
<tr>
<td>chmain</td>
<td>Details of child maintenance</td>
<td>3, 5</td>
</tr>
<tr>
<td>cohab</td>
<td>Information about previous spells of cohabitation</td>
<td>1, 6</td>
</tr>
<tr>
<td>egoalt</td>
<td>Kin and other relationships between pairs of individuals</td>
<td>1, 2, 3, 4, 5, 6</td>
</tr>
<tr>
<td>empstat</td>
<td>Employment history (half of year 1)</td>
<td>1, 6</td>
</tr>
<tr>
<td>hthexp</td>
<td>Household data from respondent households</td>
<td>1, 2, 3, 4, 5, 6</td>
</tr>
<tr>
<td>hthsamp</td>
<td>Sample and Household level data for issued households</td>
<td>1, 2, 3, 4, 5, 6</td>
</tr>
<tr>
<td>income</td>
<td>Income and payment information</td>
<td>1, 2, 3, 4, 5, 6</td>
</tr>
<tr>
<td>install</td>
<td>Enumeration information</td>
<td>1, 2, 3, 4, 5, 6</td>
</tr>
<tr>
<td>indexp</td>
<td>Data from the individual interview</td>
<td>1, 2, 3, 4, 5, 6</td>
</tr>
<tr>
<td>indsamp</td>
<td>Contains individual-level data for issued households</td>
<td>1, 2, 3, 4, 5, 6</td>
</tr>
<tr>
<td>issue</td>
<td>Information about households issued to interviewers</td>
<td>1, 2, 3, 4, 5</td>
</tr>
<tr>
<td>marriage</td>
<td>Information about previous marriages</td>
<td>1, 6</td>
</tr>
<tr>
<td>natchild</td>
<td>Information about natural children</td>
<td>1, 6</td>
</tr>
<tr>
<td>newborn</td>
<td>Information about each newborn child</td>
<td>2, 3, 4, 5, 6</td>
</tr>
<tr>
<td>partstyle</td>
<td>Details of parenting styles</td>
<td>3, 4, 5, 6</td>
</tr>
<tr>
<td>youth</td>
<td>Youth self completion questionnaire data</td>
<td>1, 2, 3, 4, 5, 6</td>
</tr>
</tbody>
</table>
Types of File (main files)

• Data from the individual interview
  Individual response files  a\_indresp  b\_indresp

• Household data from respondent households
  Household response files  a\_hhresp  b\_hhresp
Types of File (other files)

- Youth self completion questionnaire data
  Ages 10-15 a_youth b_youth c_youth

- Income and payment information a_income
Useful Files

• Stable characteristics of individuals xwavedat

• Individual and household identifiers across all waves xwaveid
Linking Up People and Households
a_hidp 68001363

Household
(resident)

pidp 68001366
pidp 68001366

pidp 68001360
a_hidp 68001363

pidp 68001366
a_sppid 68001360

Household (resident)

Household (resident)
a_hidp 68001363

pidp 68001366
a_sppid 68001360

b_hidp 68001365

pidp 68001366
b_sppid -8

c_hidp 68001367

pidp 68001366
c_sppid 6800159
a_hidp  Household identifier
a_pno   Person number in household grid
pidp    Cross-wave person identifier
fpid    Cross-wave person identifier of natural father
mpid    Cross-wave person identifier of natural mother
# A Genuine Example

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>a_hidp</td>
<td>68251603</td>
<td>Household identifier</td>
</tr>
<tr>
<td>a_pno</td>
<td>4</td>
<td>Person number in household grid</td>
</tr>
<tr>
<td>pidp</td>
<td>68251619</td>
<td>Cross-wave person identifier</td>
</tr>
<tr>
<td>fpid</td>
<td>68251607</td>
<td>Cross-wave person identifier of natural father</td>
</tr>
<tr>
<td>mpid</td>
<td>68251611</td>
<td>Cross-wave person identifier of natural mother</td>
</tr>
</tbody>
</table>
Variables and Measures
Time varying variables…
a_marstat  legal marital status

Datafile: a_indall

Related to 1 Question

hhgrid_w1.marstat
Legal marital status

Text  What is NAME's legal marital status...

Universe  if ( DVAge >= 16 ) (Age 16 or over)
<table>
<thead>
<tr>
<th>pidp</th>
<th>a_marstat</th>
<th>b_marstat</th>
<th>c_marstat</th>
<th>d_marstat</th>
<th>e_marstat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 68006807</td>
<td>divorced</td>
<td>divorced</td>
<td>divorced</td>
<td>divorced</td>
<td>divorced</td>
</tr>
<tr>
<td>2 68153691</td>
<td>married</td>
<td>separ...</td>
<td>separ...</td>
<td>divorced</td>
<td>divorced</td>
</tr>
<tr>
<td>3 68174767</td>
<td>singl...</td>
<td>singl...</td>
<td>married</td>
<td>married</td>
<td>separ...</td>
</tr>
<tr>
<td>4 68291727</td>
<td>married</td>
<td>married</td>
<td>married</td>
<td>married</td>
<td>married</td>
</tr>
<tr>
<td>5 68293095</td>
<td>singl...</td>
<td>singl...</td>
<td>singl...</td>
<td>singl...</td>
<td>singl...</td>
</tr>
</tbody>
</table>

**Legal Marital Status**
Time constant variables…
```
tab a_ukborn, missing

<table>
<thead>
<tr>
<th>born in uk</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing</td>
<td>6</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>refused</td>
<td>2</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>don't know</td>
<td>8</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>yes, england</td>
<td>33,480</td>
<td>65.65</td>
<td>65.69</td>
</tr>
<tr>
<td>yes, scotland</td>
<td>3,567</td>
<td>6.99</td>
<td>72.68</td>
</tr>
<tr>
<td>yes, wales</td>
<td>2,154</td>
<td>4.22</td>
<td>76.91</td>
</tr>
<tr>
<td>yes, northern ireland</td>
<td>2,033</td>
<td>3.99</td>
<td>80.89</td>
</tr>
<tr>
<td>not born in the uk</td>
<td>9,744</td>
<td>19.11</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total | 50,994 | 100.00 |
```
tab a_ukborn, missing

<table>
<thead>
<tr>
<th>born in uk</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing</td>
<td>6</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>refused</td>
<td>2</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>don't know</td>
<td>8</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>yes, england</td>
<td>33,480</td>
<td>65.65</td>
<td>65.69</td>
</tr>
<tr>
<td>yes, scotland</td>
<td>3,567</td>
<td>6.99</td>
<td>72.68</td>
</tr>
<tr>
<td>yes, wales</td>
<td>2,154</td>
<td>4.22</td>
<td>76.91</td>
</tr>
<tr>
<td>yes, northern ireland</td>
<td>2,033</td>
<td>3.99</td>
<td>80.89</td>
</tr>
<tr>
<td>not born in the uk</td>
<td>9,744</td>
<td>19.11</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>50,994</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

---

tab b_ukborn, missing

<table>
<thead>
<tr>
<th>born in uk</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>inapplicable</td>
<td>46,811</td>
<td>69.66</td>
<td>69.66</td>
</tr>
<tr>
<td>don't know</td>
<td>8</td>
<td>0.01</td>
<td>69.67</td>
</tr>
<tr>
<td>yes, england</td>
<td>4,834</td>
<td>7.19</td>
<td>76.86</td>
</tr>
<tr>
<td>yes, scotland</td>
<td>599</td>
<td>0.89</td>
<td>77.75</td>
</tr>
<tr>
<td>yes, wales</td>
<td>402</td>
<td>0.60</td>
<td>78.35</td>
</tr>
<tr>
<td>yes, northern ireland</td>
<td>355</td>
<td>0.53</td>
<td>78.88</td>
</tr>
<tr>
<td>not born in the uk</td>
<td>1,587</td>
<td>2.36</td>
<td>81.24</td>
</tr>
<tr>
<td>.</td>
<td>12,606</td>
<td>18.76</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>67,203</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>
```
. tab b_ukborn b_hhorig

<table>
<thead>
<tr>
<th>born in uk</th>
<th>ukhls gb</th>
<th>ukhls ni</th>
<th>bhps gb 1</th>
<th>bhps sco</th>
<th>bhps wal</th>
<th>bhps ni 2</th>
<th>ukhls emb</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>inapplicable</td>
<td>30,018</td>
<td>1,655</td>
<td>5,961</td>
<td>1,436</td>
<td>1,601</td>
<td>1,908</td>
<td>4,232</td>
<td>46,811</td>
</tr>
<tr>
<td>don't know</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>yes, england</td>
<td>3,765</td>
<td>10</td>
<td>518</td>
<td>9</td>
<td>33</td>
<td>6</td>
<td>493</td>
<td>4,834</td>
</tr>
<tr>
<td>yes, scotland</td>
<td>407</td>
<td>2</td>
<td>52</td>
<td>125</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>599</td>
</tr>
<tr>
<td>yes, wales</td>
<td>222</td>
<td>0</td>
<td>29</td>
<td>0</td>
<td>150</td>
<td>0</td>
<td>1</td>
<td>402</td>
</tr>
<tr>
<td>yes, northern ireland</td>
<td>24</td>
<td>174</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>148</td>
<td>2</td>
<td>355</td>
</tr>
<tr>
<td>not born in the uk</td>
<td>664</td>
<td>17</td>
<td>33</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>862</td>
<td>1,587</td>
</tr>
</tbody>
</table>

Total | 35,105  | 1,858    | 6,598     | 1,576    | 1,795    | 2,067     | 5,598    | 54,597 |
```
Choosing between variables (and measures)…
Understanding Society Income Measures

- Wages
- Self employment
- Second job earning
- Interest and dividends
- Pensions
- Benefits
- Educational grants
- Maintenance (alimony)
- Payments from family members
- Rent and lodgers
- Other sources
Marital Status Measures

**a_indresp**
Data from the individual interview

**Variables (5)**

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>a_pmarstat</td>
<td>marital status of respondent</td>
</tr>
<tr>
<td>a_mismatchk</td>
<td>marital status check</td>
</tr>
<tr>
<td>a_mistat</td>
<td>present legal marital status</td>
</tr>
<tr>
<td>a_marstat</td>
<td>legal marital status</td>
</tr>
<tr>
<td>a_mastat_dv</td>
<td>De facto marital status</td>
</tr>
</tbody>
</table>
a_marstat legal marital status

Datafile: a_indresp

Related to 1 Question

hhgrid_w1.marstat
Legal marital status

<table>
<thead>
<tr>
<th>Text</th>
<th>What is NAME's legal marital status...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Universe</td>
<td>if ( DVAge &gt;= 16 ) (Age 16 or over)</td>
</tr>
</tbody>
</table>

Associated variables

- a_marstat (a_indresp) legal marital status
- a_marstat (a_indall) legal marital status
Data file: a_indresp
Related to 1 Question

**hhgrid_w1.marstat**
Legal marital status

**Text**
What is NAME’s legal marital status...

**Universe**
if ( DVAge >= 16 ) (Age 16 or over)

**Associated variables**
- a_marstat (a_indresp) legal marital status
- a_marstat (a_indall) legal marital status

**Frequencies**

<table>
<thead>
<tr>
<th>Value label</th>
<th>Value</th>
<th>Absolute frequency</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>refused</td>
<td>-2</td>
<td>3</td>
<td>0.01%</td>
</tr>
<tr>
<td>don't know</td>
<td>-1</td>
<td>15</td>
<td>0.03%</td>
</tr>
<tr>
<td>single &amp; never married or in legally recog’d civil p’ship</td>
<td>1</td>
<td>16225</td>
<td>31.82%</td>
</tr>
<tr>
<td>married</td>
<td>2</td>
<td>25851</td>
<td>50.69%</td>
</tr>
<tr>
<td>in a registered same-sex civil partnership</td>
<td>3</td>
<td>90</td>
<td>0.18%</td>
</tr>
<tr>
<td>separated but legally married</td>
<td>4</td>
<td>1324</td>
<td>2.60%</td>
</tr>
<tr>
<td>divorced</td>
<td>5</td>
<td>4332</td>
<td>8.50%</td>
</tr>
<tr>
<td>widowed</td>
<td>6</td>
<td>3137</td>
<td>6.15%</td>
</tr>
<tr>
<td>separated from civil partner</td>
<td>7</td>
<td>11</td>
<td>0.02%</td>
</tr>
<tr>
<td>an ex-civil partner,civil p’ship legally dissolved</td>
<td>8</td>
<td>1</td>
<td>0.00%</td>
</tr>
<tr>
<td>surviving civil partner (partner died)</td>
<td>9</td>
<td>4</td>
<td>0.01%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>50994</td>
<td>100.00%</td>
</tr>
</tbody>
</table>
Datafile: `a_indresp`

**Derived Variable note**

De facto marital status. In contrast to _A_MARSTAT_ this variable adds two further categories for children aged under 16 and respondents who are presently living in a couple.

**Origin**

Uses _A_MASTAT_DV_ on data file _A_INDALL_

### Frequencies

<table>
<thead>
<tr>
<th>Value label</th>
<th>Value</th>
<th>Absolute frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>refused</td>
<td>-2</td>
<td>3</td>
</tr>
<tr>
<td>don’t know</td>
<td>-1</td>
<td>15</td>
</tr>
<tr>
<td>single &amp; never married or in legally recog’d civil p’ship</td>
<td>1</td>
<td>16226</td>
</tr>
<tr>
<td>married</td>
<td>2</td>
<td>25851</td>
</tr>
<tr>
<td>in a registered same-sex civil partnership</td>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td>separated but legally married</td>
<td>4</td>
<td>1324</td>
</tr>
<tr>
<td>divorced</td>
<td>5</td>
<td>4332</td>
</tr>
<tr>
<td>widowed</td>
<td>6</td>
<td>3137</td>
</tr>
<tr>
<td>separated from civil partner</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>an ex-civil partner.civil p’ship legally dissolved</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>surviving civil partner (partner died)</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>50994</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value label</th>
<th>Value</th>
<th>Absolute frequency</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>refusal</td>
<td>-2</td>
<td>3</td>
<td>0.001</td>
</tr>
<tr>
<td>Don’t know</td>
<td>-1</td>
<td>15</td>
<td>0.033</td>
</tr>
<tr>
<td>Single and never married/in civil partnership</td>
<td>1</td>
<td>12009</td>
<td>23.55%</td>
</tr>
<tr>
<td>Married</td>
<td>2</td>
<td>25847</td>
<td>50.69%</td>
</tr>
<tr>
<td>In a registered same-sex civil partnership</td>
<td>3</td>
<td>90</td>
<td>0.18%</td>
</tr>
<tr>
<td>Separated but legally married</td>
<td>4</td>
<td>1153</td>
<td>2.26%</td>
</tr>
<tr>
<td>Divorced</td>
<td>5</td>
<td>3155</td>
<td>6.19%</td>
</tr>
<tr>
<td>Widowed</td>
<td>6</td>
<td>3003</td>
<td>5.89%</td>
</tr>
<tr>
<td>Separated from civil partner</td>
<td>7</td>
<td>11</td>
<td>0.22%</td>
</tr>
<tr>
<td>A former civil partner</td>
<td>8</td>
<td>1</td>
<td>0.00%</td>
</tr>
<tr>
<td>A surviving civil partner</td>
<td>9</td>
<td>4</td>
<td>0.01%</td>
</tr>
<tr>
<td><strong>Living as couple</strong></td>
<td>10</td>
<td>5704</td>
<td>11.19%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>50994</td>
<td></td>
<td>100.00%</td>
</tr>
</tbody>
</table>
Choosing between variables (and measures)...

- Think hard about which measure(s) are best suited to your research question
- Is there a good theoretical justification for your choice of measure?
- Is the measure commonly used?
- Is using the measure practicable?
- How does the measure compare with other measures?

- UNDERTAKE (and publish) a SENSITIVITY ANALYSIS!
Part 10

Survey Weights
Survey Weight /ˈsərvəri wɛйт/ noun confusing
Adapts a sample to better represent a population of interest
Adjusting a sample to better represent a population of interest

A survey with a target population with 50% men and 50%

There are 400 respondents

75% are male (n=300)
25% are female (n=100)
Adjusting a sample to better represent a population of interest

A survey with target population with 50% men and 50%

There are 400 respondents

75% are male (n=300)
25% are female (n=100)

Could we re-scale the sample to better represent the population?
A simple example of weighting data

Adjusting a sample to better represent a (target) population of interest
# Adjusting a sample to better represent a population of interest

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Population</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Target number of interviews</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>
### Adjusting a sample to better represent a population of interest

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Population</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Target number of interviews</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Achieved number of survey interviews</td>
<td>300</td>
<td>100</td>
</tr>
</tbody>
</table>
Adjusting a sample to better represent a population of interest

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Population</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Target number of interviews</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Achieved number of survey interviews</td>
<td>300</td>
<td>100</td>
</tr>
<tr>
<td>Ratio of achieved to target</td>
<td>300/200 = 1.5</td>
<td>100/200 = 0.5</td>
</tr>
</tbody>
</table>
### Adjusting a sample to better represent a population of interest

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Population</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Target number of interviews</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Achieved number of survey interviews</td>
<td>300</td>
<td>100</td>
</tr>
<tr>
<td>Ratio of achieved to target interviews</td>
<td>$\frac{300}{200} = 1.5$</td>
<td>$\frac{100}{200} = 0.5$</td>
</tr>
<tr>
<td>Weighting factor</td>
<td>$\frac{1}{1.5} = 0.67$</td>
<td>$\frac{1}{0.5} = 2$</td>
</tr>
</tbody>
</table>
A National Survey of Pupils in England and Wales in 1990

<table>
<thead>
<tr>
<th>Population percentage</th>
<th>Less than 5 GCSEs</th>
<th>5+ GCSEs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>65</td>
<td>35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No observed in survey</th>
<th>7,800</th>
<th>6,469</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted percentage</td>
<td>55</td>
<td>45</td>
</tr>
</tbody>
</table>
## A National Survey of Pupils in England and Wales in 1990

<table>
<thead>
<tr>
<th></th>
<th>Less than 5 GCSEs</th>
<th>5+ GCSEs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population percentage</strong></td>
<td>65</td>
<td>35</td>
</tr>
<tr>
<td><strong>No observed in survey</strong></td>
<td>7,800</td>
<td>6,469</td>
</tr>
<tr>
<td><strong>Unweighted percentage</strong></td>
<td>55</td>
<td>45</td>
</tr>
<tr>
<td><strong>Ratio of achieved to target</strong></td>
<td>55/65</td>
<td>45/35</td>
</tr>
<tr>
<td><strong>Survey weight</strong></td>
<td>(1 / (55/65) = 1.18)</td>
<td>(1 / (45/35) = .78)</td>
</tr>
<tr>
<td><strong>Weighted percentage</strong></td>
<td>55 \times 1.18 = 65</td>
<td>45 \times .78 = 35</td>
</tr>
</tbody>
</table>

Youth Cohort Study of England and Wales Cohort 5; GCSEs at grades A-C
The statistical literature on survey design, sampling and weighting is dense and the terminology and concepts that are used are often confusing for social science researchers.

The web based resource Practical Exemplars and Survey Analysis (PEAS) http://www.restore.ac.uk/PEAS/about.php provides a useful and accessible introduction (although it is no longer being up-dated)
Concluding Remarks on Weights…

Gelman (2007) makes the bold assertion that ‘survey weighting is a mess’, because it is not always clear how to use weights in estimating anything other than simple descriptive statistics and because software packages do not always have procedures available for complex adjustments to be made.
Joshua Angrist and Jörn-Steffen Pischke assert that ‘few things are as confusing to applied researchers as the role of sample weights. Even now, 20 years post-Ph.D.’

(Angrist and Pischke, 2008 p.66)
Part 11

Analysing Complex Surveys using Stata
Complex Survey Analysis in Stata

svy

• The svy prefix is designed for use with complex survey data

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

• Many routine analyses can be undertaken in Stata once the characteristics of the data have been set, and the results will be suitably adjusted to represent the complexity of the survey’s design
## UKHLS Example

<table>
<thead>
<tr>
<th>Total monthly personal income gross</th>
<th>Mean (£)</th>
<th>Standard Error</th>
<th>95% Confidence Interval Lower</th>
<th>95% Confidence Interval Upper</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Simple Random Sample</td>
<td>1481</td>
<td>7</td>
<td>1467</td>
<td>1495</td>
<td>none</td>
</tr>
<tr>
<td>UKHLS Design Only</td>
<td>1526</td>
<td>8</td>
<td>1510</td>
<td>1542</td>
<td>a_psnenus_xd</td>
</tr>
<tr>
<td>UKHLS Design Only + Non-Response</td>
<td>1541</td>
<td>9</td>
<td>1523</td>
<td>1558</td>
<td>a_indpxus_xw</td>
</tr>
</tbody>
</table>
## UKHLS Example

<table>
<thead>
<tr>
<th>Total monthly personal income gross</th>
<th>Mean (£)</th>
<th>Standard Error</th>
<th>95% Confidence Interval</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Simple Random Sample</td>
<td>1481</td>
<td>7</td>
<td>1467</td>
<td>1495</td>
</tr>
<tr>
<td>UKHLS Design Only</td>
<td>1526</td>
<td>8</td>
<td>1510</td>
<td>1542</td>
</tr>
<tr>
<td>UKHLS Design Only + Non-Response</td>
<td>1541</td>
<td>9</td>
<td>1523</td>
<td>1558</td>
</tr>
</tbody>
</table>
UKHLS Example

<table>
<thead>
<tr>
<th>Total monthly personal income gross</th>
<th>Mean (£)</th>
<th>Standard Error</th>
<th>95% Confidence Interval Lower</th>
<th>95% Confidence Interval Upper</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Simple Random Sample</td>
<td>1481</td>
<td>7</td>
<td>1467</td>
<td>1495</td>
<td>none</td>
</tr>
<tr>
<td>UKHLS Design Only</td>
<td>1526</td>
<td>8</td>
<td>1510</td>
<td>1542</td>
<td>a_plenus_xd</td>
</tr>
<tr>
<td>UKHLS Design Only + Non-Response</td>
<td>1541</td>
<td>9</td>
<td>1523</td>
<td>1558</td>
<td>a_indpxus_xw</td>
</tr>
</tbody>
</table>
A Warning Against
A Naïve Approach to Inference

‘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)
Part 12

Remarks and Conclusions
Human beings, who are almost unique in having the ability to learn from the experience of others, are also remarkable for their apparent disinclination to do so.

-Douglas Adams
Complex Social Surveys

You might never conduct a complex survey…

You might never analyse a complex survey dataset…

But you may need to read results from the analysis of complex social survey and therefore wish to understand more
Things you might consider when reading analyses using complex social surveys....

• Was the most appropriate survey used?

• Did the researchers think about the survey design and selection strategy?

• Were an optimal set of cases used?
Things you might consider when reading analyses using complex social surveys….

- Could extra information have been incorporated (e.g. from other household members)?

- Did the researchers think about the effects of missing data (i.e. units and items)?

- Were the most suitable $Y$ and $X$ vars used?

- Is there sufficient information about the analytical process?
Things you might consider when commissioning an analysis….

• Is the data analyst suitably trained?

• How well do they know the dataset?
  • How many dates have they been on!

• Is there sufficient time to get to know the dataset?

• Is there sufficient (funded?) time for data enabling?
Things you might consider when commissioning an analysis....

• Can you and your organisation publically defend the work?

• Will the data analysis process be transparent?

• Will the work be archived and rendered reproducible?
Things you might think about before analysing complex social surveys...

• Learn how to use an suitable software package (e.g. Stata)

• Come to graduate school at either Edinburgh or Warwick

• Attend a course run by AQMeN (http://aqmen.ac.uk/)

• Consult the resources page on our website
  http://www2.warwick.ac.uk/fac/soc/sociology/staff/connelly/cognitiveinequalities
WHAT IS QUANTITATIVE LONGITUDINAL DATA ANALYSIS?

VERNON GATTLE AND PAUL LAYBERT

RESEARCH METHODS SERIES
The Future

The analysis of surveys with complex designs is an emerging area

There will undoubtedly be new theoretical insights which, alongside advances in software functionality and improvements in computational power, will influence data analysis practices
Angrist and Pischke (2008) playfully remarked that if applied research was easy then theorists would do it.

They also reassure readers that applied research is not as hard as the dense pages of *Econometrica* might lead us to believe.

Most of all *don’t panic*.
Resources on Stata

Stata home page  http://www.stata.com/
Stata Bookstore  http://www.stata.com/bookstore/books-on-stata/
Stata List  http://www.statalist.org/
Stata on Twitter  http://twitter.com/#!/stata
Stata Journal  http://www.stata-journal.com/

UCLA Academic Technology Services  http://www.ats.ucla.edu/stat/stata/

Princeton Stata resources  http://data.princeton.edu/stata/

The website of the ESRC 'Longitudinal Data Analysis for Social Science Researchers' project much of my earlier Stata training resources are available on this site  www.longitudinal.stir.ac.uk
Introductions to Stata

(An excellent introduction to Stata)

(A very clear new text on using Stata)

(Students often like this book as an introduction)
Other Useful Texts

(The authoritative text on the workflow and organizing data analysis)

(An excellent book on statistical modelling)

(A great book on analysing panel data)

(This is a comprehensive guide to statistical modelling).


More Advanced Texts


This work was supported by the Economic and Social Research Council [Grant Number ES/N011783/1]. More details of the project are available here: http://tinyurl.com/h3nj59b