Introduction to Complex Sample Design in UK Government Surveys

ESDS Government

Author: Anthony Rafferty

Updated by Pierre Walthery

Version: 1.3

Date: September 2011

University of Essex   MANCHESTER 1824 The University of Manchester   JISC   E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

CONTENTS

1. Introduction

Overview

Standard commands in statistical software typically treat data as simple random samples. The vast majority of ESDS Government supported surveys however employ complex sample design features such as clustering or stratification. Software commands designed for simple random samples do not take into account the statistical implications of complex sample design. An important consequence of this is that the magnitude of standard errors may be underestimated (or in some cases, overestimated). Techniques are available in statistical packages such as Stata or SPSS that incorporate complex design features into your analysis to take into account such effects.

This introductory guide provides an overview of the survey design features of ESDS Government datasets. It also provides information on analysing complex samples using Stata. Focussing on design-based approaches, practical examples are given using the Health Survey for England (HSE).

The guide is organised as follows. Chapter 2 outlines key concepts in complex survey design such as clustering, primary, and secondary sampling units, stratification, and post-stratification. Chapter 3 outlines the design features of ESDS Government supported datasets. Chapter 4 summarises the differences between design and model-based approaches to incorporating complex survey design features into your analysis. Using examples from the Health Survey for England (HSE), Chapter 5 provides an introduction to commands in Stata, focussing on design-based methods using the **svy** suite of commands.

Information about ESDS Government supported datasets is provided for the latest data available at the time of writing. Therefore readers are advised to check the user documentation for the specific year they are interested in for a given survey as their definitive source of information.

About ESDS Government and supported surveys

The Centre for Census and Survey Research provides: user support for ESDS Government, user meetings on specific surveys, training courses on key topics of interest, specific statistical packages and on methods of statistical analysis, topic-related online course materials, and a range of teaching datasets. The service aims to promote and facilitate effective use of ESDS government datasets in research, learning and teaching across a range of disciplines.
The surveys that ESDS Government supports are:

- Annual Population Survey
- British Crime Survey
- British Social Attitudes
- Continuous Household Survey (Northern Ireland)
- Expenditure and Food Survey
- English Housing Survey (EHS), formerly Survey of English Housing
- Living Costs and Food Survey (formerly the Family Expenditure Survey)

- Family Resources Survey
- General Lifestyle Survey (formerly the General Household Survey)
- Health Survey for England
- Households Below Average Income
- Integrated Household Survey
- Labour Force Surveys
- Life Opportunities Survey
- Living in Wales Survey
- National Food Survey
- National Survey for Wales
- National Travel Survey
- Northern Ireland Family Expenditure Survey
- Northern Ireland Labour Force Survey
- Northern Ireland Life and Times Survey (and the former Northern Ireland Social Attitudes Survey)
- ONS Opinions Survey (formerly the Omnibus Survey)
- Scottish Crime and Justice Survey (formerly the Scottish Crime and Victimisation Survey)
- Scottish Health Survey (SHeS)
- Scottish Social Attitudes
- Survey of English Housing
- Time Use Survey
- Vital Statistics
- Welsh Health Survey
- Young People's Social Attitudes (periodic offshoot of the BSA)

Accessing the data and directing your queries

New users can get information about access to the data on the following web page: **http://www.esds.ac.uk/support/newuser.asp**

Queries regarding access to, or analysis of, ESDS Government data can be directed towards the ESDS Government Helpdesk:

E-mail: govsurveys@esds.ac.uk
Tel: +44 (0)161 275 1980
Fax: 0161 275 4722
ESDS Government web site: **www.esds.ac.uk/government**

Other ESRC Resources

The ESRC Research Methods Practical exemplars on the analysis of surveys (PEAs) project at Napier University provides both theoretical background and practical exemplars on complex sample design.

2. Key Concepts in Complex Survey Design

Introduction

The intended learning outcomes of this section are to:

- Understand the basic principles involved in simple random sampling and systematic sampling;
- Understand how samples may be stratified and/or clustered;
- Know what design effects are and how these influence the standard errors of estimates;
- Know what multi-stage sampling is. Understand the differences between one-stage and two-stage cluster samples and the meaning of 'primary sampling unit' (PSU) and 'secondary sampling unit' (SSU);
- Know why post-stratification is used;
- Understand the potential pitfalls of ignoring complex sample design in your analysis. Particularly, we will focus on the potential misspecification of standard errors.

Samples are fractions of populations

In survey research, we typically have a population of interest about which we wish to make inferences based on data obtained from it. Obtaining data for every person can be costly; therefore, we use a sample of the population. In standard notation, the units (e.g. people, households etc, depending on level of analysis) in a population we are interested in are denoted by N whereas the sample units are denoted by n. A sample is a fraction of the population. This can be expressed as a ratio of the size of our sample to the size of the population:

$$\text{Sampling fraction } (f) = n/N.$$

For example if a population N= 50,000,000 and we take a sample n =5000, our sampling fraction (f) is 5000/50000000= 0.0001.

Random samples: with or without replacement

Survey research typically uses some form of random sampling or process for approximating a random sample. The reason we select random samples is to try to ensure that our sample is representative of the population of interest and therefore not biased because of our selection procedure. In a simple **random sample with replacement** (SRSWR), one unit is randomly selected from the population, then put back into the population and then a second unit is drawn. This procedure is repeated until the desired n units are obtained, and due to replacement there may be duplicates in the sample (i.e. the same unit may be used twice).

A **simple random sample without replacement (SRS)** is often preferred to the above approach to avoid duplicates, as people do not provide any further information

about the population when they are included more than once. In SRS, every potential unit has an equal probability of being selected.

Using sampling frames to draw systematic samples from populations

Typically, Government surveys use some form of **systematic sampling** to proxy simple random samples. **Sampling frames** are used for this. These consist of lists of identifiers that (ideally) uniquely identify the elements or units in a population. The most commonly used sampling frames in ESDS Government datasets are the **Postal Address Files (PAFs)** for Great Britain, and the **Land and Property Services Agency's (LPSA)**[1] list of domestic addresses for Northern Ireland. A random starting point and fixed interval can be taken across the sampling frame to obtain a sample. Lists of pseudo-random numbers can be used to draw a simple random sample without replacement with roughly equal selection probabilities for each unit.

Rather than lists of individuals, the PAF provides a lists of postal 'delivery points' which are used to identify addresses.

To understand the PAF, consider the following postcode: **M13 9PL**

- **M13** is the 'post code district'
- M13 **9** is the 'postcode sector'. This is typically used as the level of Primary Sampling Units (PSUs) (discussed below)
- M13 9 **PL** is the 'delivery point' (used to identify addresses).

Design Effect (Deff) and Design Factor (Deft)

Standard statistical commands in software such as Stata or SPSS assume that data is drawn using simple random sampling methods. However, in most cases, ESDS Government supported datasets have complex survey designs, that incorporate **stratification** or **clustering**. This means that the 'correct' estimation of variances and standard errors requires specialised techniques to incorporate the **design effects** of complex sample design.

Complex samples are usually understood in terms of how they influence standard errors of estimates in comparison to what would occur if we had used a simple random sample (SRS) of the same size. Below, we will discuss in more detail how in comparison to a SRS, clustering can increase standard errors, whereas stratification will almost always decrease standard errors. The comparison of sampling errors under different sample designs is carried out using design effect statistics. For complex samples, this is typically carried out by drawing comparisons to a hypothetical simple random sample (SRS) of the same size.

The **design effect (Deff)** is the ratio of the design-based variance of an estimate θ (from a complex sample) to the variance of an estimate θ from a simple random sample (SRS) of the same size:

---

[1] This used to be called the Land Valuation Agency (LVA) list. However, since April, 2007 the LVA became the Land and Property Service Agency.

Design Effect = Var (θ design) / Var (θ srs)

The square root of the design effect gives the **design factor (deft)**. This puts things back into the scale of the standard errors so we can consider the effects of the design factor on the standard errors (n.b. the square root of the variance similarly gives the standard error).

In a simple random sample, the 95% **confidence interval** is defined as the mean +/- 1.96 of the standard error. For surveys with complex design features such as clustering and stratification, these figures need to be multiplied by the **design factor (Deft)** which takes into account the effect of complex design features on sampling error. For example, a *deft* of 3 indicates that the standard errors are three times as large as they would have been had the design been a simple random sample.

Interpretation of Deft

Deft = 1:   No Effect of sample design on standard error.
Deft>1:   Sample design inflates the standard error of the estimate.
Deft<1:   Sample design increases efficiency (reduces s.e.) of estimate.

Software such as SPSS and Stata have commands that allow you to automatically calculate design effects. For example, in Stata, **Svyset** commands (discussed in Chapter 4) allow you to estimate variances and standard errors adjusted for the design effects of stratification and clustering.

Stratification

The design effect of a complex survey design depends on whether, and if so how, stratification and clustering is incorporated into the design. Stratification is commonly used to reduce the standard errors of survey estimates or ensure that sample sizes for a given set of characteristics or 'strata' are of the expected size. Through stratification, we can enforce the sample to be representative based on the characteristics we use to stratify the sample.

Stratification can be 'explicit' or 'implicit'. In **explicit stratified sampling**, the population is partitioned into groups (strata) based on variables, such as regions, and a sample is selected by some design (e.g. randomly or systematically) within each strata. In this manner, prior knowledge about some characteristics of sampling units in the population is used to control the number of units sampled within each stratum[2]. For example, we know that the population is made up of different regions. Therefore, we might stratify by region to obtain samples of regions proportional to their size.

Implicit stratification refers to where the population of sampling units is sorted by some characteristic(s) and then a sample is selected from the sorted list using a fixed

---

[2] See http://www2.napier.ac.uk/depts/fhls/peas/sratheory.asp

sampling interval and random start. As we choose an interval that allows us to cover the sorted sampling frame, this allows us to draw a sample representing our different stratum.

Large-scale surveys often use a combination of explicit and implicit stratification. The sampling frame is first grouped into a number of explicit or 'major' strata, and within each of these, sorted by a continuous stratum variable or one with many classes[3]. For example, a population of adults might be sorted by geographical strata such as region, then within each region ordered by socio-economic classification (e.g. National Statistics Socio Economic Classification (NS-SEC[4])). Suppose every *nth* person is then selected from across the sampling frame by taking a random start between 1 and n and then every nth person after that, working down the list. Such a sample is described as a stratified sample with explicit stratification by region and implicit stratification by NS-SEC. Details of stratification variables used in ESDS Government supported datasets are provided in Chapter 3.

A further distinction made is between **proportionate** and **disproportionate stratification**. Recall that the sampling fraction is the size of the sample (*n*) divided by the size of the population (*N)*. If the *same* sampling fraction is used per stratum this is referred to as proportionate stratification, whereas if the sampling fraction is not the same in each stratum, this is referred to as disproportionate stratification. The latter can be used to create a larger sample of a sub-population of interest (e.g. ethnic minority or child booster samples). In a simple random sample, as minority ethnic groups have smaller population sizes, we might obtain an ethnic minority sample, which is too small for our analytical purposes, particularly if we are interested in exploring differences by ethnicity. We might instead therefore use disproportionate sampling to over-sample ethnic minority groups, such as by using a sampling fraction of 1 in every 1000 for the majority white population, and 1 in every 500 people for the minority ethnic population. ESDS Government supported datasets that include ethnic boost samples include the Health Survey for England (1994 and 2004), and some years of the British Crime Survey.

Design effects of stratification

Proportionate stratified sampling leads to an increase in survey precision (smaller standard errors), when compared to a design with no stratification. Disproportionate stratification in contrast can have varying effects, increasing or decreasing precision, depending on the level of variance for a given characteristic within the over-sampled stratum[5]. Disproportionate stratification also requires weights to give unbiased cross-strata estimates. Otherwise, if weights are not used, the over-sampled strata will have an influence on overall population estimates disproportionate to their actual population size.

---

[3] See exemplar 2 on the PEAS website for an example of this for the Scottish Household Survey: http://www2.napier.ac.uk/depts/fhls/peas/exemplar2.asp#stratification
[4] See http://www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/soc2010/soc2010-volume-3-ns-sec--rebased-on-soc2010--user-manual/index.html
[5] See: http://www2.napier.ac.uk/depts/fhls/peas/sratheory.asp#disperrors

Clustering

Clustering is often used to generate survey samples as it can be more cost effective than simple random sampling, cutting down fieldwork costs through making interviews more geographically concentrated. Rather than taking a random sample of the overall population, clustering involves first breaking down a population into a higher level characteristic or unit than the population elements, such as by geographical unit. A random sample of these units is randomly selected then population elements are drawn from the selected units.

This process can involve one or more stages. In a **one-stage cluster sample** (of individuals), a random sample of a population unit is selected, such as households, then individuals are selected from households, and all the elements in the sample of clusters are taken as the element level sample.

Sample designs with more than one stage of clustering are referred to as **multi-stage samples**. For example, in **two-stage cluster samples**, a second stage of selection is involved. Consider the following example, giving a common clustered design for a Postal Address File (PAF) sample. At the first stage, a random sample of postcode sectors is selected. Next, in the second stage, a random sample of households is selected within postcode sectors. Finally, individuals ('population elements') are selected within households. In this design, adults are clustered in households, and households are clustered in postcode sectors. Note that in the final stage, all the population units in the lowest level sampling unit are selected. The higher-level cluster is commonly referred to as the **Primary Sampling Unit (PSU)** (i.e. 'first sampling unit'), and the lower level (in the current example, households) as the **Secondary Sampling Unit (SSU)**. Of course, it is possible to have a greater number of stages in clustering resulting in different multi-stage designs, but we will focus on one stage and two-stage designs, as most ESDS Government datasets do not go beyond this number of stages.

Design Effect of clustering

Clustered samples tend to increase the standard error of survey estimates relative to a simple random sample of the same size, giving a design effect above one, because cases in the same clusters are more similar than cases in different clusters, on average.

The size of the increase of the design effect is dependant on: 1) the sample size per cluster, and 2) the homogeneity of the clusters (see Lohr, 1999, pg. 138-141). As cluster sizes increase, standard errors tend to increase. The homogeneity of the cluster is measured by the **intra-cluster correlation coefficient** (ICC, or rho). If the individuals within a cluster have more in common than individuals have in general, then ρ will be greater than zero. If, at the extreme, all individuals within a cluster are identical yet there is some between-cluster variation, then ρ will be equal to 1. As ρ increases so does the standard error.

The two observations above make intuitive sense. As individuals in clusters tend to share common characteristics then any decrease to the number of clusters increases

the risk of drawing a sample that is different to the population which in turn increases standard errors of survey estimates. Similarly, if the homogeneity of individuals clusters increases (i.e. individuals within clusters become more alike compared to individuals more generally) then there is an increased risk of drawing a sample that happens to be different to the population, and this risk is reflected in the increased standard error.

Example: Design effects in the UK Labour Force Survey

Even though some surveys are referred to as simple random or systematic random samples, the sampling of addresses means that clustering may still occur at the household level for *individual* level analysis (where people are population elements):

"In the case of the LFS sample design, there is a clustering effect. This reflects the fact that addresses are sampled, but that results are shown for individuals. For example, ethnicity is particularly clustered, since it is likely that all members of a household living at a particular address will share the same ethnicity. This results in, for example, the design factor for the Pakistani and Indian ethnic groups being 1.71, which is higher than for the other ethnic groups because of the relatively large household sizes for Indians and Pakistanis. The design factor for part-time employees on the other hand is 0.95, reflecting the fact that part-time employee status is not clustered within a household. By itself clustering would tend to increase the design effect of LFS estimates. However, the LFS sample design employs stratification. Since addresses are stratified by postcode sector there is a reduction in the standard error of estimates related to the factors used in stratification. For the standard errors of change and redundancy rates interviewer areas are used as strata." (pg. 36, LFS User Guide, vol. 1 – at the time of writing the most recent User Guide, vol 1. is available at: http://www.esds.ac.uk/doc/6782%5Cmrdoc%5Cpdf%5Clfs_user_guide_vol1_background2009.pdf).

Post-stratification

Post-stratification may be used once the data is collected to attempt to reduce any bias in the survey due to sampling error or non-response. This is a weighting method that adjusts for any differences between the survey data and the population in terms of a key population variables (such as age or sex) based on auxiliary information about the make-up of a population (e.g. census information). More information on this can be found in the ESDS guide to Weighting the Social Surveys (p11) which is available at: http://www.esds.ac.uk/government/docs/weighting.pdf

Ignoring complex sample design

There are some potential pitfalls of ignoring complex sample design. If you use standard commands in Stata or SPSS without accounting for complex survey design, your software will treat your data as a simple random sample. Many statistical procedures assume that observations are independent and identically distributed (iid). The effects of clustering, stratification, and unequal selection probabilities however may mean this is not the case. As well as incorrectly estimating standard errors, your estimates may be biased if the value of variables is related to selection probability.

The under-estimation of standard errors when we ignore clustering can mean that we find some results statistically significant at a given level (e.g. $p<0.05$) which, once design effects are taken into account, are not. Skinner, Holt, and Smith (1989: 29, Table 2.1) for example show that ignoring the Deff, which is the same as treating the Deff as 1.0, skews significance levels of obtained p- values quite substantially.

3. Sample Design of UK Government Surveys

Introduction

Although featuring complex sample designs, the majority of ESDS Government supported datasets unfortunately do not currently contain variables for identifying sampling units. Notwithstanding this (!), the aims of this section are:

- To outline the main survey design features of major ESDS supported Government datasets;
- To indicate, where available, the key sample variables (focusing on primary sampling unit (PSUs)) in the datasets.

Overview of sample design features

Figure 1 provides an overview of the main sample design features and key sample variables of ESDS Government supported datasets. Information is provided for the latest data available at the time of writing. Readers are therefore advised to check the user documentation for the specific year they are interested in for a given survey to obtain information on historical features or recent developments.

Examining the listed datasets, a number of common features are identifiable:

- The most common sampling frame for Great Britain (GB) is the Postcode Address File (PAF). In one-stage designs, addresses (delivery points) are used as primary sampling units (to select households); whereas in two-stage designs postcode sectors are used as primary sampling units (PSUs), and addresses form the secondary sampling units (SSUs). A more detail explanation of the PAF can be found in Chapter 2;
- For surveys that include Northern Ireland (NI), the NI sample is usually obtained via systematic random or stratified random sampling using the Land and Property Services Agency (LPSA) list of addresses[6]. This means that there is no comparable PSU at a higher level of geography than address for the NI sample, when compared to equivalent GB samples for a given survey. However, for individual level analysis, clustering may still occur at the household level;
- Household level clustering does not apply (either in one or multi-stage designs, or in NI or GB) in surveys where only one person per household is interviewed, or where a household level unit of analysis is opted for;
- The most common stratification variables used are geographical (e.g. Government Office Region); socio-economic (e.g. NS-SEC, proportion of people in the area in non-manual occupations; car ownership) or demographic (e.g. proportion of people who are pensioners, population density).

---

[6] This used to be called the Land Valuation Agency (LVA) list. However, since April, 2007 the LVA becamce the Land and Property Service Agency.

Figure 1. Sample design features of ESDS Government supported datasets

Data set	Sample Design	Clustering	PSU variable	Strata[7]
APS **April 2008- March 2009**	Stratified Random (with household clustering and weak stratification by postcode area).	Individuals clustered in households (individual level analysis).	Same as QLFS (see below). Clustering variable can be derived from data deposit	Geographic using Postcode address file (PAF). Stratification variable can be derived from data deposit
BCS **2008-09**	Multi-stage stratified random with partial clustering.	PSU is middle super output areas (MSOA) SSU: Address (from PAF) Tertiary Sampling Unit: One individual selected per household (Kish Grid).	PSU not included in deposit.	Stratification variable not in deposit. Police Force area, Population density cluster type, Crime and disorder deprivation index Population density index
BSAS **2008**	Multi-stage stratified random.	PSU: Postcode Sector, SSU: Address, Tertiary SU: One person per household selected.	Variable '**spoint**'.	Sub-region; population density; home ownership. Variable '**stratid**' (see description below).
CHS[8] (NI) **2008-09**	Stratified random.	Although stratified random sample, individuals cluster in households for individual level analysis.	Households identified by '**caseid**'.	Stratification variable not in deposit. Geography (see description below).
FES **2000-01**	Multi-stage stratified random (GB); systematic random sample (NI).	(For GB) PSU: Postcode sector.	PSU (postcode) not included in deposit.	Stratification variable not in deposit. Socio-economic status and car ownership (introduced 1996-7).

---

[7] For UK-wide surveys, information does not apply to Northern Ireland samples unless stated.
[8] Continuous Household Survey (Northern Ireland).

Data set	Sample Design	Clustering	PSU variable	Strata[1]
FRS **2007-08**	Multi-stage stratified random (GB); stratified systematic. random (NI).	Postcode sector (PAF) (GB); NI sample uses Valuation and Land Agency list.	PSU (postcode) not included in deposit.	Stratification variable not in deposit. GB: Government Office Region; NS-SEC; economically active population 16-74 yrs; male unemployment NI: District council and ward.
General Lifestyle Survey – GLF (Formerly General Household Survey) **2006**	Multi-stage stratified random	PSU: Postcode sector. SSU: Address.	PSU (postcode) not included in End User Licence dataset but available through Special License dataset (cluster)	Stratification variable not in End User Licence dataset but available through Special License dataset (major_strata) Government Office Region; car ownership; socio-economic group; pensioners (see below).
HSE **2008**	Multi-stage stratified random.	PSU: Postcode sector. SSU: Address.	PSU: '**area**' (except for 2 years: Called '**psu**' in 2006 and 2008)	Local Authority; NS-SEC. Strata identified by variable '**cluster**' (confusingly).
Integrated Household Survey - IHS **2010**	Made of several surveys, has incorporate several sample design characteristics.	See the component surveys (QLFS, GLF, LCF, EHS, LOS) for details about their respective designs.	Not available	Not available
QLFS **Jan- March 2009**	Systematic with random start. HH clustering. (weak stratification due to use of PAF, sorted by postcode, to select addresses).	PSU: Address. Individuals are clustered within households (individual analysis)	See derivations below of address (PSU) and postcode area (strata).	Weak stratification by postcode sector (derivation described see below).
Life Opportunities Survey (LOS) **2010**	Single stage stratified random	PSU: Addresses. Individuals are clustered within households (individual analysis)	Not available	Not available
LCF **2008**	Multi-stage stratified random (GB); systematic random (NI).	(For GB) PSU: Postcode sector SSU: address.	PSU (postcode) not included in deposit.	Stratification variable not in deposit. GOR; socio-economic group; car ownership.
NFS **2000**	Multi-stage stratified (GB); systematic random (NI).	(For GB) PSU: Postcode sector (PAF). SSU: address.	PSU (postcode) not included in deposit.	Stratification variable not in deposit. Government Office Region; Socio-economic group; car ownership

Data set	Sample Design	Clustering	PSU variable	Strata'
NILFS **2000**	Stratified systematic random.	None with the exception of household clustering for individual level analysis.	Not applicable (household indicator needed for individual analysis)	Stratification variable not in deposit. Geographical stratification: Sample sorted by district council and ward.
NILTS **2008**	Systematic random sample of addresses (then one person per household selected)	Not applicable.	Not applicable.	Not applicable.
National Survey for Wales (NSW)– Pilot Study **2009-2010**	Single stage stratified random (Phase 1) Two stage stratified random with selection probability proportional to size (Phase 2)	PSU are addresses. During the first phase, they were randomly selected; during the second stages, addresses were selected according to size of the Welsh Assembly Constituencies	ADDNO (address number)	DV_WAG_PART_CONST (Welsh Assembly constituencies), PHASE (phase of the survey)
NTS **2002-06**	Multi-stage stratified random.	PSU: Postcode sector. SSU: Address.	PSU not included in deposit.	Stratification variable not in deposit. Region; car ownership; population density.
ONS Opinions Survey 2007 (formely ONS Omnibus Survey)	Multi-stage stratified random sample. One person per household interviewed.	PSU: Postcode sector.	PSU not included in deposit.	Stratification variable not in deposit. Region; NS-SEC; Pensioners (people aged 65+ yrs).
Scottish Crime and Justice Survey (SCJS) **2008**	Multi-stage stratified random.	PSU: In rural areas datazones are used as PSUs. In urban areas the sample design is unclustered; delivery points were randomely selected from PAF. One person selected per household interviewed.	PSU not included in deposit.	Stratification variable not in deposit. Police force area; Criminal Justice Authority Area; urban/rural classification, Output area, datazone, local authority, intermediate geography area
Scottish Health Survey (SHeS) **2008**	Multi-stage stratified random.	PSU: Datazone. SSU: Address.	PSU and stratification variables are included in 2008 deposit ('psu' 'strata').	Health boards and 2006 Scottish Index of Multiple Deprivation

Data set	Sample Design	Clustering	PSU variable	Strata[7]
		Four year sample (2008-2011) will be unclustered)		
English Housing Survey household dataset – EHS (formerly survey of English Housing) **2007-08**	Single stage stratified random	PSU: Addresses	PSU not included in deposit.	-
TUS **2000**	Multi-stage stratified random.	PSU: Postcode sector (GB); Wards (NI). SSU: Address.	PSU not included in deposit.	Stratification variable not in deposit. Government Office Region; Socio-economic group; population density
WHS (Welsh Health Survey) **2008**	Multi-stage stratified random.	Households selected directly from PAF. Archhsn identifies households enabling household clustering to be accounted for. Stratification variable not available.	Archhsn	Stratification variable not in deposit. Unitary Authority (UA).

Annual Population Survey (APS)

Description: http://www.esds.ac.uk/government/aps/index.asp

Sample design: Multi-stage stratified random sample.

The APS comprises key variables from the *Labour Force Survey* (LFS), all its associated LFS boosts and the APS boost. Thus, the APS combines results from five different sources: the LFS (waves 1 and 5); the English *Local Labour Force Survey* (LLFS), the *Welsh Labour Force Survey* (WLFS), the *Scottish Labour Force Survey* (SLFS) and the *Annual Population Survey Boost Sample* (APS(B) - however, this ceased to exist at the end of December 2005, so APS data from January 2006 onwards will contain all the above data apart from APS(B)).

Key sample design variable(s): Same as LFS (see below)

Weighting: For later studies in the series, due to the removal of the sample APS(B)[9] cases, there is now a single weighting variable for use with all variables. This weights for non-response and to gross to population estimates. In 2009, ONS undertook a reweighting project, whereby APS and LFS data were reweighted using population estimates for 2009. As a result, reweighted editions of APS datasets were re-deposited at UKDA during 2010 and have been added to the collection.

---

[9] See http://www.esds.ac.uk/findingData/snDescription.asp?sn=6068 for a description of the history of the sample structure.

British Crime Survey (BCS)

Description: http://www.esds.ac.uk/government/bcs/

Sample design: Multi-stage stratified random sample with partial clustering.

The 2008-9 BCS followed a revised sample design. Full details of previous sample designs can be seen in previous BCS technical reports. The 2008-9 sample design, shares some common ground with previous surveys including:

- A sample of approximately 46,000 interviews per year with adults aged 16+ living in private households in England and Wales;
- A minimum of around 1,000 interviews per year within each of the 42 Police Force Areas in England and Wales;
- A sample design that provides nationally representative estimates on a quarterly and annual basis; and
- One adult in each household selected at random for interview.

The new features of the BCS design were as follows:

- Adopting a partially clustered design with different levels of clustering in different population density strata in an effort to reduce PSU-level cluster effects;
- Using ONS Middle Super Output Areas (MSOA) as the Primary Sampling Units in the strata where the sample was clustered;
- Using new stratification variables based on an analysis of BCS data from 2004-2007; and
- Allocating sample collection between quarters to ensure the sample was nationally representative on a quarterly basis but front-loading the sample within each quarter to reduce the spill over of cases which are issued in one year but are interviewed in the next.

Three different sampling strategies were pursued dependent upon strata defined according to population density of MSOA:

1. In the *most densely populated* areas of each PFA an unclustered sample of addresses would be drawn (Stratum A);
2. In areas of *medium population density* a two-stage design would be employed, first sampling Medium Layer Super Output Areas (MSOAs) as the primary sampling units and then selecting 32 addresses within each PSU (Stratum B); and
3. In areas of *low population density* a three-stage design would be employed by first sampling Medium Layer Super Output Areas (MSOAs), then selecting 2 Lower Level Super Output Areas (LSOAs) within each sampled MSOA as the primary sampling units, and finally selecting 16 addresses within each PSU (Stratum C);

The PSUs in the BCS were stratified according to 4 levels:

1. Police Force Area

2. Density cluster type
3. 'Crime and disorder' deprivation index (3 bands)
4. Population density index (3 bands)

In the rare cases where more than one dwelling unit was associated with a single address the interviewers randomly selected one dwelling unit. Once the dwelling unit was selected, one resident was randomly sampled and no substitutes were allowed.

For more information on the BCS 2008-9 sampling design with justification for the changes to the sample design in 2008-9 see section 2 of the BCS technical report (vol 1) http://www.esds.ac.uk/doc/6367%5Cmrdoc%5Cpdf%5C6367_bcs_2008-09_technical_report_vol1.pdf . The latest version of the technical guide at the time of writing can be found here.

Key sample design variable(s):   Dependant of population density. Medium layer super output area is the PSU in medium to high population density areas. A PSU variable is not included in the deposited data.

Weighting: There are three main reasons for weighting the BCS (1) to compensate for unequal selection probabilities (2) to compensate for differential response rates (3) to ensure that quarters are equally weighted for analyses that combine data from more than one quarter..

The BCS 2008-9 includes four weights. **Indivwgt** should be used for individual based analysis (attitudinal questions and estimates of personal crime rates). **Hhdwgt** should used for household based analysis (estimates of household crime rates). For incident-based analysis, the weight **weighti** should be used. For analysis confined to 16-24 year olds a weight based on 16-24 year olds from the main sample and those in the young adults boost sample should be used (**ypcwgt**).

For more information see section 7 of the 2008-09 technical report or the ESDS guide to Weighting Social Surveys: http://www.esds.ac.uk/government/docs/weighting.pdf

British Social Attitudes Survey (BSAS)

Description*: http://www.esds.ac.uk/government/bsa/

Sample Design: Multi-stage stratified random sample.

The British Social Attitudes survey is designed to yield a representative sample of adults aged 18 or over. Since 1993, the sampling frame for the survey has been the Postcode Address File (PAF). The sampling method involved a multi-stage design, with three separate stages of selection.

At the first stage, postcode sectors were selected systematically from a list of all postal sectors in Great Britain. Before selection, any sectors with fewer than 500 addresses were identified and grouped together with an adjacent sector; in Scotland all sectors north of the Caledonian Canal were excluded (because of the prohibitive costs of interviewing there). Sectors were selected, with probability proportional to the number of addresses in each sector from strata based upon:

- 37 sub-regions;
- Population density with variable banding used, in order to create three equal-sized strata per sub-region; and
- Ranking by percentage of homes that were owner-occupied.

In the second stage, addresses were selected by starting from a random point on the list of addresses for each sector, and choosing each address at a fixed interval. The Multiple-Occupancy Indicator (MOI) available through PAF was used when selecting addresses in Scotland. The MOI shows the number of accommodation spaces sharing one address. Thus, if the MOI indicates more than one accommodation space at a given address, the chances of the given address being selected from the list of addresses would increase so that it matched the total number of accommodation spaces. The MOI is largely irrelevant in England and Wales, as separate dwelling units generally appear as separate entries on PAF. In Scotland, tenements with many flats tend to appear as one entry on PAF. However, even in Scotland, the vast majority of MOIs had a value of one. The remainder were incorporated into the weighting procedures (described below).

In the third stage, interviewers called at each address selected from PAF and listed all those eligible for inclusion in the British Social Attitudes sample - that is, all persons currently aged 18 or over and resident at the selected address. The interviewer then selected one respondent using a computer-generated random selection procedure. Where there were two or more 'dwelling units' at the selected address, interviewers first had to select one dwelling unit using the same random procedure. They then followed the same procedure to select a person for interview within the selected dwelling unit.

Key sample design variable(s): Variable '**spoint**' can be used to identify primary sampling units to account for clustering effect. The stratification variable is 'stratid'.

Weighting: The BSAS has been weighted since 1983. In 2005 the BSAS moved to a more sophisticated set of weights that included two new components to correct for non-response and to calibrate the sample to regional sex and age population profiles. As was the case for surveys prior to 2005 the weights also take into account differing selection probabilities.

The 2008 survey has a weight called wtfactor which must be used in all analysis – the data is not preweighted.

When reporting time-series analysis, there is a small possibility that the change of weighting scheme (in 2005) could disrupt the time-series. As a precaution, NATCEN recommend that when reporting time-series analysis figures from 2005 onwards the calculations should be rerun using the old weighting structure (oldwt) to check that this does not present a radically different picture. The figures produced using the new weights (wtfactor) should still be the ones used in reporting, but any substantial differences should be mentioned in a note.

The latest user guide can be found at: http://www.data-archive.ac.uk/doc/6390/mrdoc/pdf/6390userguide.pdf

Continuous Household Survey (CHS) (Northern Ireland)

Description: **http://www.esds.ac.uk/government/nichs/**

Further description can be found on the Northern Ireland Statistics and Research Agency (NISRA) web site.

Sample design: Stratified random sample.

Information from user guide and NISRA website. Sample of 4,500 addresses drawn each year from the Land and Property Services Agency's (LPSA) list of domestic addresses. The sample is drawn from three strata. The first of these strata is the Belfast District Council area. The other two are formed by dividing the remainder of Northern Ireland into East and West along district council boundaries. Within each of these strata, a simple random sample of addresses is drawn, with size proportional to the distribution of domestic addresses on the rating list. Random samples are drawn within strata so LPSA addresses within stratum have equal selection probabilities.

Key sample design variable(s): For household level clustering in individual level analysis, the variable **caseid** identifies households.

Weighting: None used.

Living Costs and Food Survey (LCF) (formerly Expenditure and Food Survey)

Description: http://www.esds.ac.uk/government/efs/

Sample design: Multistage Stratified random sample (GB); Single stage random sample (NI).

See user guide Volume A, pg. 1[10] for further details. The sample drawn from Great Britain is stratified by Government Office Regions (sub-divided into metropolitan and non-metropolitan areas) and two 2001 Census variables; socio-economic group and ownership of cars. Postcode areas are the Primary Sampling Unit (PSU). 638 postal codes postcodes areas are randomly drawn each year. The Northern Ireland (NI) sample is a random sample drawn from the Land and Property Services Agency list. The Living Costs and Food Survey is designed predominantly for the analysis of household level expenditure.

Key sample design variable(s): Deposit does not contain the PSU variable. Appendix B3 of *Family Spending, 2009* discusses design effects in the EFS[11]. Also see: *Sampling Errors Manual* (B Butcher and D Elliot, ONS 1987).

Weighting: The LCF is weighted to adjust for non-response and to gross to population estimates. The non-response component is calculated using 2001 Census-linked data and the grossing component is calculated using population projections based on the 2001 Census.

The 2008 LCF dataset contains two weights: weighta and weightq. Weighta is an annual weight and weightq is a quarterly weight. The quarterly weight was introduced because sample sizes vary from quarter to quarter as a result of re-issuing addresses where there had been a non-contact or refusal to a new interviewer after an interval of a few months, so that there are more interviews in the later quarters of the year than in the first quarter. Spending patterns are seasonal and quarterly grossing counteracts any bias from the uneven spread of interviews through the year. Note, that the LCF survey contains an Northern Ireland boost and so the weight also counteracts the disproportionate size of the NI sample.

---

[10] http://www.data-archive.ac.uk/doc/6385/mrdoc/pdf/6385_volume_a_introduction_2008.pdf
[11] http://www.ons.gov.uk/ons/rel/family-spending/family-spending/2009-edition/family-spending.pdf

Continuous Household Survey (CHS) (Northern Ireland)

Description: **http://www.esds.ac.uk/government/nichs/**

Further description can be found on the Northern Ireland Statistics and Research Agency (NISRA) web site.

Sample design: Stratified random sample.

Information from user guide and NISRA website. Sample of 4,500 addresses drawn each year from the Land and Property Services Agency's (LPSA) list of domestic addresses. The sample is drawn from three strata. The first of these strata is the Belfast District Council area. The other two are formed by dividing the remainder of Northern Ireland into East and West along district council boundaries. Within each of these strata, a simple random sample of addresses is drawn, with size proportional to the distribution of domestic addresses on the rating list. Random samples are drawn within strata so LPSA addresses within stratum have equal selection probabilities.

Key sample design variable(s): For household level clustering in individual level analysis, the variable **caseid** identifies households.

Weighting: None used.

Life Opportunities Survey (LOS)

The Life Opportunities Survey (LOS) is a survey of people with disability in Britain. Part of the data collected also feed into the Integrated Household Survey.

Description: http://www.esds.ac.uk/government/los/ General information is also available on the ONS website : **http://www.ons.gov.uk/ons/about-ons/surveys/a-z-of-surveys/life-opportunities-survey/index.html,** as well as in the technical report of the 2009/10 edition: http://www.ons.gov.uk/ons/rel/los/life-opportunities-survey/life-opportunities-survey/technical-report.pdf.

Sample design: single-stage stratified random sample. Sampling frame is the small users Postcode address (PAF)

Key sample design variable(s): Deposit does not contain PSU (address) variable. However, only 1.5 of the originally issued sample included multi-household addresses. As a result, households can be used as PSUs, in combination with the weight variable.

Weighting: the LOS is weighted to correct for unequal selection probabilities of the households in the sample and for non-response. The former stage allows to correct for multi-address households. Non response is corrected for by computing response probabilities based on small-area level census aggregate indicators. Finally, weight were grossed to populations estimate based on the mid-year population estimates (split by age ranges, sex and Government Office Regions). The three weights are included as three separate variables: SEL_WEIGHT, NR_WEIGHT, and CAL_WEIGHT. Users who need to use these weights simultaneously will need to create a new variable, made of the value of the three weights multiplied by each other.

Family Expenditure Survey (FES)

Description: **http://www.esds.ac.uk/government/fes/**

Sample design: Multi-stage stratified random sample (GB); Systematic Random sample (NI).

See description from FES user guide[12] and Expenditure and Food Survey description (EFS)[13]. The FES sample for Great Britain is a multi-stage stratified random sample with clustering. It is drawn from the Small Users file of the Postcode Address File.

Postal sectors are the primary sample unit. 672 postal sectors are randomly selected during the year after being arranged in strata defined by standard regions (sub-divided into metropolitan and non-metropolitan areas) and two 1991 Census variables – socioeconomic group and ownership of cars. These were new stratifiers introduced for the 1996-97 survey. The Northern Ireland sample is drawn as a random sample of addresses from the Valuation and Land Agency list.

From 2000, this survey was incorporated into the EFS.

Key sample design variable(s): PSU variable not present in datasets.

Weighting: Since 1998/99 the FES data has used one weighting variable which adjusts for non-response and grosses to population estimates. The 2000-2001 weighting variable is called "weight." Appendix F of the 2000 FES Report 'Family Spending'[14] contains further details on weighting. Note that survey is over sampled so that weight also counteracts the disproportionate size of the NI sample.

---

[12] http://www.esds.ac.uk/doc/4490%5Cmrdoc%5Cpdf%5Ca4490uab.pdf

[13] Sample design of FES identical to EFS, although socio-economic and car ownership stratifiers were only introduced in the 1996-7 survey see volume 1 of the 2000-2001 FES user guide.

[14] http://www.ons.gov.uk/ons/rel/family-spending/family-spending/2000-edition/family-spending.pdf

National Food Survey (NFS)

Description: **http://www.esds.ac.uk/government/nfs/**

Sample design: Multi-stage stratified (GB) and systematic random sample (Northern Ireland).

The latest Defra report available is for the 1998 survey, which was published in 1999[15]. Appendix A of this document provides information on the sample design of the survey. From January 1997, the primary sampling unit was postcode sectors with addresses being drawn from the Small Users Postcode Address File (PAF). The sample is stratified by three variables:

- The 24 regions that comprise the Government Office Regions Metropolitan split;
- The proportions of heads of household in Socio-Economic Groups 1 - 5 or 13 (in 3 bands);
- The proportions of households with no car.

Within each selected postcode sector, 28 addresses were sampled. Three hundred and seventy two postcode sectors (or groups of postcode sectors) were selected annually with probability proportional to size of the sector (measured as the number of addresses in England and Wales, and by multiple occupancy indicator, which gives the number of households at an address, in Scotland) and allocated equally to months. Each year half of the selected sectors were retained from the previous year's sample and half were replaced by a new selection from the same stratum.

Key sample design variable(s): PSU not included in deposit.

Weighting: Prior to the inclusion of the Northern Ireland data into the National Food Survey (1996), there was no weighting. The weight accounts for the deliberate oversampling of Northern Ireland and for differential response rates among different household types. This is described in detail in the NFS User Guide[16]. The datasets for 1996 onwards contain an Excel file called nfsweights.xls[17], which provides the weights that users should add to the files if using the NI data. From 2000, the NFS was incorporated into the EFS.

---

[15] See https://statistics.defra.gov.uk/esg/publications/nfs/1998/default.asp ????
[16] http://www.esds.ac.uk/doc/4512/mrdoc/pdf/a4512uab.pdf
[17] See http://www.esds.ac.uk/doc/4512/mrdoc/excel/nfsweights.xls

Family Resources Survey (FRS)

Description: http://www.esds.ac.uk/government/frs/

Sample design: Multi-stage stratified random sample (GB); stratified systematic random sample (NI).

The 2007-08 technical report provides the following details[18]. For GB, a two-stage sample with postcode sectors as primary sampling units, and addresses as secondary sampling units. For Northern Ireland a stratified systematic random sample was drawn from Valuation and Land Register. PSUs in Scotland are over-sampled by sampling twice the number of PSUs in Scotland than would be required under an equal-probability sample of the UK.

The stratification factors for Great Britain the 2007-08 survey, based on 2001 Census, were:

- A regional stratifier, based on the Government Office Region (GOR);
- Proportion of Household Reference Persons (HRPs) in NS-SEC groups 1-3 (8 bands);
- Proportion of adults aged 16-74 economically active (2 bands);
- Proportion of economically active males 16-74 unemployed.

The regional stratifier classified the PSUs into 27 regional strata. England was divided into 19 strata, based on the old metropolitan versus non-metropolitan county split within GOR, with London divided into four quadrants. Wales was divided into two groups of unitary authorities: the more populous southern belt and the remainder. Finally, Scotland was divided into six regions. The Northern Ireland sample is stratified by district council and ward.

Key sample design variable(s): PSU not included in deposit.

Weighting: Since 1992, the FRS used one weighting variable for two purposes (1) to gross to population and (2) to compensate for non-response. However, the 1994-1995 to 2001-2002 datasets were re-released due to the inclusion of a new (interim) grossing factor introduced to make adjustments to the FRS for low income households in Scotland. These datasets contain two weighting variables: Gross1 is the original variable and Gross2 is the new variable. From 2003-04 onwards there have been further revisions to the grossing scheme - Following a review, ONS produced revised grossing factors, incorporating both the new grossing regime and the revised population counts, have been calculated for all the years for which full-year FRS data is available, from 1994-95 onwards[19].

---

[18] Daniel, E & Chenary V. (2008) *FRS Annual Technical Report* ???
[19] See FRS 06 User guide vol. 1, pg 31-47

General Lifestyle Survey - GLF (formerly General Household Survey)

Description: http://www.esds.ac.uk/government/ghs/

Sample design: Multi-stage stratified random sample.

Appendix B of the ONS GHS 2006 Annual Report provides the following information on sample design. Two stage sample (postal code sector and address). The sample design was revised in 2000, introducing new stratifiers. (see footnote 2 and 3 in the report). The GHS is formed by 'major' and 'minor' strata:

- Postcode sectors are allocated into 30 major strata based on 10 Government Office Regions in England, 5 subdivisions in Scotland, and 2 in Wales. The English regions were divided between the former Metropolitan and non-Metropolitan counties. In addition London was subdivided into quadrants (Northwest, Northeast, Southwest and Southeast) with each quadrant being divided into inner and outer areas;
- These sectors are then ranked according to the proportion of households with no car, then divided into three bands containing approximately the same number of households;
- Within each band, sectors were re-ranked according to the proportion of households with household reference person in socio-economic groups 1 to 5 and 13, and these bands were then sub-divided into three further bands of approximately equal size;
- Finally, within each of these bands, sectors were re-ranked according to the proportion of people who were pensioners. In order to minimise the difference between one band and the next, the ranking by the pensioners and socio-economic group criteria were in the reverse order in consecutive bands.

Major strata were then divided into minor strata with equal numbers of addresses, the number of minor strata per major strata being proportionate to the size of the major stratum. Since 1984, the frame has been divided into 576 minor strata and one PSU has been selected from each per year. Of the 576 PSUs selected, 48 are randomly allocated to each month of the year. Each PSU forms a quota of work for an interviewer. Within each PSU, 23 addresses are randomly selected. In 2005, the frame was divided into 720 strata. In 2006, 588 of these were rolled forward to the next wave in the longitudinal design There were 132 pseudo wave 4 strata which were replaced and an additional 96 strata added, giving 816 for 2006. The number of PSUs has increased over time to counteract the attrition in the longitudinal sample.

Further Information on Changes in the 2006 data: The GHS methodology has changed to longitudinal data collection. The design changed in 2005 but the 2006 dataset is the first wave where a proportion (68%) of the sample are people who were also interviewed the year before. It should be noted however that the dataset is still cross-sectional as it contains data from 2006 only.

Key sample design variable(s): PSU and stratification variables are not included in standard End User License data but they are available through the Special License datasets. The primary sampling unit is given by the variable 'cluster' and the variable

'major strata' gives some of the stratification information. GHS Special License datasets can be found at:

http://www.esds.ac.uk/findingData/ghsSL.asp

Weighting: 'weight06' is the variable you should use to weight the data[20]. This weight applies to both household and individual level data.

---

[20] For a discussion of weighting, see Appendix D of the ONS (2006) GHS Annual Report

Health Survey for England (HSE)

Description: http://www.esds.ac.uk/government/hse/

Sample design: Multi-stage stratified random sample.

The description of the 2008 HSE is obtained from Craig, Mindell and Hirani (2009).[21] The sample for the HSE was drawn in two stages. Primary Sampling Unit (Stage 1): 1176 postcode sectors with probability of selection proportional to population of sector, Secondary Sampling Unit (stage 2): Addresses. At the first stage a random sample of primary sampling units (PSUs), based on postcode sectors, was selected. Within each selected PSU, a random sample of postal addresses was then drawn. Postcode sectors with fewer than 500 PAF addresses were combined with neighbouring sectors to form the PSUs.

Stratification: The list of PSUs in England was ordered by local authority and, within each local authority, by the percentage of households in the 2001 Census with a head of household in a non-manual occupation (NS-SEC groups 1-3). The sample of PSUs was then selected by sampling from the list at fixed intervals from a random starting point.

Once selected the 1,176 PSUs were randomly allocated to one of three sample groups:

- Group 1: 792 PSUs were allocated to a group with core and child boost sample and no accelerometry data collected;
- Group 2: 180 PSUs were allocated to a group with core only sample and accelerometry data collected; and
- Group 3: 204 PSUs were allocated to a group with core and child boost sample and
- accelerometry data collected.

Once selected, the PSUs in each group were randomly allocated to the 12 months of the year (e.g. 66 per month in Group 1 with core and child boost sample, no accelerometer) so that each quarter provided a nationally representative sample. Note that the PSUs selected for Group 3 (the core and child boost sample with accelerometry) were issued to two months. The second month was selected to be six months from that originally allocated within 2008. In one of the selected months the core addresses were issued, and in the other the child boost addresses were issued.

Key sample design variable(s): Variable '**area'** identifies PSUs apart from in 2006 and 2008 where variable is called '**psu**'. Strata variable is called '**cluster**'.

Weighting: Weighting variables are year specific owing to the variable sample design and the survey topic. For example, in 2000 weights are added for different probabilities of selection in care homes.

---

[21] Craig, R. & N. Shelton, Health Survey for England, volume 2, Methodology and Documentation, London: NatCen

In 2003, non-response weighting was introduced to the HSE data. Although the HSE has generally presented a good match to the population, this decision was taken to keep up with the recent changes on many large-scale government sponsored surveys, and with the aim of reducing the possible biases.

The 2009 HSE follows the same general weighting strategy as developed in 2003. Four types of non-response weights have been generated pertaining to analysis of household, individual, nurse, blood and saliva data. These are described in the user documentation:
http://www.data-archive.ac.uk/doc/6397%5Cmrdoc%5Cpdf%5C6397userguide.pdf

Integrated Household Survey (IHS)

The Integrated Household survey is a composite dataset made of data from the Labour Force Survey, the General Lifestyle Survey (formerly the General Household Survey), the Living Costs and Food Survey (formerly the Expenditure and Food Survey) , the English Housing Survey, and the Life Opportunities Survey (LOS) all of which share a common set of 'core' variables.

Description: http://www.esds.ac.uk/government/ihs/

Sample design: there is not a unique IHS sample design. See the respective sections of each of the component survey for specific details about their sample design.

Key sample design variable(s): analyses with the IHS that would include design variables are not possible, given that some the component survey are single stage stratified random sample, whereas others are multistage random samples.

Weighting: Weighting variables are included, which correct for the heterogeneous sample designs of the component surveys. It should be noted that this variable (HHWTxx[22]) is a *household*-level weight. More information is available in the user guide:
http://www.esds.ac.uk/doc/6584%5Cmrdoc%5Cpdf%5Cihs_user_guide_volume_1.pdf

---

[22] The last two digits indicate the year of the most recent reweighting exercise. Please note that this is not always reflected in the user guides.

Quarterly Labour Force Survey (QLFS)

Description: http://www.esds.ac.uk/government/lfs/

Sampling design: The LFS is officially described as a simple random sample, although because addresses are sampled, household level clustering occurs for individual level analysis[23]. Weak stratification also exists due to the systematic selection of postal addresses in postcode sectors. Page 36 of volume 1 of the user guide[31] discusses clustering and stratification. The effects of clustering are further discussed in Annex A of volume 1 of the user guide and tables of selected design effects are reported.

Key sample design variable(s):

Using household identifiers in the QLFS has become more complex due to the experimental introduction of new variables, in order to decrease the perceived risk of identity disclosure of the respondents. As of September 2011, the situation is the following:

In the End User License of the QLFS from the January - March 2011 quarter (Study Number 6782) onwards, a new pseudo anonymised variable (HSERIALP) allows to uniquely identify households. This variable is specific to each quarter, and does not allow linking users between quarters;

In the End User QLFS between April-June 2009 and October-December 2010, no household identifiers are available;

In the End User License edition of the QLFS until 2009 as well as in the Special License version of the QLFS (all issues), administrative variables within the LFS can be used to derive PSUs (household unit).

Households are indicated by:

From spring 1992- spring 2000
HSERIAL = (QUOTA*10000000) + (WEEK*100000) + (THISWV*10000) + (ADD*100) + HHLD

From autumn 2000
HSERIAL = (QUOTA*1000000000) + (WEEK*10000000) + (W1YR*1000000) + (QRTR*100000) + (ADD*1000) + (WAVFND*100) + HHLD (see Annex D of vol. 1 of the user guide for more details).

23

http://www.esds.ac.uk/doc/6782%5Cmrdoc%5Cpdf%5Clfs_user_guide_vol1_background2009.pdf – p38

The later definition was developed because of a mistake that meant the initial definition did not uniquely identify for later years. Please consult the LFS page on ESDS website for recent updates: http://www.esds.ac.uk/government/lfs/.

Based on information in the user guide, the following variables are used to construct a variable to identify strata (postcode sectors): wave, quarter, quota, week, and address number (ADD). The standard errors of the UK LFS estimates shown in Annex A are produced using a linearized jackknife approach by treating paired addresses (sorted by wave, quarter, quota, week and address number) as strata, and the address as a primary sampling unit (PSU).

Weights: Since 1984 the LFS has been weighted (grossed) to produce population estimates and to compensate for non-response among sub-groups. Additionally, the earnings data is also grossed. As part of the 2009 reweighting exercise new weights were released for all LFS datasets from mid-2001 onwards. A similar reweighting exercise is also planned for 2011.

The 2011 Quarterly LFS datasets have two weights (Pwt10 and Piwt10), (1) Pwt10 is the weight for individual data - this compensates for non-response and grosses to population estimates. (2) Piwt10 is the weight for income data - this weights so that that the weight of a sub-group corresponds to that sub-group's size in the population and also weights to give estimates of the number of people in certain groups. This is restricted to employees' earnings: other income data are not (yet) weighted. NB: in the near future, Pwt10 and Piwt10 will replace the weights pwt09 and piwt09 because of the re-weighting exercise to bring LFS data in line with the 2010 mid-year population estimates.

National Survey for Wales (NSW)

The NSW is a new general purpose survey commissioned by the Welsh Assembly, aimed at providing detailed information about the Welsh population at a level of precision not available in other surveys. At the time of writing this guide (September 2011), only results from the pilot study are currently available. Fieldwork for the larger size survey will begin in 2012.

Description: information is available on the website of the Welsh Assembly http://wales.gov.uk/about/aboutresearch/social/ocsropage/nationalsurveyforwales/nswtechnicalinformation/?lang=en **????**

Sample design: the originally planned consisted in a simple random survey of addresses and households. As in the LFS, individuals are clustered within these, which users should ideally take into account when analysing the data. The sampling frame was the Postcode Address File, from which all institutions and non residential addresses (except farms) where removed, using information from the Valuation Office Agency.

The actual sample design consists in a two phases stratified sample survey. During the first stage, households were randomly selected on the PAF, and all individuals aged 16 or above were interviewed within them. Failure to meet the interview targets triggered a second phase during which a subsample of 2,000 households was randomly drawn from the remaining sample with selection probability proportional to the size of the Welsh Assembly Constituency which amounted to stratified sampling with selection probability proportional to population. Only one respondent was randomly selected for interview by household.

Key sample design variable(s): ADDNO (address number), DV_WAG_PART_CONST (Welsh Assembly constituencies). The latter only applies to the second phase of the survey (which can be identified by PHASE)

Sample size: originally planned sample size was about 13,200 individuals and 8,500 households. Issues with the fieldwork forced to revise the target to 4,559 households and 6,385 individual interviews.

Weighting: Given the smaller than anticipated sample size, the data producers recommend that users of the pilot survey carry their analyses only for Wales as a whole. Two (individual-lelve) weights variables are available: WALES_WEIGHT and WALES_WEIGHT_GROSSED.

National Travel Survey (NTS)

Description: http://www.esds.ac.uk/government/nts/

Sample design: Multi-stage stratified random sample.

Pages 228 of volume 1 of the NTS user guide provide information on sample design[24]. The NTS is based on a stratified two-stage random probability sample of private households in Great Britain. The sampling frame is the Postcode Address File (PAF). For practical reasons, the Scottish islands and the Isles of Scilly were excluded from the sampling frame. This excludes 2.2% of addresses in Scotland and 0.2% in Great Britain. Each sample was drawn firstly by selecting the Primary Sampling Units (PSUs), and then by selecting addresses within PSUs. The sample design employs postcode sectors as PSUs. There were 684 PSUs in 2003 and 2004.

Following a review of the NTS methodology[25], it was decided that the NTS should introduce a quasi-panel design from 2002 onwards. According to this design, half the PSUs in a given year's sample are retained for the next year's sample and the other half are replaced. Hence 342 of the PSUs selected for the 2002 sample were retained for the 2003 sample, supplemented with 342 new PSUs. The PSUs carried over from the 2002 sample for inclusion in 2003 were excluded from the 2003 sample frame, so they could not appear twice in the sample. The dropped PSUs from 2002 were included in the sample frame. The same procedure of dropping PSUs was carried out to create the 2004 sample.

Whilst the same PSU sectors might appear in different survey years, no single addresses were allowed to be in more than one year. The PSUs which were carried over each year had different addresses selected to those selected in the same PSU in the previous year. Each year, NatCen provided the sampling company with a list of the addresses selected for the previous year's survey. These addresses were excluded from the sampling frame before either the 2003 or 2004 addresses were selected. This meant respondents to the previous year's survey in the PSUs which were carried over could not be contacted again.

This list of postcode sectors in Great Britain was stratified using a regional variable, car ownership and population density:

- The regional strata for Great Britain are based on the NUTS2[26] areas, grouped in a few cases where single areas are too small. NUTS2 roughly relates to counties or groups of counties in England, and groups of unitary authorities or council areas in Scotland and Wales. The 40 regional strata for the survey are shown on volume 1, page 28 , Figure 2-1 of the user guide.

---

[24] http://www.esds.ac.uk/doc/5340%5Cmrdoc%5Cpdf%5C5340userguide.pdf
[25] Also see: Elliott, D. (2000) ONS Quality Review of the National Travel Survey: Some Aspects of Design and Estimation Methods.
[26] Nomenclature of Units for Territorial Statistics: A European-wide geographical classification developed by the European Office for Statistics (Eurostat).

- Within each region, postcode sectors were listed in increasing order of the proportion of households with no car (according to the 1991 Census). Cut-off points were then drawn approximately one third and two thirds (in terms of delivery points) down the ordered list, to create three roughly equal-sized bands.
- Within each of the 120 bands thus created (40x3), sectors were listed in order of population density (people per hectare). 342 postcode sectors were then systematically selected with probability proportional to delivery point count. Differential sampling fractions were used in Inner London, Outer London and the rest of Great Britain in order to oversample London
- These sectors were then added to the 342 sectors carried over from the previous year's survey to make the final sample of 684 sectors for each year.

Within each selected sector, 22 addresses were sampled systematically, giving a sample of 15,048 addresses (684 postcodes x 22).

Key sample design variable(s): PSU variable not present due to perceived disclosure risk, although NTS collects information measured at the PSU level ('P-level'). The value of a P-level variable applies to all households living within that PSU. The P-level is therefore the highest level at which the data may be analysed, coming just above the H (Household) level in the analysis hierarchy. The 2003 and 2004 NTS included seventeen P-level variables. A more detailed account of the derivation of PSU-level variables is given in volume 1 of the user guide.

Weighting: The NTS does not currently employ a weighting scheme. However, the NTS over-samples in London due to the lower response rates achieved so there exist three variables weighted to gender, age and residency in London. These are very seldom used within the NTS and are not available from ESDS. They do not constitute a 'weighting scheme'. They are weighted versions of the main 'individual', 'number of stages' and 'number of journeys' variables. They are minor variables, and the corresponding unweighted variables are usually used. Details of this and the NTS sampling procedures can be found in the 2001 technical report [27]. Weighting the NTS is not straightforward because of the many levels used for analysis (household, individual, vehicle, trip etc).

---

[27] http://www.esds.ac.uk/doc/5340%5Cmrdoc%5Cpdf%5C5340userguide.pdf

Northern Ireland Labour Force Survey (NILFS)

Description: http://www.esds.ac.uk/government/nilfs/

Also see NISRA web site: **http://www.csu.nisra.gov.uk/survey.asp53.htm**

Sample design: Stratified systematic random sample.

The sample for the NILFS is drawn from the Land and Property Services Agency (LPSA) list. A systematic random sample of 650 addresses is drawn each quarter from the LPSA list. Geographical stratification: Sample sorted by district council and ward.

Key sample design variable(s): Stratified random sample although clustering at household level for individual level analysis.

Weighting: See details for UK Labour Force Survey (UKLFS).

Northern Ireland Life and Times Survey

Description: http://www.esds.ac.uk/government/nilts/

Sample design: Systematic random sample. Land and Property Services Agency provides sampling frame.

Key sample design variable(s): No household clustering. 1 person per household selected.

Weighting: All analyses of the adult data should be weighted in order to allow for disproportionate household size. In 2008 the weighting variable is called WTFACTOR. The only exceptions are the few household variables (for example, tenure and household income), which do not need to be weighted.

Latest userguide:
http://www.esds.ac.uk/doc/6546%5Cmrdoc%5Cpdf%5C6546userguide.pdf

ONS Opinions Survey (formerly Omnibus Survey)

Description: http://www.esds.ac.uk/government/omnibus/

Sample design: Multi-stage stratified random sample. One person per household interviewed.

The following draws on the description of the sample given in the 2007 user guide[28]. The Opinions Survey uses the Postcode Address File (PAF) as its sampling frame. A new sample of 67 postal sectors is selected for each month the survey is conducted.

This is stratified by:

- Region;
- The proportion of households where the household reference person is in the National Statistics Socio-economic Classification (NS-SEC) categories 1 to 3;
- The proportion of people who are aged over 65.

The postal sectors are selected with probability proportionate to size and, within each sector, 30 addresses (delivery points) are selected randomly. If an address contains more than one household, the interviewer uses a standard ONS procedure to randomly select where to interview – this may be at one, two or three households depending on the exact circumstances. Within households with more than one adult member, just one person aged 16 or over is selected.

Key sample design variable(s): PSU not included in deposit.

Weighting: Weighting factors are applied to Opinions data to correct for unequal probability of selection caused by interviewing only one adult per household, or restricting the eligibility of the module to certain types of respondent. The weighting system also adjusts for some non-response bias by calibrating the Opinions sample to ONS population totals. See user guide for further details[29].

The February 2007 dataset has two weights (indwgt and hhwgt). Indwgt should be applied if the unit of analysis is the individual because the weight makes the sample representative of British adults. Hhwgt should be applied if the unit of analysis is the household reference person or spouse.

---

[28] http://www.esds.ac.uk/doc/5813%5Cmrdoc%5Cpdf%5C5813userguide.pdf
[29] http://www.esds.ac.uk/doc/5813%5Cmrdoc%5Cpdf%5C5813userguide.pdf

Scottish Crime and Victimisation Survey (SCVS)

Description: **http://www.esds.ac.uk/government/scs//**

Sample design: Multi-stage stratified random sample.

In April 2008 the Scottish Crime and Justice Survey (SCJS) replaced the Scottish Crime and Victimisation Survey (SCVS) which had replaced the Scottish Crime Survey (SCS) in 2004. The following information was obtained from the 2009-10 Technical report[30].

The SCJS 2009-10 sample design differed from those of the preceding SCVS and SCS surveys in a number of important respects:

- Firstly, its planned annual sample size of 16,000 interviews was considerably larger, for example the 2006 SCVS had a sample size of 5,000;
- Secondly, the survey design required the equivalent of at least 1,000 simple random sample interviews in each Police Force Area (PFA);
- Lastly, whereas the previous surveys had completely clustered designs, the majority of the SCJS sample was un-clustered; clustering only occurred in the more sparsely populated rural areas of Scotland.

The survey follows a disproportionate sample design with at least 1000 interviews carried out in each Police Force Area regardless of population size. In urban areas the sample was systematically selected within PFA with a fixed interval giving an unclustered sample. In rural areas, data zones were selected as primary sampling units with probability proportional to population size and the sample was clustered within those areas.

The sample design included explicit stratification by Community Justice Authority Area, Police Force Area and by an urban/rural classification. In addition the sample design also employed implicit stratification whereby the addresses for the un-clustered (urban) sample were first combined into groups of contiguous data zones which formed final strata within each PFA. In those strata, addresses were ordered by postcode within output area within data zone. The target number of addresses to be sampled for the un-clustered sample in the PFA was allocated to the strata in proportion to their total addresses. Sampling intervals were calculated as total addresses divided by the target and the selected addresses were determined using a fixed sampling interval from a random start point.

For the clustered (rural) sample, data zones were used as the primary sampling units. These were ordered by the census codes for the data zones within intermediate geography area and selected with probability proportional to their populations. The addresses in the selected data zones were ordered by postcode within output area and selection was conducted by a random start and sampling interval method similar to

---

[30] http://www.esds.ac.uk/doc/5784%5Cmrdoc%5Cpdf%5C5784technicalreport.pdf

that for the un-clustered selections given above. Sixteen addresses were selected from each data zone to be sampled. Each batch of sixteen formed an interviewer assignment.

Only one adult was interviewed in each household. The majority of households contain more than one adult. Hence to avoid any bias in selection the respondent to be interviewed was determined by a random method. That random selection was implemented using an algorithm in the CAPI script. Once a selection was made, no substitutions were permitted under any circumstances.

Sampling Unit Variable: PSU not included in deposit. Page 93 of the technical report provides examples of design effects and a generalised design effect estimate for the survey.

Weighting: The SCJS is weighted for three reasons:

1. To correct the sample for unequal probabilities of selection that arose from various aspects of the sample design (e.g. disproportionate sampling of PFAs)
2. To correct the sample for differing response rates by sub-groups within the sample
3. To gross up the sample data to allow the results to be expressed as population values

The survey has a number of different weights which should be applied in different circumstances. For example, the 2009-10 SCJS has the following weights:

Weight	Files[*]	Description
WGTGHHD	RF and VFF	Gross household weight (grossed to population)
WGTGINDIV	RF and VFF	Gross individual weight (grossed to population)
WGTGINC_SCJS	VFF	Gross incident weight SCJS crimes (The values are the products of the appropriate household or individual weight and the number of incidents (the incident count), capped at five)
WGTGHHD_SC	SCF	Self-completion household weight (grossed to population)
WGTGINDIV_SC	SCF	Self-completion household weight (grossed to population)

RF = Respondent form. VFF = Victim form file. SCF = Self-completion form file

Separate weights are calculated for the self completion form (SCF) because of the higher levels of non-response compared to the respondent form. It is thought that the sensitive nature of questions in the SCF is responsible for this higher non-response. More details on the calculation of the weights in the SCJS can be found in section 8 of the SCJS 2009-2010 Technical report:
http://www.esds.ac.uk/doc/6685%5Cmrdoc%5Cpdf%5C6685_scjs_2009-10_tech_report_110321_f3671729.pdf

Scottish Health Survey (SHeS)

Description: http://www.esds.ac.uk/government/shes/

Sample design: Multi-stage stratified random sample (first stage: datazone, second stage: address, stratified by Health Board areas and the Index of Multiple Deprivation).

Section 1.2 of The Scottish Health Survey (Volume 2): Technical report informs the material given here and gives more detail on the survey sample design. It can be downloaded from: http://www.scotland.gov.uk/Resource/Doc/286063/0087159.pdf

The SHeS uses stratification with the sample design involving twenty-five strata comprising each of the three island health boards (Orkney, Shetlands, and the Western Isles), and 22 other strata constructed by dividing the 11 mainland Health Boards into separate strata containing "deprived" and "non-deprived" data zones.

Whilst the 2008 SHeS was a clustered sample the survey design is intended to provide an unclustered sample for the four year period starting in 2008 (2008-2011). In order to achieve this datazones are sampled for the four years and addresses are sample each year. The sampling procedure is described below:

1. Firstly, the numbers of addresses needed to be issued in each stratum over a four-year period were calculated.

2. Next, the number of addresses needed to be sampled in each datazone over the four-year period to achieve the numbers in (i) was calculated.

3. To ensure that each year's sample was geographically clustered the datazones were put into batches, with each batch containing datazones geographically close to each other. A quarter of the batches (approximately) were randomly assigned to each of the four survey years.

4. The Year 1 (2008) addresses were randomly selected from the batches assigned to the first year and once the addresses were chosen they were clustered into interviewer assignments. Each assignment consisted of approximately 20 addresses.

5. Finally, each assignment was allocated at random to a quarter, and then to a survey month. (Year 1 consisted of 11 survey months – February to December – but years 2-4 will include all 12 months).

For the main sample all adults aged 16 years and over at each household were selected for the interview (up to a maximum of ten adults). However, in order to limit the burden on households with three or more children (aged 0 15), two of the children were randomly selected for inclusion in the survey. No interviews were attempted with the other children in the household.

Key sample design variable(s): PSU is included in 2008 deposit (variable name: 'psu') as is information on stratification (variable name: 'strata')

Weighting: Weighting has been used to correct for different selection probabilities and for non-response. The non-response weights were designed to adjust for non-contact and for refusals of entire households, the non-response of individuals within responding households, and non-response to specific aspects of the study (the KAM module, the nurse visit, the blood sample). Separate weights exist for adults and children.

Pages 4-5 of the 2009 user guide provide more detail on the various weights in the SHeS and when each should be used. The user guide can be downloaded from: http://www.esds.ac.uk/doc/6713%5Cmrdoc%5Cpdf%5C6713dataset_documentation.pdf

English Housing Survey (EHS)

In April 2008 the Survey of English Housing (SEH) merged with the English Housing Condition Survey (EHCS) to form the new English Housing Survey (EHS). To find out more go to the EHS section of the Communities and Local Government web site.

The data consists of two datasets: the 'full household dataset' where interviews were carried out with the household reference person and another one, the dwellings dataset, were a subsample of the former are paired with results from property valuation. Part of the English Housing Survey data are fed into the Integrated Household Survey datasets.

Description: See: http://www.esds.ac.uk/government/EHS/index.asp

Sample design: The full household dataset of the English Housing Survey has a stratified random sample design (PSU are addresses, selected from the Small Postcode Area file). The dwelling or paired dataset, was stratified according to tenure type: households renting to local authorities or housing associations were oversampled. Achieved sample size in 2009-10 was 17,042 in the full household sample, and 7,872 in the paired dataset.

Key sample design variable(s): PSUs (ie dwellings) are not available in the data. Sampling error tables are provided (sampling_error_tables.xls)

Weighting: See the Sampling and Weighting Technical advice note for detailed information: http://www.communities.gov.uk/documents/housing/pdf/1799086.pdf. The weighting variable is AAGFH09, and is a household-level weight. Given the focus of the survey, no individual level weights are provided.

Survey of English Housing

In April 2008 the Survey of English Housing (SEH) merged with the English House Condition Survey (EHCS) to form the new English Housing Survey (EHS). The final fieldwork year for the SEH was 2007/08. To find out more go to the EHS section of the Communities and Local Government web site.

Description: See: http://www.esds.ac.uk/government/SEH/

Sample design: the SEH was a multi-stage stratified random sample (PSU: postcode sector, SSU: addresses). The information for the 2007-2008 dataset is drawn from the user guide[31]. The sample design involved stratification by Government Office Region with a distinction between metropolitan and non-metropolitan areas (except for London); proportion of households in privately rented accommodation; and proportion of household reference persons in non-manual occupations.

Once the stratification was carried out the first stage of sampling took place. This involved sampling 1,176 sectors with probability proportional to the size of the sector (i.e. the number of addresses in that sector). As the sectors were selected with probability proportional to size, the selection of an equal number of addresses per sector resulted in a sample with equal selection probabilities. This means that every address in England had the same probability of selection. Within each sector 23 addresses were selected.

Eight per cent of the sampled addresses were ineligible, most commonly because they were empty but addresses were also ineligible if they did not contain any private households (for example, institutions and addresses used solely for business purposes). Most of the remaining eligible addresses contained just one household. Where an address was multi-occupied (i.e. occupied by more than one household), interviews were sought with all households at the address.

Interviews took place week by week throughout the year beginning 12 April 2007 using computer assisted personal interview (CAPI). Interviews were sought with the household reference person or their spouse/partner at each household.

Key sample design variable(s): Not included in deposit. Appendix D of the user guide[42] includes information on standard error calculation and design effects.

Weighting: See Appendix C of user guide. The SEH has been weighted since 1994/95 to produce population estimates and to compensate for different response rates among households. The 2007-2008 dataset has two weight variables (H4b and H4bt), both of which combine weights for non-response and grossing. H4b weights for non-response and grosses to households in England (in 000s) and h4bt: weights for non-response and grosses to tenancy groups in England (in 000s).

---

[31] http://www.data-archive.ac.uk/doc/6399%5Cmrdoc%5Cpdf%5C6399userguide.pdf

Time Use Survey (TUS)

Description: http://www.esds.ac.uk/government/timeuse/

Sample design: Multi-stage stratified random sample.
See User Guide vol 1[32] and the technical report[33]. The primary sampling unit consisted of postcode sectors stratified into five Government Office Region combinations. Within these postcode sectors, account was taken of the population density and the social-economic group of the head of the household.
Postcode Address File (PAF) was employed as the sampling frame in England, Wales and Scotland. As this is not available in Northern Ireland the Value and Lands Agency (VLA) list was used. The primary sampling units (PSUs) consisted of postcode sectors in Great Britain, and Wards in Northern Ireland. Any postal sectors with less than 500 addresses were amalgamated with the adjacent sector. The 52 week year was divided into thirteen fieldwork months, each of which covered a nationally representative sample. To this end, PSUs and wards were systematically allocated to fieldwork months and then to weeks.

Key sample design variable(s): PSU not included in data archive deposit. Section 7.3 of technical report discusses complex sampling errors[34].

Weighting: In July 2003, new weights (grossed and ungrossed) were introduced using 2000 population estimates based on the 2001 Census. See user guide for further details.

---

[32] http://www.esds.ac.uk/doc/4504%5Cmrdoc%5Cpdf%5C4504userguide1.pdf
[33] http://www.statistics.gov.uk/downloads/theme_social/UKTUS_TechReport.pdf . Also see http://www.statistics.gov.uk/TimeUse/sampling_design_and_fieldwork.asp
[34] http://www.statistics.gov.uk/downloads/theme_social/UKTUS_TechReport.pdf

Welsh Health Survey (WHS)

Description: http://www.esds.ac.uk/government/whs/

Sample design: Multi-stage stratified random sample.
The information reported here is drawn from the User guide[35]. The 2008 WHS followed a different sampling design to previous years to take account of different Unitary Authority (UA) response rates and to increase effective sample size in order to improve precision of estimates at UA level.

The sample for the 2008 WHS used the PAF as the sampling frame with stratification by Unitary Authority. Unlike previous surveys, which first selected postcodes then addresses, in 2008 an unclustered sample of addresses was selected from each of the 22 UAs. The decision to select an unclustered sample was based on the need for improved precision against a background of falling response rates.

A small proportion of addresses in the PAF contained more than one household. If the number of households found by the interviewer at an address selected for the WHS was three or less, then all the households were included in the WHS. However, if more than three households were found, then the interviewers were instructed to select three households to be included in the WHS. The households to be included were selected at random using a Kish grid.

Key sample design variable(s): Archhsn identifies households which allows account to be taken of household clustering when analysing data at individual level. Information on stratification not in deposit. Page 34 of the User guide discusses sampling errors.

Weighting: Weights were calculated for the WHS data to correct for unequal selection probabilities and also for survey non-response. Two sets of weights were generated, household weights (wt_hhld) and individual weights (wt_adult and wt_child). The household weights adjusted for noncontact and refusals of entire households. The individual weights, calculated separately for adults and children, adjusted for non-response among individuals within responding households.

See page 29 of the User guide for further details on weights.

---

[35] http://www.esds.ac.uk/doc/6372%5Cmrdoc%5Cpdf%5C6372userguide.pdf

4. Incorporating Complex Survey Design into your Analysis

Overview

There are two common approaches to incorporating complex survey design into your analysis: **design-based** and **model-based approaches.** Design-based methods use procedures that are robust or make adjustments to standard errors to account for design effects. Such methods can be implemented within statistical packages such as Stata (examples of which are provided in Chapter 5). Sometimes tables of design effects are also included in survey documentation, allowing you to manually adjust your standard errors. Model-based approaches in contrast account for complex samples through modelling population structure, such as by defining sampling units as levels in a hierarchical model using a multi-level modelling framework.

The main learning outcomes of this chapter are to:

- Know the main differences between model-based and design-based approaches;
- Conceptually understand design-based methods such as linearization (taylor series) methods, and replicate methods (Balanced Repeated Replication, Jackknife).

Manual adjustment using design effects (or factors)

Some ESDS surveys report design effects or factors in their user guides that can be used to adjust standard errors. We can use this information to multiply the standard errors obtained using standard software commands, which assume simple random sampling, by the design factor.

Design statistics can be reported in different formats. Sometimes, there may be an average, or 'generalized' design effect or factor reported, such as based on the average design factor across a series of variables, or the value of the Deff for the dependent variable of interest in a regression analysis. The reporting of generalised design effects in user documentation however is rare for ESDS supported datasets. Generalised design effects can also be inaccurate. For example, some characteristics cluster more than others, meaning that the design effects for some variables are likely to be bigger than the generalised figure, and for others, smaller.

An alternative approach involves using tables of reported individual Deffs associated with each variable. Once again, a limitation of this approach is that most ESDS Government surveys do not contain full lists of design effects in their user documentation.

Design-based approaches

Rather than relying on reported statistics, if you have information in your dataset on sample design, such as the cluster (and stratification) variables, you can calculate your own design adjusted statistics. Design-based approaches handle complex sample designs through adjusting estimates for the design effects of clustering and stratification in a sample design. The most common design-based techniques are **linearization (taylor series)** and **replication methods**. These different approaches can be specified using the **svyset commands** in Stata. Other techniques include the method of random groups (Wolter, 1985) and PSU-level bootstrapping (Shao and Tu, 1995, chapter 6). Although available in other commands in Stata, at the time of writing, bootstrapping techniques were not built into Stata **svyset** commands.

Linerization techniques make mathematical adjustments so that standard 'linear estimators' can be applied to data. A linear estimator is a linear function of the sample observations. In simple random samples, many estimators are linear estimators where the sample size *n* is fixed. However, in cluster sampling, situations arise where the sample size cannot be assumed fixed across different clusters, for example, in one-stage clustering where the sizes of clusters vary[36]. The **taylor series linearization** method provides a linear approximation for non-linear estimators, so that linear estimator formulas for estimating variances can be applied. In practice, your statistical software will apply linearization and estimate variances automatically in one command.

Replication methods use information on variability between estimates drawn from different subsamples of an overall sample to make inferences about variance in the population. In replication approaches, a defined number (K) of subsets (replicate samples) of the full sample are selected, and the estimation procedure is repeated for each subsample and the variance calculated. Information on the sum of these subset samples is used to estimate variances. Statistics of interest are calculated for each subsample ('replicate statistics') and then the variability between these subsample replicate statistics is used to estimate the variance of the full sample statistic. Because the formulation of the probability samples is based on the complex design, unbiased, design corrected variance estimates can be derived. This means that because we know the probability rules (from the sample design) by which the sample was drawn, we can work out what the distribution of an estimator is. This is obtained by repeating the chosen sampling method on subsamples from approximating the selection of the different possible combinations of population members to form the sample of a particular design through sub-sampling. These subsamples can be considered as independently and identically distributed, so that simple variance estimators can be applied (see Lohr, 1999: 298-315).

Replication techniques include '**Balanced Repeated Replication (BRR)** and **'Jackknife'**. Balanced Repeated Replication (BRR) requires relatively few computations when compared to jackknife and bootstrap techniques. However, it requires your sample to have only two PSU per stratum design. **Jackknife** techniques

---

[36] See Lehtonen and Pahkinen, 1994, section 5.2, pg. 137 for a discussion of the properties of ratio estimators.

in contrast can be used for where there are two PSUs per stratum or for a greater number of PSUs per stratum. Using Jackknife for unstratified surveys, one PSU at a time is ommited from the sample and the others reweighted to keep the same total weight (known as the JK1 Jackknife). For stratified designs, Jackknife removes one PSU at a time, but reweights only the other PSUs in the same stratum. Jackknife is the most computationally intensive of the different replicate methods, although is sometimes preferred to other estimation techniques for some functions such as quantiles (Lohr, 1999: 207). Generally speaking, the different approaches perform equally well for most purposes[37].

Model-based approaches

Whereas in design-based approaches, the sampling design is used to specify how variability is estimated, in model-based approaches, a model is estimated that determines how variability is estimated. For model-based inference, the sample design is generally viewed as irrelevant to estimation- as long as the specified model is correct (see Lehtonen & Pahkinen, 1994: 82). Typically, model-based approaches can incorporate clustering by specifying sampling units as levels in a multi-level model such as by using a variance components model. Stratification variables can be included as covariates. In Stata, the **xtreg** command can be used to specify a random effect at the PSU level in a linear regression analysis. Other **xt** commands allow similar implementation for a broad range of techniques (in Stata, see 'help xtset').

Implementing such approaches, we can make our estimates conditional on their membership to PSUs and so our specified population structure. Recall that whereas in simple random sampling, every person has an equal probability of being selected in the sample, in complex surveys, observations may have different selection probabilities (see Lohr, 1999: 352). Running standard regression models that assume simple random sampling on complex survey data may result in standard errors that are likely to be incorrectly estimated as they do not take into account the design effects of complex samples.

Comparing design and model-based approaches

In design-based methods, the variance is the average squared deviation of the estimate from its expected value, averaged over all samples that could be obtained using a given design (Lohr, 1999: 87). In model-based estimates, the variance is again based on the average squared deviation from the expected value, but the average is over all possible samples that could be generated from the population model (Thompson, 1997). Design-based 'survey' approaches start from the position that there is a single often finite population that we want to describe (e.g. UK population). Model-based inference in contrast assumes that the finite population values are, in fact, realizations of a super-population distribution based on a hypothetical model with super-population (model) parameters. In model-based inference, the estimators are said to

---

[37] Lehtonen, R. and Pahkinen, E, 1994: 165-168, for example discuss literature on the comparison of different approaches.

be **model unbiased**[38]. This means that if the model is correct then our estimator can be considered as unbiased.

Advantages and disadvantages of approaches

- Model based approaches are generally more efficient (smaller standard errors) than design-based approaches;
- However, model approach draws assumptions on population structure- wrong model can mean wrong results. I.e. Model-based estimates are unbiased within the structure of a particular model, so if the model is wrong, then the model-based estimators are likely to be biased. We may also under-estimate variances;
- In model based approaches, stratification variables(s) are included as covariate (s). However we may be interested in an estimate unadjusted by the stratification variable(s). For example, we might just be interested in a simple average of the wage of men and women. In a model based framework, the correct specification would mean we would obtain an estimate adjusted by our stratification variables (for example, average wage after controlling for region if the sample is stratified by region and we want to attempt to account for this).
- At the same time, there may be circumstances where your sampling units are of substantive interest, such as when we are interested in geography and place, which make model-based approaches attractive for including PSU level variables as levels in our model to estimate area level effects.

---

[38] See Lehtonen & Pahkinen, 1994: 47.

5. Design-based Complex Survey Analysis in Stata

Overview

Focussing on design-based approaches, the following section contains three workshops, introducing command for analysing complex samples in Stata:

- Workshop 1 looks at how to declare your complex sample design using **svyset** to apply linearization of replication methods (discussed in Chapter 2). Here we focus on declaring weights and Primary Sampling Units (PSUs), giving examples of their effects on our standard errors;
- In Workshop 2, we consider the topic of stratification and how to inspect your stratification variable for singleton stratum (i.e. where there is only one PSU in a stratum);
- Workshop 3 considers whether (and if so, when) you should specify sampling units beyond the PSUs, such as secondary sampling units (SSUs). Different approaches to design-based methods are also compared (linearization and replicate methods) and a brief example of model-based approaches given.

Getting Started

- Examples are given using the 2007 and 2006 Health Survey for England (HSE). For information about the datasets see: http://www.esds.ac.uk/findingData/hseTitles.asp . For details on how to register and access the data see: http://www.esds.ac.uk/support/newuser.asp .
- Workshop 1 uses the HSE 2007 individual level data (file '**hse07ai.dta'**). Workshop 2 uses the HSE 2006 data (file '**hse06ai.dta'**).
- **Bold bullet points are instructions for the workshop tasks.**

- **Create the following directory (folder) on your computer and copy the HSE06 and HSE07 data into it:**

        C:/work/

- Alternatively, change the syntax below to load the files from another directory of your choice.
- The workshops are written for Stata 10 and are still valid in Stata 11.

Workshop 1. the svyset commands in Stata: Weighting and Clustering

In this workshop, we will consider:

- How to declare the complex sample design features of you survey to Stata using the **svyset** command. We will focus for now on identifying the primary sampling units and weights (as this often satisfies for most purposes). Stratification and secondary sampling units are considered in workshop 2.
- How to create summary statistics such as frequencies, means, and cross-tabulations incorporating complex survey design (**svy**: commands).
- Conducting sub-population analysis correctly.
- Basic modelling and estimating design effects using svyset. In particular, we will focus on the effects of clustering on standard errors and on the statistical significance of findings.

Health Survey for England (HSE)

The Health Survey for England (HSE) is a series of annual surveys about the health of people living in England. Since 1994, the survey has been carried out by the Joint Health Surveys Unit of the National Centre for Social Research and the Department of Epidemiology and Public Health, Royal Free and University College Medical School, London. The survey is sponsored by the Department of Health.

Sample Design

The sample for the HSE 07 was drawn in two stages. At the first stage a random sample of primary sampling units (PSUs), based on postcode sectors, was selected. Within each selected PSU, a random sample of postal addresses (known as delivery points) was then drawn. To maximise the precision of the sample, it was selected using a method called stratified sampling. The list of PSUs in England was ordered by local authority and, within each local authority, by the percentage of households in the 2001 Census with ahead of household in a non-manual occupation (NS-SEC groups 1-3).

The sample of PSUs was then selected by sampling from the list at fixed intervals from a random starting point. 900 PSUs were selected with probability proportional to the total number of addresses within them. Selecting PSUs with probability proportional to number of addresses and sampling a fixed number of addresses in each ensures that an efficient (equal probability) sample of addresses is obtained. Once selected the 900 PSUs were randomly allocated to one of two groups; 720 PSUs were allocated to the core sample and 180 PSUs were allocated to the additional child boost sample. The core PSUs contained sampled addresses for both the core and child boost sample, the additional child boost PSUs only.

- **Q. In a sentence, how would you describe the sample design of the HSE?**
  - single stage/multi-stage-
  - (implicitly/explicitly/) (proportionately/ disproportionately) stratified/unstratified?

- **Let us take a quick look at the data. Type:**

```
clear
set memory 400m
cd c:\work\
use hse07ai.dta

log using workshop
```

- The log file will keep a record of your output
- **In command box, type:**

```
describe
```

- This gives a description of the data.
- The key variables we will be using are:

VARIABLE NAME	DESCRIPTION
Sample Design	
area[39]	Primary Sampling Unit (PSU)
hserial	Secondary Sampling Unit (SSU)
int_wt	Interview weight. (other weights for different analyses are available[40] but we will focus on the main weight.
cluster	Stratification variable (original format)
samptype	Whether main or child boost sample
Other Variables Used	
age	age
sex	sex
topqual2	Highest Qualification
econact	Economic Activity Status
cksalt	Whether add salt whilst cooking? (yes/no/salt alternative)
gor07	Government Office Region (GOR) (called gor06 in the 07 data we will use in workshop 2 and 3)

---

[39] In 2006, for one year it was called 'psu'.

[40] In 2007, non-response weights have been calculated for both adults and children. Four sets of non-response weights have been generated in total. Firstly a household weight was calculated to adjust for non-contact and for refusals of entire households (hhld_wt). In addition, three sets of weights have been calculated to adjust (a) non-response among individuals in responding households, interview weight (int_wt), (b) non-response to the nurse visit stage and (nurse_wt), (c) refusal to give a blood sample (blood_wt). There are also weights specific to individual years to use with given booster samples.

1.1. Declaring your sample design using svyset

The command **svyset** (declare data as survey data) is used to identify the sample design features of your data to Stata.

Single-stage design syntax:

svyset [psu] [weight] [, design_options options]

1.1.1. Weighting using svyset

- **We will first consider how to specify the weighting variable and consider its implications for disproportionate stratification/ sampling. Type:**

  ```
  svyset [pweight= wt_int]
  ```

- A summary of the specified design will be displayed.
- pweight means probability weight
- **Note in the 07 HSE, there are several weights for different purposes. Type:**

  ```
  describe wt_*
  ```

- **Recall there is also a child booster sample and the core sample:**

  ```
  tabulate samptype
  ```

The sample for the HSE 2007 thus comprises of two components the core (general population) sample and a boost sample of children aged 2-15. The weight (wt_int) is constructed for analysing the core sample (ignoring the boost) and so assigns the child boost sample a zero weight:

- **Type:**

  ```
  tabulate wt_int if samptype == 2
  ```

The weight for the analysis of the core and child boost together (wt_child) in contrast assigns a positive weight value to the child boost cases. However, as children are disproportionately over-sampled, this weight (wt_child) will act to weight down their influence so that they are proportional in the sample to their population size. Otherwise, they will have an impact on estimates disproportionate to their population size. This demonstrates the importance of applying weights where disproportionate stratification or sampling has taken place.

1.1.2. Including the Primary Sampling Unit (PSU)

- **'area' is the 07 PSU variable. Type:**

  ```
  svyset area [pweight= wt_int]
  ```

1.2 Descriptive statistics and sub-population analysis

Once you have used the svyset command with your data, most survey design commands can be executed by prefixing command lines with **svy:**

We will give examples of commands here in the workshop, but a more exhaustive list is provided Stata manual or by typing : help svyset.

1.2.1. Summary statistics

- **First, estimate the mean of age assuming simple random sampling (but omitting child boost sample). Type:**

  ```
  mean age if samptyp =1
  ```

- **Now see what happens when we apply the weights (but not the PSU). Type:**

  ```
  svyset [pweight= wt_int]
  ```

  ```
  svy: mean age
  ```

- Note the standard error remains similar but a greater effect of weighting is witnessed on the point estimate (i.e. the mean).

- **Now applying weights and PSU ('area'. Type:**

  ```
  svyset area [pweight= wt_int]
  ```

  ```
  svy: mean age
  ```

- The estimate of the mean remains identical to when only the weights are applied but the standard errors (and so confidence intervals) are much bigger.

- **To find out the design effect and factor, type:**

  ```
  estat effects
  ```

- The standard error of our estimate is around 43% bigger when we take into account clustering.

1.2.2. Frequencies (one-way tables)

- **These are produced using the 'svy: tab' command (tabulate). Type:**
  ```
  svy: tab sex
  ```

- **Compare standard simple random sample approach. Type:**

  ```
  tab sex if samptyp ==1
  ```

1.2.3. Cross-tabulation (two-way tables)

- **The tab command is also used for two-way tables (cross-tabs). Type:**

  ```
  svy: tab topqual2 sex
  ```

- **Standard table options such as row percent, column percent, and confidence intervals can be specified:**

  ```
  svy: tab topqual2 sex, row cell
  ```

1.2.4. Sub-population analysis

In many analyses, you may wish to focus on a sub-population, such as men or women, or a specific age group. A standard approach to this would be to use the **'if'** command (e.g. tab age if sex ==1), or to drop unwanted cases. However, svyset commands require information on the entire population size to calculate standard errors. Such approaches should therefore be avoided, and instead, the '**subpop**' command should be used (although in practice it often does not make much difference).

- **We first need a binary variable coded as 1=subpopulation of interest, 0 = not subpopulation of interest (missing if we don't know). In the data, we will generate a recode of our variable 'sex' (currently 1=men 2==women, recoding women to 0). Type:**

  ```
  recode sex (2=0), ge(male)
  ```

- **Next, apply the subpop command:**

  ```
  svy, subpop (male): mean age
  ```

- **This can be applied to most commands. For means, alternatively, we can use the 'over' command:**

  ```
  svy: mean age, over (male)
  ```

1.3 Basic multivariate analysis

In the final part of the first workshop, we will consider an example of using svyset for logistic regression. Similar procedures apply to most standard modelling techniques:

- **We will first create a binary outcome variable (1=adds salt when cooking, 2=doesn't add salt/uses salt alternative). Type:**

  ```
  ta cksalt

  recode cksalt (3=0) (2=0), ge(cksalt_b)
  ta cksalt_b cksalt
  ```

- **Model without clustering applied:**

  ```
  *(enter code as one line)

  xi: logistic cksalt_b sex i.gor07 i.econact age
  i.topqual2 [pweight= wt_int]
  ```

- **Store estimate so we can use it later (call it 'model1):**

  ```
  est store model1
  ```

- **Model with PSU specified:**

  ```
  svyset area [pweight= wt_int]

  xi: svy: logistic cksalt_b sex i.gor07 i.econact age
  i.topqual2

  est store model2
  ```

- **Compare two models using estimates table command to examine standard errors and significance. Type:**

  ```
  est table model1 model2 , se p eform
  ```

- Specified table options: se gives standard error, p give p value, eform gives exponentiated coefficients (odds ratios)

- **Compare just with stars indicating significance. Spot the difference better between the models?:**

  ```
  est table model1 model2 , star eform
  ```

- **Examine design effects (will estimate for last model run):**

  ```
  estat effects
  ```

Due to clustering in the selection procedure, individuals are not selected independently. This results in correlation within clusters that can inflate variances of estimates compared to those obtained from a simple random sample (SRS) of the same size.

Workshop 2: Stratification

2.0. Introduction

Workshop 2 considers the sometimes more labour intensive, but often analytically less rewarding topic of stratification. Stratification, as we shall see, typically has less of an impact on design effects.

The topics covered are:

- How to inspect and, if necessary, prepare a stratification variable for inclusion in your analysis using the **svydes** command and other general Stata commands.
- The effects of stratification on standard errors.

2.1. Effects of Stratification

Generally speaking, the effects of clustering on the efficiency of survey estimates tend to be greater than those of stratification, meaning that ignoring stratification can be less of a concern than ignoring clustering, although weights should not be ignored, often even more so where disproportionate stratification has been employed (see workshop 1).

In the syntax for svyset, stratification is declared as a design option. Recall the command structure:

- svyset [psu] [weight] [, design_options options]
- we will use the 06 HSE for this practical
- PSU is called PSU in 06 not area; and Government Office Region variable is gor06, not gor07.
- **Load data, type:**

  clear

  set memory 400m

  cd c:\work\

  use c:\work\hse06ai.dta

- **Recreate dependent variable from workshop 1:**

  recode cksalt (3=0) (2=0), ge (cksalt_b)

- **Confusingly, the stratification variable in the HSE is called cluster (!). In the command box, type:**

  svyset psu [pweight= wt_int], strata(cluster)

- A summary of the specified design will be displayed.

- **Run model with stratification specified:**

```
xi: svy: logistic cksalt_b  sex i.gor06 i.econact age
i.topqual2
```

```
est store m1
```

```
estat effects
```

- **Now lets compare to when we just specified the PSU (and not the strata variable)**

```
svyset psu [pweight= wt_int]
```

```
xi: svy: logistic cksalt_b sex i.gor06 i.econact age
i.topqual2
```

```
est store m2
```

```
estat effects
```

```
est table m1 m2, eform  se
```

- **stars to see significance easily**

```
est table m1 m2, eform star
```

Standard errors were slightly smaller when we included stratification, but overall there was no big enough change in the current example to alter statistical significance of coefficients?

Identifying singleton stratum and inspecting your sample structure

A common problem encountered when including stratification in your analysis is that of **singleton stratum**. When there is only one PSU within a stratum, there is not enough information from which to compute the stratum's variance, making it impossible to compute the variance of an estimated parameter in a stratified clustered design.

When this happens, the standard errors fail to be calculated, and at the bottom of the table, you should get the following message:

Note: missing standard errors because of stratum with single sampling unit.

- In such cases, further work is therefore required to detect and handle singleton stratum.

- **Firstly, the svydes (survey describe) command can be used to inspect your stratification variable, looking for singleton stratum. Type:**

```
svydes
```

```
Survey: Describing stage 1 sampling units

      Survey: Describing stage 1 sampling units

      pweight: wt_int
          VCE: linearized
  Single unit: missing
     Strata 1: cluster
         SU 1: psu
        FPC 1: <zero>


                                    #Obs per Unit
                             ----------------------------
Stratum    #Units     #Obs      min      mean      max
--------  --------  --------  --------  --------  --------
       1         2        24         8      12.0        16
       2         2        61        30      30.5        31
       3         2        51        21      25.5        30
       4         2        64        29      32.0        35
       5         2        12         2       6.0        10
       6         2         9         2       4.5         7
       7         2        13         6       6.5         7
       8         2        20         9      10.0        11
       9         2        31        15      15.5        16

 [Omitted]……..etc
```

- Explanation of Output: **Stratum** is the stratum id number given by the strata variable;
- **#units** is the number of PSUs in the strata and **#Obs** the number of observations in a given stratum. The other columns give some summary statistics on the number of observations.
- The important thing to note here: if strata have singleton PSUs then #units =1. This means they only include one PSU- its also indicated by a **(*)**
- In our current example, our stratification variable looks fine (no *)

Strata with singleton PSUs can arise for several reasons:

- For an estimator, or list of covariates, singleton strata may result from missing cases. For example, for a mean, there may be missing cases for all the observations in a particular stratum except for those in a single PSU.
- **In such cases, use the svydes command with a list of variables you are interested in to check whether there are singleton stratum. Type**

  svydes sex gor06 econact age topqual 2 cksal t_b

- Another source of singleton stratum is if observations are dropped from a model as they are not in an estimation sample such as in logit or probit models because a variable or group of variables perfectly predicts an outcome. In such a case, **logit** or **probit** would just terminate with an error message. To verify that this is the problem and to identify which observations are being dropped, use **logit** or **probit** with **pweight**s and the **cluster()** option (the

clusters are the same thing as PSUs). You can then use the **e(sample)** function to identify the estimation sample.

- Another possible reason : The data depositor has made an error when deriving the strata variable is another potential reason. If the strata variable on an ESDS supported dataset looks suspect, contact ESDS and we will check it out for you.

Example where there are singleton stratum:

- So everything is ok above. However, I will work through the following 'hypothetical' example where this is not the case to show you how to handle the problem, as this is a query we get:
- On another dataset, through svydes, I get:

```
                                          #Obs per Unit
                                 -------------------------
Stratum    #Units      #Obs       min       mean      max
--------   --------   --------   --------   --------   --------
   1          2          37        18        18.5       19
   2          2          30        14        15.0       16
   3          2          27        10        13.5       17
   4          1*          3         3         3.0        3
   5          2          33        12        16.5       21
   6          2          29        13        14.5       16
   7          2          26        12        13.0       14
   8          2          23        10        11.5       13
   9          2          25        10        12.5       15
  10          2          43        16        21.5       27
  11          2          59        22        29.5       37
  12          2          42        19        21.0       23
  13          2          21         4        10.5       17
```

- Note the * indicating singleton stratum (this is on 06 data)
- To deal with singleton stratum, we need to identify which cases belong to them. The list command can be used to identify case numbers (its 06 data example: `List area cluster if cluster== 4`

```
        +-----------------+
        | area    cluster |
        |-----------------|
  7022. |  539          4 |
  7023. |  539          4 |
  7024. |  539          4 |
        +-----------------+
```

On the Stata web site a number of ways of dealing with singleton Strata[41] are discussed:

- Firstly, you can treat them as missing and error, deleting them from your sample.
- Singleton strata can be specified as 'certainty units' that are centred and/or scaled using the **singleunit (method)** option on for the svyset command[42]. See help svyset.
- Alternatively, you can group with other singleton strata to treat them like they belong to the other strata. For example, we could use the **recode** cases in cluster=4 to collapse value 4 in the cluster variable into 3 :

```
ge cluster_c = cluster
replace cluster = 3 in 7022/7024
```

- We can then repeat this procedure for all singleton strata identified by svydes, collapsing them into other stratum.

Workshop 3 Further topics

3.0. Introduction

In this workshop, we will consider the following:

- Incorporating (or ignoring) multi-stage design (e.g. secondary sampling unit features) into your analysis and the 'ultimate cluster method'.
- Comparing linearization and replicate methods to complex sample analysis.
- A brief comparison of model-based and design-based approaches.

As in Workshop 2, we are using the 06 HSE data.

3.1. Incorporating (or ignoring) multi-stage design features

The Health Survey for England (HSE) is a multi-stage stratified random sample. The primary sampling unit variable is **area** and the Secondary Sampling Unit (SSU) is **address (hserial).** The following example considers what happens when we try to specify the secondary sampling unit:

Recall that the syntax for specifying multi-stage designs is as follows:

svyset psu [weight] [, design_options] [|| ssu , design_options] ... [options]

The key thing here to note is that the different stages of the sample are separated by to lines ||. You can therefore specify the design options of different stages.

- **In the following example, we will specify the primary sampling unit, secondary sampling units, and weights. Type:**

svyset psu [pweight= wt_int], strata (cluster)||hserial

- **In the output screen, you should get the following message:**

```
stage 1 is sampled with replacement, all further stages
will be ignored
```

- When using statistics packages that compute complex standard errors from multi-stage clustered samples it is generally only necessary to have a PSU variable in the dataset. Any clustering after the first stage generally does not

65

have to be identified - the variance between PSUs automatically incorporates later stages of clustering[43]. This is referred to as the 'ultimate cluster method'.

- The main exception to this rule is where a multi-stage sample design and Finite Population Correction (FPC) is specified, then information from further sampling units is required.
- Finite Population Correction adjusts design effects downwards to account for the effects of increased sample sizes, although to have a substantial effect a very large sample size is needed.
- However, because you did not specify a Finite Population Correction (FPC) at the first stage, the sampling information at the second stage is irrelevant to variance estimation.
- Overall FPC often has little or minor impact on survey research. We will not be going into this in any further detail here, but those interested in pursuing the topic further outside the workshop are directed to the PEAS website: http://www2.napier.ac.uk/depts/fhls/peas/finitepop.asp

3.2. Linearization and Replicate Methods compared

There are two common alternative approaches used in Stata for variance/ covariance estimation for complex survey designs/ These are linearization (Taylor-Series) techniques, and replicate methods (BRR, Jackknife). Linearization is the default method, although alternative estimation approaches can be specified easily in **svyset** using the **vce ()** option. In the following example, we will compare results from these different methods:

- **First we will specify the default (linearization) method. Type:**
  (This is default so we do not have to explicitly specify the vce() option).

  ```
  svyset psu [pweight= wt_int]

  xi: svy: logistic cksalt_b sex i.gor06 i.econact age
  i.topqual2

  estat effects
  ```

- **We will now repeat the example using the Jackknife replicate method. Type:**

  ```
  svyset psu [pweight= wt_int], vce(jackknife)
  ```

- **What is different about the svyset description given in the output screen when you enter this command?**

  ```
  xi: svy: logistic cksalt_b sex i.gor06 i.econact age
  i.topqual2
  ```

---

[43] See: http://www2.napier.ac.uk/depts/fhls/peas/clustering.asp#intro

```
estat effects
```

The replicate methods are computationally demanding and generally take longer to run. In the above models, the estimation method chosen (reassuringly) does not make much of a difference to our estimates of standard errors.

3.3. A brief example of the model-based approach

The following example uses xtreg.

```
 xi: xtlogit cksalt_b sex i.gor06 i.econact age i.topqual2 ,
i(psu)
```

Interpretation

- Lnsigu : psu level variance
- Sigmau standard deviation (square root lnsigu)
- rho: intra class correlation coefficient (portion of total variance accounted for by clustering (%)).
- Chi bar: significance of random effect.

Appendix: References and Web Resources

ESRC Research Methods Programme Resource:
http://www.restore.ac.uk/PEAS/index.php

Lohr, S. (1999) "Sampling Design and Analysis", Pacific Grove: Duxbury.

Lehtonen, R. and Pahkinen, E. (1994) "Practical Methods for the Design and Analysis of Complex Surveys, New York, John Wiley.

Nathan, G. and Smith, TMF (1989) "The effects of selection on regression analysis", in H. Skinner et al (eds) *The Analysis of Complex Surveys*, New York: Wiley, p 149-163.

Shao, J. and Tu, D. (1995) *The jackknife and boostrap*, New York: Springer-Verlag.

Skinner, CJD et al (1989) *The Analysis of Complex Surveys*, New York: Wiley.

Thompson, S.K. (1992) *Sampling*. New York: Wiley.

Wolter, KM (1985) *Introduction to Variance Estimation*, New York: Springer-Verlag.