



Economic and Social Data Service

Introductory Guide to the Integrated Household Survey

ESDS Government

Author: Pierre Walthery

Version: 1.0

Date: July 2011



Table of contents

1. Introduction and frequently asked questions.....	3
What is the IHS?.....	3
Components surveys and themes of the IHS	3
What datasets are currently available?	4
Themes covered by the IHS datasets	4
IHS data is currently experimental. What does that mean?.....	5
Why use the IHS data?	5
Can I pool annual IHS datasets together to get even larger sample sizes?	6
How can I can gain access to the IHS?	6
2. Technical information	7
Sample design	7
The questionnaire.....	9
Bridging variables.....	9
Imputation	10
Weights	10
Using the IHS for teaching.....	10
Precision and confidence intervals.....	11
3. Accessing and unpacking the data.....	12
3.1 Downloading the data.....	12
3.2 Unpacking the data and accessing the IHS files	14
3.3 Browsing through the documentation.....	14
3.4 The IHS User Guides	14
4. Analysing the IHS using Stata (End User License).....	17
4.1 Using Stata	17
4.2 Opening the dataset	19
4.3 Searching for a variable	19
4.4 Obtaining summary statistics: age by ethnicity	20
4.5 Simple and cross tabulations: ethnicity and health.....	24
4.6 Plotting categorical variables: health, gender and ethnicity	29
5. A more advanced example (Special License)	32

1. Introduction and frequently asked questions

This short guide is meant to provide essential information about the new Integrated Household Survey (IHS) dataset produced by the ONS as well as examples of analysis using the new features it offers to users. It is targeted at beginners and intermediate users. Some of it is compiled from the more extensive documentation available on the [Office for National Statistics' IHS webpage](#) to which users needing more information can refer. As of July 2011, due to the experimental status of the IHS some of the information provided in this may change.

What is the IHS?

The IHS is a special dataset in which the responses to similar questions from 6 existing surveys -- the 'core' questions -- are merged together into a single dataset. It is made possible by greater standardisation and integration of variables from existing Government surveys. This allows users to benefit from very large sample size (n=433,410 individuals), and produce reliable estimates even when analysing smaller groups in the population or small geographical units. Thus, strictly speaking, the IHS is not a survey in its own right since it does not perform its own specific fieldwork.

Components surveys and themes of the IHS

The 6 surveys feeding into the IHS, as of 2009 are:

- *Opinions Survey (OPN)*
- General Lifestyle Survey (GLF)
- Living Cost and Food Survey (LCF)
- *English Housing Survey (EHS)*
- Labour Force / Annual Population Survey (LFS/APS)
- *Life Opportunities Survey*

The biggest contributor is the APS, itself a compilation of several issues of the Labour Force Survey. With a sample size of 334,206 in 2010, it represents 74% of the IHS sample.

The ONS Opinions Survey was part of the IHS in 2009, but was removed from 2010 onwards. Due to funding restriction in the English Housing Survey and changes to the survey design of the Life Opportunities Survey, these surveys will not be part of the IHS datasets from April 2011 fieldwork period onwards. It is also expected that the General Lifestyle Survey is due to be phased out from January 2012, and so its data will stop contributing to the IHS from autumn 2012. It is expected that the composition of the IHS will be flexible with some surveys leaving the IHS and others entering each year.

What datasets are currently available?

The IHS is available under either an End User License or a Special License. The Special License IHS includes additional variables perceived by the ONS as causing a risk of identity disclosure for respondents for example:

- Local Authority
- household relationships/grid
- type of family unit
- 2 and 3 digits industry and occupation
- Sexual identity
- Civil partnerships
- Detailed country of birth

Users interested in these variables as well as in detailed geographical analysis are likely to need the Special License. Information about how to apply for a Special License is available on the ESDS web page [here](#). The list of variable in either datasets might evolve in the future.

At the moment three editions of the IHS, listed below in order of release, are available for download:

- April 2009 - March 2010
- July 2009 - June 2010
- October 2009 - September 2010

Each is available under End User License and Special License. Once the experimental period has ended, IHS datasets will continue to be released on a quarterly basis.

Themes covered by the IHS datasets

The IHS currently provides information about:

- Employment and the labour market: such as economic activity
- Education
- Perceived Health
- Family and household characteristics
-
- Detailed ethnicity
- Sexual identity

The design of the IHS is the result of a compromise between sample size and number of variables versus the precision of the information available. As a result, the IHS is appropriate for producing general analysis of small groups or areas (the latter in the

case of the Special License). Users interested in more specific information about health or employment should use the component surveys of the IHS directly.

IHS data is currently experimental. What does that mean?

At the moment, the IHS is considered 'experimental statistics' by the ONS. This means that the data are still being tested, do not yet meet the UK Statistics Authority standards of quality and are subject to improvement following user feedback. The definition of some variables, or the weights, for example are still subject to change without notice. The IHS is expected to reach the final status of 'national statistics' in early 2012. Feedback from users is valued and any comments about any aspects of the IHS are welcomed by ONS and ESDS-Government, using the feedback form on the ESDS website.

Why use the IHS data?

Analyses of sub-populations

The large sample size of the IHS allows detailed analysis of sub-populations, such as ethnic groups -- more than 15 main groups are available, with additional differentiation, for instance within mixed-ethnicity respondents. In the October 2009 - September 2010 dataset there were 2,517 Bangladeshi respondents, 1,575 Chinese respondents and 5,144 Black African respondents. An example of analyses with the ethnicity variables is provided in Section 4.

Smoking data

The IHS contains the largest ever sample for smoking data (158,000 in the October 2009 - September 2010 issue).

Geographical comparisons

The Special License version of the IHS is the largest dataset since the Sample of Anonymised Records of the Census that can be used for geographical comparisons (for example 2,463 respondents in Salford, 1,342 respondents in Tower Hamlets). The Special License Dataset contains data at the following geographical levels: Local Authority, County, NUTS2, NUTS3 and Government Office Region. The End-User License dataset contains data at Government Office Region only.

Sexual Identity data

A variable about sexual identity is included (allowing people to declare whether they consider themselves gay, lesbian or bisexual). This is available in the Special License dataset.

Household linkage

The Special License IHS allows linkage of household members and family units.

Harmonised with Eurostat

The IHS variables are harmonised across its component surveys. In future releases, from 2012, some IHS data will also be harmonised for Eurostat definitions so enabling common variables across the European Union, which should improve the quality of cross-national comparative analysis.

Can I pool annual IHS datasets together to get even larger sample sizes?

The IHS datasets are not designed to be pooled (i.e. combined together) to increase sample sizes. The individual id variables and the weights are not set up for pooling, as there would be double counting of some of the respondents.

How can I can gain access to the IHS?

To access the IHS data, all users must [register](#) with the Economic and Social Data Service (ESDS). You will need a username and password to register. If you are affiliated with a UK higher education institution your usual central username and password will suffice, allowing you to authenticate using the UK Access Management Federation. If in doubt, contact the IT service in your organization.

If you do not have a username and password issued by a UK HE/FE institution, you will need to [apply for a UK Data Archive username and password](#). If you need further advice go to [Login help](#) on the ESDS web site. Registered users can download/order the datasets direct from the [ESDS web site](#). Data is currently available in SPSS, Stata or tab-delimited formats. R users can download the data in Stata format and then use the *Foreign* package to convert it¹.

The IHS data is also available via the [Nesstar](#) system, which allows online exploration of data and basic exploratory analysis before choosing to download all, or a subset, of the data. Nesstar can save data into formats suitable for SPSS, Stata, SAS, Statistica, DIF (suitable for use in Excel), Dbase and NSDStat. Non-registered users of Nesstar can view descriptions of variables in datasets and basic frequency distributions, whereas access to more advanced functions requires registration.

All users requiring the IHS data for non-commercial purposes can download it free of charge. Where data is required for commercial purposes fees apply. At the time of writing these are a per usage/project fee of £450 and an additional per study number fee of £50. For all CD orders there is a flat media fee of £7.50 per study number, handling fee of £2.50 and a flat rate postage and packing fee (£3 in the UK, £4 rest of EU, £5 rest of world). All packages are sent first class via Royal Mail. See [Charges](#) on the ESDS web site for more information. Commercial users are also advised to read the [special advice](#).

1 Please note that this is still an experimental feature and that no support is provided for it

2. Technical information

As we have just seen, the IHS is a combined dataset made of 6 different surveys each with its own specific design. At the moment, partial harmonisation of the sample design was achieved between the Living Costs and Food Survey and the General Lifestyle Survey. As a result, extra care is required when using the IHS. In particular, *it is not recommended to produce tables or estimates of the data without using the weights*, given the heterogeneity of possible source or error within each variable.

In order to gain a better knowledge of the IHS sample design, one needs to look at the design of each component survey. In the following section we will only cover the surveys that are still part of the IHS as at Spring 2011. Users interested in the sample design of the Opinion Survey, the English Housing Survey, or the Life Opportunities Survey can consult the [Survey page](#) on ESDS website in order to access each of these surveys documentation.

Sample design

The sample design of the component surveys of the IHS can be divided into two categories:

- stratified random sample (Annual Population Survey, Living Opportunities Survey). In these surveys, addresses are randomly selected from of Royal Mail's Postcode Address File (PAF). The Primary Sampling Units are thus *addresses*.
- multi-stage clustered random sample (Living Costs and Food Survey , General Lifestyle Survey): the sample proceeds in two or more stages. In the first stage, 638 Postcode Sectors are randomly sampled (ie up postcodes up to the 1st digit after the space). Sectors are stratified by metropolitan/non metropolitan areas and 2001 Census estimates of proportion of head of household in each Socio-Economic Group (SEG) and car ownership, and a first stage sample is randomly drawn from these. At the second stage, individual addresses within the Postcode sectors are then sampled. The Primary Sampling Units are Postcode Sectors and the secondary sampling units are addresses².

2 The sample in Northern Ireland was designed differently. Please refer to the documentation for more detail

Figure 1 Sample design of the General Household Survey



Figure 1 provides an illustration of this sample design coming from the General Lifestyle Survey. As a result of the harmonisation undertaken between the GLF and the LCF, the number of selected addresses is very close to each other (11,482 in the LCF, 11,598 in the GLF in 2009). Response rates are however different: whereas in 2009 5,019 households completed the full LCF questionnaires, which is a response rate of just under 50%, this raised to 66% in the GLF.

Whereas stratified sampling tend to reduce the standard errors and improve the precision of the estimates by comparison with simple random sampling, clustered sampling tends to increase it. The fact that two thirds of the observations in the IHS come from the Annual Population Survey is therefore a guarantee of the precision of the data -- but this does not mean that sampling error due to clustering may be ignored.

The following table gives the number of observations of each component survey in the IHS October 2009-December 2010 dataset.

Table 1 Size of the modules in the Oct 2009 - Sep 2010 IHS

	Observations	%
Annual Population Survey	334,206	74%
Living Opportunities Survey	23,369	5%
English Housing Survey	40,753	9%
Living Costs and Food Survey	11,989	3%
ONS Opinions Survey	20,981	5%

Source: Integrated Household Survey User Guide – Volume 1: IHS Background & Methodology 2010

Please note that the origin survey for each respondent cannot be identified in the IHS, even in the Special License version. This is because the IHS design and weighting is intended to be used across the whole IHS dataset, not for a dis-aggregated component.

The questionnaire

As such there is not a specific IHS questionnaire that is asked to IHS respondents. Instead there are standardised IHS questions or modules that are integrated as much as possible in the set of questions in each one of the respective component surveys. The full list of these is detailed in the Volume 2 of the IHS documentation. Not all the IHS questions are included in their component questionnaires, however. Some of them are derived at a later stage (bridging, see below), or there might be slight differences in the phrasing in some others. Below is an example using the self-declared health variable, QHEALTH1 in the IHS as it appears in the documentation.

Figure 2 Self-rated health question in the IHS documentation

Integrated Household Survey User Guide – Volume 2: 2011 Questionnaire	
HEALTH	
86. QHealth1	NLFS UK
How is your health in general; would you say it was...	
① <i>Running prompt</i>	
1. "very good,"	
2. "good,"	
3. "fair,"	
4. "bad,"	
5. "or very bad?"	
Asked if DVage >15 in LCF and OPN, Asked to all in EHS and GLF	

The documentation provides the exact phrasing of the question, and their respective derivation path in the component surveys. We can see that the question was only asked to respondents aged over 15 in the Living Costs and Food Survey, and the ONS Opinions Survey, whereas it was asked to all respondents in the English Housing Survey and the General Lifestyle Survey.

Information on the left-hand side of the page provides further detail about the availability of the questions in the component surveys: *NLFS* means that the question was asked to LFS respondents but is not included in the LFS datasets distributed by the Data Archive, whereas *UK* states that it was asked to all respondents in the UK (which is not necessarily the case, since questions such as those related to ethnicity and health are specific to the component countries of the UK).

When an IHS question differs in a component survey, the original question from which it is derived is presented in a box.

Bridging variables

In the IHS, variable bridging refers to an intermediate stage before full harmonisation of the questionnaires in the component surveys is achieved. It refers to variables in the IHS that gather information at different levels in the component datasets (i.e. with

more /less detail). In other words, bridging is about defining the smallest common denominator between the component surveys.

Imputation

Some variables with missing data had imputed values; the imputation is based on information from other individuals within the same households or similar households. More information about imputation methods used in the IHS is available in the documentation (Volume 1, p16). Variables flagging observations with imputed values are included in the IHS datasets.

Weights

Using weights when analysing IHS data is indispensable, given the heterogeneous sample design of the component surveys. IHS weights are designed to simultaneously correct mainly for

- Uneven address selection probability;
- Uneven selection probability of multi-household addresses;
- Non-response within selected households;
- Attrition (ie respondents dropping out) where a IHS component survey such as the APS has a longitudinal design and includes data from several waves);
- Heterogeneous sample size between component surveys (which makes scaling necessary);

When weights correcting for these factors have been computed, a final calibration is performed to adjust the sample totals to the UK population.

Two types of weights are available in the IHS

- the Integrated Household Weight (e.g. HHWT09x), which is the weight recommended for most analyses. Please note that it is a *household-level* weight: all respondents within the same household share the same weight;
- the Sexual Identity Weight (e.g. SIWT09n) when the analysis involves the sexual identity variable SEXID. This weight is only available in the Special License dataset.

IHS weights are computed on a quarterly basis. The number of the quarter is indicated by the last digit of the weight name. The previous two digits refer to the year. For example, in the October 2009 - September 2010 IHS dataset used later in this guide the Integrated Household Weight is named HHWT094.

Using the IHS for teaching

Although in theory, the IHS (End User License version) could be used for teaching, this is not recommended at this stage given that it still has experimental status. There

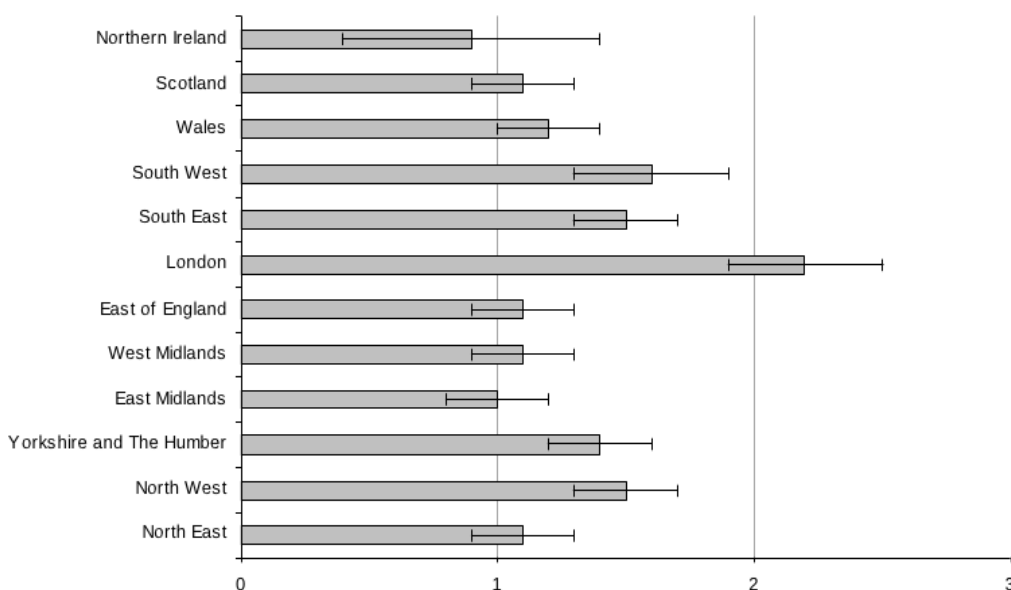
is a special procedure to be followed if datasets needs to be downloaded for teaching purposes (or if using previously downloaded datasets for teaching). All students need to be registered as users and sign the Access Agreement for Teaching. More information about this, as well as a PDF version of the Agreement are available on [the Restrictions on Use page](#) of the ESDS Website. As such, Special License versions of ESDS Government datasets cannot be used for teaching.

Precision and confidence intervals

The large number of observations of the IHS is meant to provide robust estimates and reduced confidence intervals even for small groups within the British population. Figure 1 shows the proportion (and confidence intervals) of adults respondents who declared they were either Gay, Lesbian, or Bisexual, broken down by Government Office Region

Figure 1
Proportion of Gay/Lesbian or Bisexual adults: by Government Office Region (GOR) of England and Countries of the United Kingdom, April 2009 to March 2010

Percentages



1 The total number of eligible responders to the question was 247,623 of which 238,206 provided a valid response. The question was asked to respondents aged 16 and over and was not asked by proxy.

2 The Gay / Lesbian and Bisexual categories have been combined for this analysis.

3 The whisker bars represent the confidence intervals for each estimate (See annex 3 for more details)

Source: Office for National Statistics

3. Accessing and unpacking the data

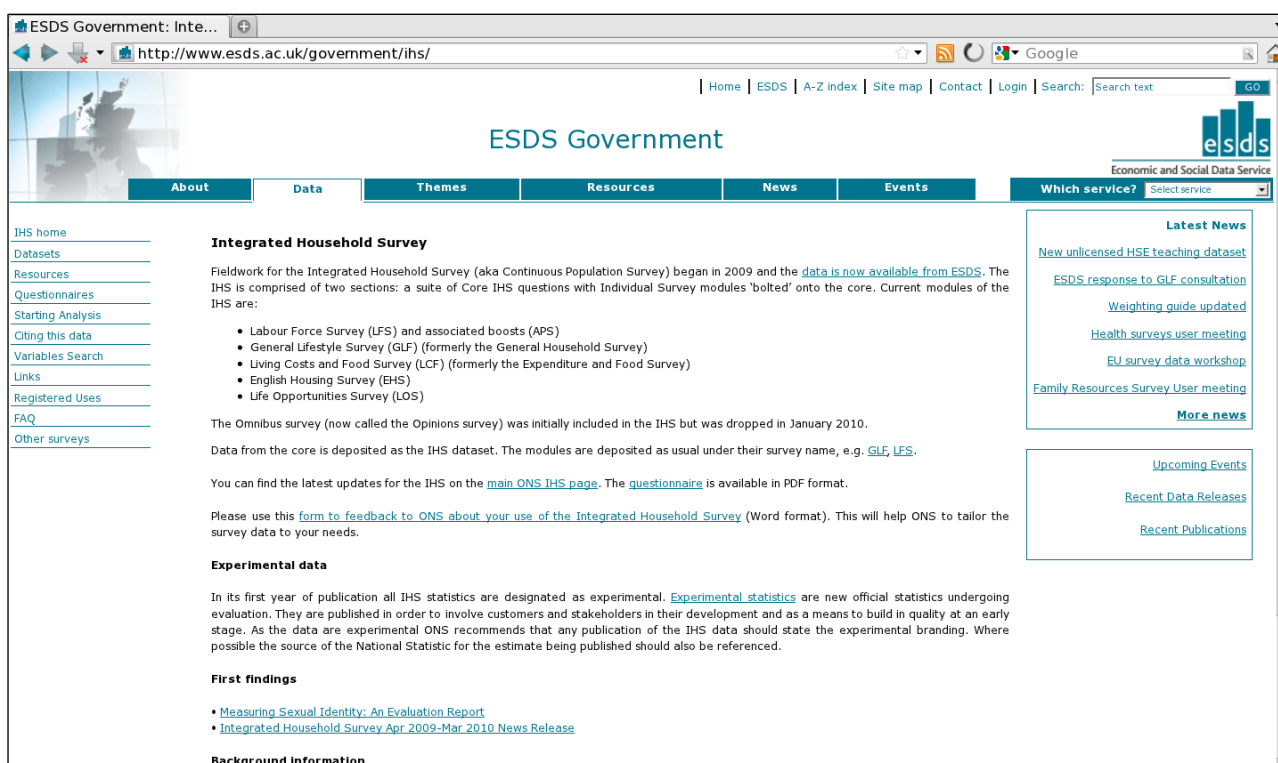
This section provides a step-by-step guide to downloading and opening the IHS datasets. It does not cover the procedure to access data via [ESDS](#).

3.1 Downloading the data

Accessing the IHS page on the ESDS Government Website

- *Browse to <http://www.esds.ac.uk/government/ihs> page.*

It should look like this:



The menu on the left of the page gives access to useful resources related to the IHS. Click on Datasets, which should lead you to the ESDS IHS Data page. The first paragraph looks like this:



- Click on the Integrated Household Survey link.

On the top of the next page, choose the version of the IHS you want to download (ie End User License or Special License). This will take you to the current list of the IHS datasets. At the time of writing (June 2011) only 3 datasets are available. When the

IHS reaches production stage, new releases will be issued at each quarter.

Integrated Household Survey list of datasets

Revised dataset due
 During March 2011, the Office for National Statistics plan to release a revised dataset for the *Integrated Household Survey* (IHS) April 2009 - March 2010. The revision is due to changes in the weighting methodology, and the file will replace the current version. The documentation (specifically Volume 1 of the user guide) will be updated to explain the change in weights.

Users should obtain the data and documentation using the tables below.

Users are advised to visit the [Integrated Household Survey](#) web pages for support in using these data, additional resources, and news and events.

SN	Study Description	Explore Online	Doc	Download / Order
6584	Integrated Household Survey, April 2009 - March 2010			<input type="checkbox"/>
6743	Integrated Household Survey, July 2009 - June 2010			<input type="checkbox"/>
6775	Integrated Household Survey, October 2009 - September 2010			<input checked="" type="checkbox"/>

▶ ADD THESE DATASETS TO MY ORDER

- Choose the version of the IHS needed by ticking the Download/order box, then click on Go.

You will now need to login into ESDS using your institution credentials. Please note that since 2009, login into ESDS via Athens is not possible any more.

After you are authenticated, you will see the following screen. The usages visible on this page are provided as examples. You will need to create your own if you haven't done this already.

Your Datasets

* If all your usages below have expired then either [register a new use of data](#) or email help@esds.ac.uk to extend the expiry date.*

To proceed with your order, please select a usage or [register a new use of data](#) and click on Add datasets.

	ID Number	Title	Expiry date
<input type="radio"/>	51824	E-stat	3/2/2013
<input type="radio"/>	12456	ESDS Government - Special Licence Datasets	31/12/2012
<input type="radio"/>	48542	PhD	10/9/2012
<input checked="" type="radio"/>	19105	ESDS Government - End User Licence Data	11/1/2012

- Select the usage to which you wish to add the IHS dataset, then Click on 'Add Datasets'

This will take you to the following screen:

Datasets for usage 48542 - PhD

Use the download column to download now or request for download or the other media column to e.g. request data on CD. Add more datasets to this usage below.

SN	Study Description	Status	Download	Explore Online	Other media
6775	Integrated Household Survey, October 2009 - September 2010		<input type="button" value="Download"/>		<input type="checkbox"/>

After you have clicked on Download, you will be offered to choose between Stata, SPSS and Tab formats. R users can choose Stata format then open the dataset using the 'read.dta()' routine from the Foreign package.

3.2 Unpacking the data and accessing the IHS files

For the sake of this example, we will assume that the dataset was downloaded in the 'My Documents/Data/IHS' folder.

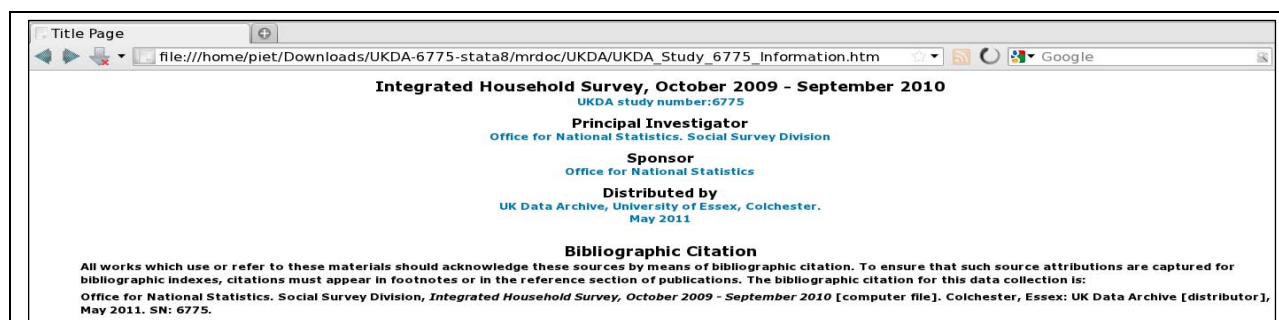
- Unzip the file using your Windows application (right click on the file and choose 'Expand').

You will now be able to access the 'UKDA-6775-stata8' folder. The documentation is accessible in the 'mrdoc' subfolder, and the data in the 'stata8' subfolder.

3.3 Browsing through the documentation

The IHS documentation is made of three types of files:

1. An abstract of the study, the content of the datasets and instructions about how to cite the dataset in bibliographies can be found in the file 'UKDA_Study_6775_Information' in the UKDA subfolder. This file needs to be opened with an internet browser.



2. A full variable list, including positions, variable names and labels (including missing values) in the `ihs_o09s_eu_ukda_data_dictionary.rtf` file. This file is useful for users who want to quickly check for the availability of specific variables or sets of variables.
3. The IHS user guides proper are available in the 'pdf' subfolder:

3.4 The IHS User Guides

The IHS User Guide is made of 4 separate files:

- *Volume 1* is the main reference to the IHS. Users are strongly advised to refer to it if they have any questions that remain unanswered by the present guide, or if they need more detail on a specific topic. Below is the Table of Content.

BACKGROUND AND METHODOLOGY
2010

CONTENTS

	PAGE
HISTORY OF THE IHS	
SECTION 1 – WHAT IS THE INTEGRATED HOUSEHOLD SURVEY.....	2
SAMPLE, DESIGN, QUESTIONNAIRE, FIELDWORK AND PROCESSING	
SECTION 2 – SAMPLE DESIGN.....	4
SECTION 3 – THE QUESTIONNAIRE.....	13
SECTION 4 – FIELDWORK.....	14
SECTION 5 – PROCESSING THE DATA - DERIVED VARIABLES AND IMPUTATION.....	16
SECTION 6 – EXPERIMENTAL STATISTICS.....	19
DATA QUALITY	
SECTION 7 – STATISTICAL QUALITY AND SAMPLING ERRORS.....	21
SECTION 8 – WEIGHTING THE IHS SAMPLE USING POPULATION ESTIMATES.....	25
PUBLICATION AND DISSEMINATION	
SECTION 9 – IHS DISSEMINATION AND PUBLICATION.....	28

- *Volume 2* contains the IHS harmonised sets of question that were asked to respondents in each of the the component surveys (the LFS, the LCF, the General Lifestyle Survey).
- *Volume 3* contains the comprehensive list of variables available in the End User and Special License datasets and where applicable basic information about the derivation of the variables.
- *Volume 4* Contains the full details of the derivation paths for all derived variable

Additional information, not included in the User guides, as well as a copy of the User Guides themselves may also be found on the [Office for National Statistics' IHS Web page](http://www.statistics.gov.uk/statbase/Product.asp?vlnk=15381)

The screenshot shows a web browser window with the address bar displaying <http://www.statistics.gov.uk/statbase/Product.asp?vlnk=15381>. The page header includes the Office for National Statistics logo and navigation links: UK Snapshot, Neighbourhood, Economy, Census, About ONS, and Jobs. The date is 7 June 2011. A banner for the 'Launch of new ONS website 28 August 2011' is visible, along with a search bar and a 'Find out more' link.

The main content area is titled 'Integrated Household Survey' and includes a 'Product' tab. A 'Note' section states: 'On 31 March 2011, the Office for National Statistics will release a revised Statistical Bulletin and Evaluation Report, using data from the experimental Integrated Household Survey (IHS) April 2009 to March 2010.' It also mentions changes to weighting methodology and provides a link to 'View more information about this product'.

A list of publications is provided, including:

- ▶ **Integrated Household Survey Apr 2009-Mar 2010 Statistical Bulletin (Updated 31/03/11)** (131Kb - Pdf)
- ▶ **Measuring sexual identity: an evaluation report** (294Kb - Pdf)
- ▶ **Integrated Household Survey Apr 2009-Mar 2010 News Release** (Pdf)
- ▶ **User Guide volume 1: IHS background & methodology 2010** (228Kb - Pdf)
- ▶ **User Guide volume 2: April 2011 questionnaire** (313Kb - Pdf)
- ▶ **User Guide volume 2: 2011 questionnaire** (292Kb - Pdf)
- ▶ **User Guide volume 2: 2009/10 questionnaire** (299Kb - Pdf)
- ▶ **User Guide volume 3: Details of IHS variables 2011** (1.1Mb - Pdf)
- ▶ **User Guide volume 4: IHS derived variables 2011** (927b - Pdf)
- ▶ **IHS Statistical Bulletin Appendix 1: Geographic Breakdown** (222Kb - Pdf)
- ▶ **IHS Statistical Bulletin Appendix 2: Sampling Errors** (232Kb - Pdf)
- ▶ **Apr 2009 - Mar 2010 User Guide complete publication** (1.5Mb - Pdf)
- ▶ **Integrated Household Survey further information** (Web link)

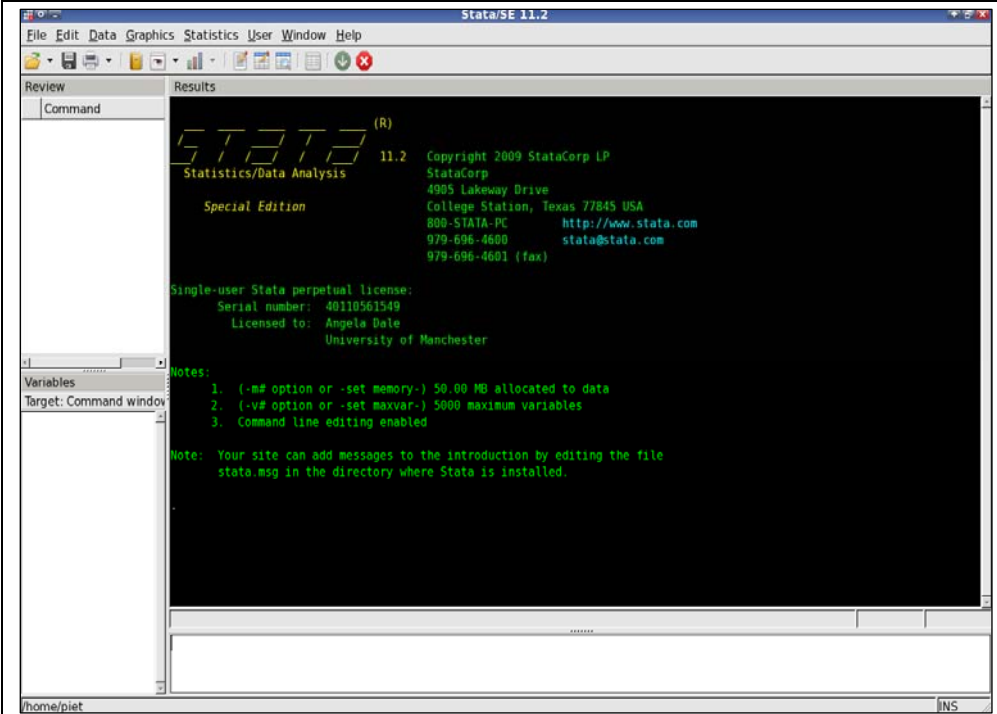
At the bottom, there are links for Accessibility, FAQs and Contact Us, Terms & Conditions, Privacy Statement, and a note that Crown Copyright applies unless otherwise stated.

4. Analysing the IHS using Stata (End User License)

In this section, a step by step typical example of analysis that can be carried out with the IHS is presented (using End User Licence data). Section 5 then goes on to explore the analysis further using the Special Licence data. The remaining sections of this guide will assume that we have downloaded a Stata version of the IHS dataset. Basic computer skills, but no prior knowledge of Stata are assumed. However, users who have never used Stata can consult the ESDS ['Introduction to Stata'](#) guide. Although any Stata command may be accessed using the pull down menus most users use the syntax version that can be saved as a separate '.do' file and reused later

4.1 Using Stata

Once you have launched Stata 11, you will be presented with this screen:



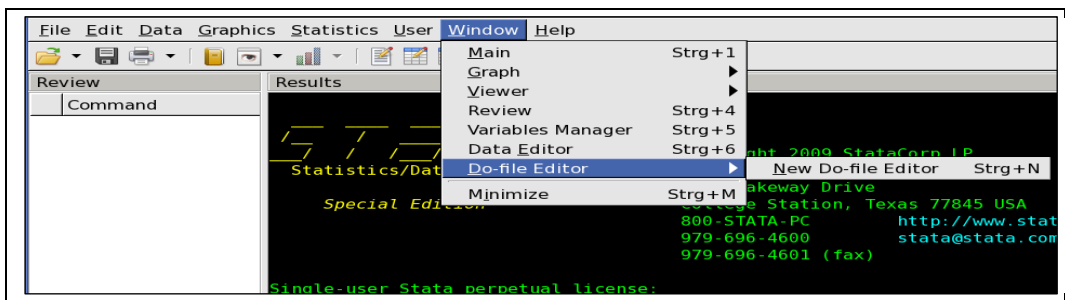
```
Stata/SE 11.2
File Edit Data Graphics Statistics User Window Help
Review Results
Command
STATA (R)
Statistics/Data Analysis 11.2 Copyright 2009 StataCorp LP
Special Edition StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
Single-user Stata perpetual license:
Serial number: 40110561549
Licensed to: Angela Dale
University of Manchester
Notes:
1. (-m# option or -set memory-) 50.00 MB allocated to data
2. (-v# option or -set maxvar-) 5000 maximum variables
3. Command line editing enabled
Note: Your site can add messages to the introduction by editing the file
stata.msg in the directory where Stata is installed.
/home/piet INS
```

The standard interface of Stata is made of 4 Windows:

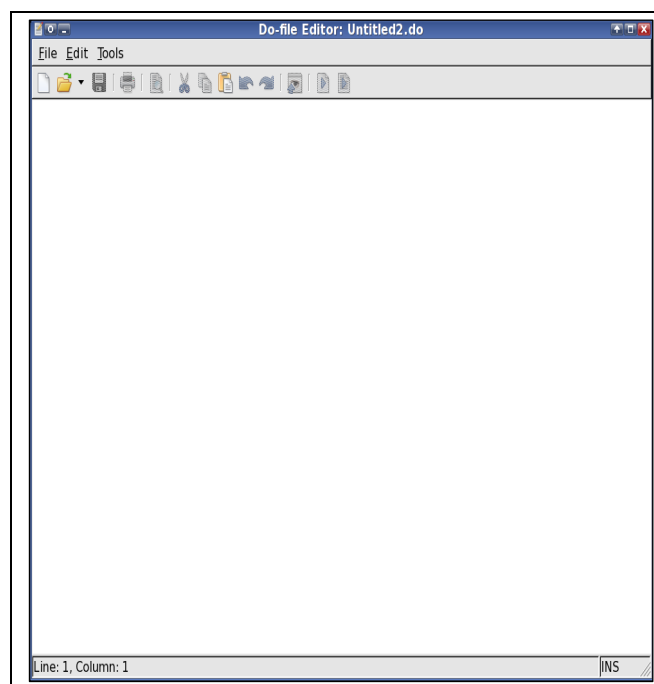
- A Review window where all command are recoded and can be repeated with a simple click;
- A Variable window where all the variables are listed by names;
- A Result window where the output of any command is printed
- A Command window at the bottom where all syntax is typed in

Alternatively, user may create a .do file where they can store all the command they will have typed in:

- Select Window: Do-file Editor: New Do-file Editor



This will open a new window in which you can type in your Stata commands:



Any command typed into the Do-file editor can be executed by selecting the corresponding line, and either typing Control-D or selecting the *Tools: Execute (do)* menu item of the Do-file Editor. The output of the command will then be printed in the Result Window.

You can then save the .do file using the File menu or by typing the following into the Command window:

save "My Documents/do_files/ihs_intro.do", replace

4.2 Opening the dataset

Let's open the IHS October 2009 - December 2010 data file we previously downloaded.

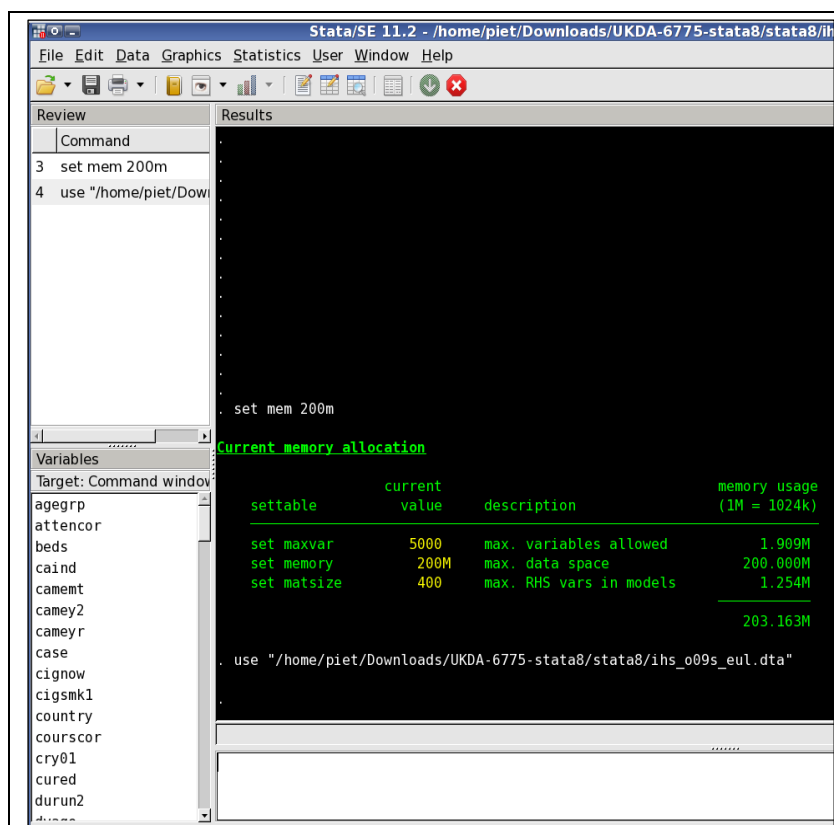
First we need to make sure there is enough memory available to Stata. We will ask Stata to give us 200 mb:

```
.set memory 200m
```

To open the file, type:

```
.use "My Documents/Data/IHS/UKDA-6775-stata8/stata8/ihs_o09s_eul.dta",  
clear
```

The previously empty Variable Windows is now filled with the names of the IHS variables, sorted alphabetically:



4.3 Searching for a variable

If we are lazy and we want to quickly find the name of variables, rather than browse through the codebook we can use the `lookfor` command, which will search all variable names and description for a keyword provided. Let's assume we are interested in cigarette smoking. We could type:

`.lookfor smok`

The following would be displayed in Stata's Results window:

```
. lookfor smok

variable name      storage   display   value   variable label
                  type     format   label
-----
cignow             byte     %8.0g    cignow  current cigarette smoker
cigsmk1            byte     %8.0g    cigsmk1 smoking status (ever smoked)
smokever           byte     %8.0g    smokever ever smoked cigarette, cigar or pipe
```

Looking at Volume 3 of the IHS documentation, we can find more information about these different variables (p. 101):

SMOKING

SMOKEVER – Ever Smoked

(1)	Yes
(2)	No

FREQUENCY: First contact on IHS module surveys
COVERAGE: Applies to all respondents aged 18 and over.
NOTES: This variable is available on the ONS research, GSS client, Special License and End User License datasets. Includes only ordinary tobacco which is smoked. Exclude any reference to snuff, tobacco or tobacco products that are chewed or sucked or herbal tobaccos. By 'ever smoked', we mean even just once in their life.

CIGNOW – Smoke at all nowadays

(1)	Yes
(2)	No

FREQUENCY: First contact on IHS module surveys
COVERAGE: Applies to all respondents aged 18 and over and when response in SmokEver is 'Yes'.
NOTES: This variable is available on the ONS research, GSS client, Special License and End User License datasets. Includes only ordinary tobacco which is smoked. Exclude any reference to snuff, tobacco or tobacco products that are chewed or sucked or herbal tobaccos.

CIGSMK1 – Smoking Status

(1)	Current cigarette smoker
(2)	Ex-cigarette smoker
(3)	Never smoked
(-6)	Child/proxy/NI
(-8)	Don't know/refusal
(-9)	DNA

FREQUENCY: First contact on IHS module surveys
COVERAGE: Applies to all respondents aged 18 and over
NOTES: This variable is available on the ONS research, GSS client, Special License and End User License datasets. This variable is derived from SMOKEVER and CIGNOW.

4.4 Obtaining summary statistics: age by ethnicity

Although the current EUL version of the IHS October 2009 - September 2010 does not contain many continuous variables we will begin this example with a brief illustration of how to use traditional descriptive statistics.

We will begin with age. Age in the IHS is provided in 2 formats:

- a categorical variable with five-year age intervals
- a continuous variable (DVAGE)

We will use the latter in the following example. In order to get the summary statistics, we will type the following command in Stata:

`sum dvage`

'Sum' is the shortcut command for 'summary'. By default, it provides the means, standard deviation, minimum and maximum value of the variable:

```

sum dvage

```

Variable	Obs	Mean	Std. Dev.	Min	Max
dvage	433410	40.43513	23.50779	0	109

We can see that the average age of the respondents is close to 40.5 years with a typical difference of 23 years. the minimum value of 0 refers to newborn babies.

We may want to go a little bit further and examine the mean age by ethnicity -- providing detailed information about ethnic groups is one of the strengths of the IHS. The IHS provides seven different ethnicity variables:

`lookfor ethn`

```

variable name      storage      display      value      variable label
type              format

```

variable name	storage type	display format	value label	variable label
eth01	byte	%8.0g	eth01	which ethnic group do you consider yourself to belong to?
ethas	byte	%8.0g	ethas	which asian ethnic group?
ethbl	byte	%8.0g	ethbl	which black ethnic group?
ethcen15	byte	%8.0g	ethcen15	15 level ethnicity coding
ethcen6	byte	%8.0g	ethcen6	6 level ethnicity coding
ethmx	byte	%8.0g	ethmx	which mixed ethnic group?
ethwh	byte	%8.0g	ethwh	which white ethnic group?

In order to gain a general overview of the age of all ethnic groups present in Britain, we will use ETHCEN15. Before looking at the results, let's consult the detailed variable description in the users' manual (Volume 3):

ETHCEN15 - Ethnicity revised	
(1)	British
(2)	Other White
(3)	White and Black Caribbean
(4)	White and Black African
(5)	White and Asian
(6)	Other Mixed
(7)	Indian
(8)	Pakistani
(9)	Bangladeshi
(10)	Other Asian
(11)	Black Caribbean
(12)	Black African
(13)	Other Black
(14)	Chinese
(15)	Other
FREQUENCY: First contact on IHS module surveys	
COVERAGE: Applies to all respondents.	
NOTES: This variable is available on the ONS research, GSS client, Special License and End User License datasets.	
ETHCEN6 and ETHCEN15 are the new variables covering Ethnic origin. They are fully in line with the Census definitions of ethnicity. ETHCEN15 is a detailed ethnic classificatory variable based on answers contained at the questions Eth01, EthWh, EthMx, EthAs and EthBl. Data in this variable has had 'other' type verbalim responses coded and re-allocated to the appropriate category.	
Please note that APS respondents in Northern Ireland who state that their ethnicity is white are not asked the detailed level question EthWh. Therefore all Northern Ireland cases have been excluded from this DV.	

Using ETHCEN15 is an improvement from previous analyses in that it uses categories identical to those in the Census and allows to monitor changes between two editions of the Census. We also learn that at the moment, due to issues with harmonisation, the information is not available for Northern Ireland. Any results will only hold for Great Britain.

The table command in Stata will allow us to get the mean age value for each ethnic category:

```
table ethcen15, c(mean age)
```

The results are displayed in the Output window of Stata. We can observe major age differences between ethnic groups.

15 level ethnicity coding	mean(dvage)
dna (ni cases)	37.9903
british	41.9088
other white	39.2236
white and black caribbean	17.7206
white and black african	18.349
white and asian	19.0299
other mixed	22.6124
indian	34.1793
pakistani	25.9527
bangladeshi	24.7537
other asian	30.1944
black caribbean	37.5414
black african	25.569
other black	26.8318
chinese	33.6971
other	30.5549

However, as we saw above, it is important that we use weights in order to get correct estimates for the reference population -- here Great Britain. In Stata, this can be done by adding the name of the weight variable to the command we just used. We saw above that at the moment this variable is named HHWT094.

Using weights in Stata is straightforward: one just needs to add the weights specification to the original command within square brackets. Please note that the weight specification *always* comes at the end of a command but *before* the comma that signals the options:

```
table ethcen15 [pw=hhwt094], c(mean age)
```

In Stata, several types of weights are allowed. PW refers to probability weights and are used when the provided weights represent unequal probability of sample selection. For more information about using weights in Stata, users might want to

consult the Stata help file by typing:

help weights

Or in Stata 11 the more detailed pdf guide can be accessed by clicking on the hyperlink to the relevant section of the Stata Manual at the bottom of the Help window opened by the previous command.

```
. table ethcen15 [pw=hhwt094], c (mean dvage)
```

15 level ethnicity coding	mean(dvage)
dna (ni cases)	36.9287
british	40.7036
other white	37.5966
white and black caribbean	18.5659
white and black african	19.2258
white and asian	19.4385
other mixed	22.804
indian	34.0367
pakistani	26.308
bangladeshi	25.0868
other asian	30.2086
black caribbean	37.3077
black african	26.1914
other black	26.4224
chinese	33.0115
other	30.5113

The White group, whether British or not, tend to be oldest group with an average age of 40. Those with mixed ethnicities are the youngest groups. Among the other ethnic groups, Bangladeshi is the youngest group with an average age of 26, closely followed by the Black African and Other Black groups. The Chinese, Indian and Black Caribbean groups are older on average and closer to the White group.

Comparing the above table of weighted means with the previous unweighted table illustrates that failing to use the weights leads to an underestimation of the average age of the Black African, Pakistani and Bangladeshi groups and an overestimation of that of the White population on the other, by about one year in each case.

In both tables, the missing Northern Irish respondents are grouped together into the first category. In case one wish a cleaner output in which they would be omitted, These cases they can be selected out of the results by adding a condition to the initial command:

```
table ethcen15 if ethcen15!=-9 [pw=hhwt094], c(mean age)
```

The `!=` operator in Stata and many other programs means 'different from'. Although the `-9` code used for the 'Do not Apply / Northern Irish' Category is not provided in the documentation, it can easily be found by typing in the command for a descriptive one way table of the ETHCEN15 variable, adding the `nolabel` option:

```
tab ethcen15,nolabel
```

4.5 Simple and cross tabulations: ethnicity and health

The majority of the variables used in the IHS are categorical, which means that we need to use univariate and bivariate frequency tables in order to get descriptive statistics of their distributions. We have just seen that simple tabulations can be obtained using the `tab` command (which is distinct from `table` we used previously to compute summary statistics). For example, if we want to know the weighted distribution of ethnic groups in Great Britain, we could type

```
tab ethcen15 [aw=hhwt094] if ethcen15>0, row nofreq
```

```
. tab ethcen15 [aw=hhwt094] if ethcen15>0, sort
```

15 level ethnicity coding	Freq.	Percent	Cum.
british	353,397.89	82.87	82.87
other white	23,642.87	5.54	88.41
indian	9,723.655	2.28	90.69
pakistani	7,444.37	1.75	92.44
other	6,896.0849	1.62	94.05
black african	6,278.9361	1.47	95.52
black caribbean	4,278.6593	1.00	96.53
other asian	4,247.0052	1.00	97.52
bangladeshi	2,800.476	0.66	98.18
chinese	1,924.0688	0.45	98.63
white and black caribbean	1,771.2591	0.42	99.05
white and asian	1,450.4377	0.34	99.39
other mixed	1,310.4522	0.31	99.69
white and black african	747.8478059	0.18	99.87
other black	556.992235	0.13	100.00
Total	426,471	100.00	

Please note that given the way the `tab` command is designed we need to use the variance weights (`aw` command) in Stata instead of the probability weights (`pw` command) if we want to produce weighted proportions and have an indication of the observed frequencies that remain close to their actual numbers in the sample. This does not affect the results -- the weights are computed by Stata in this case³.

3 Please note that using variance weights is not recommended for calculating population estimates. Should we be interested in producing population counts, the `fw` command is more appropriate in most cases. However, at the moment IHS weights are made of non-integer values, which are not managed by the `fw` command, and as a result, the more general `iw` has to be used. The proportions estimated will not change but population count estimates are produced that are more realistic. `iw` provides identical results as weights in SPSS.

```
. tab ethcen15 [iw=hhwt094] if ethcen15>0
```

15 level ethnicity coding	Freq.	Percent	Cum.
british	49305592.9	82.87	82.87
other white	3,298,621.1	5.54	88.41
white and black caribbean	247,123.66	0.42	88.82
white and black african	104,338.71	0.18	89.00
white and asian	202,363.1	0.34	89.34
other mixed	182,832.51	0.31	89.65
indian	1,356,631	2.28	91.93
pakistani	1,038,628	1.75	93.67
bangladeshi	390,718.61	0.66	94.33
other asian	592,536.39	1.00	95.33
black caribbean	596,952.74	1.00	96.33
black african	876,028.63	1.47	97.80
other black	77,710.8	0.13	97.93
chinese	268,443.47	0.45	98.38
other	962,132.4	1.62	100.00
Total	59500654.6	100.00	

Let's now move on to the next stage and consider the relationship between health and ethnicity. The IHS dataset provides a variable that measures self-reported health. We can find it quickly in Stata by typing as before:

`lookfor health`

```
. lookfor health
```

variable name	storage type	display format	value label	variable label
qhealth1	byte	%8.0g	qhealth1	general health

Let's now look at the documentation:

Integrated Household Survey User Guide – Volume 3: Details of IHS variables 2009 / 2010

HEALTH

HEALTH PROBLEMS

QHEALTH1 – How is the respondents health

(1) very good,
 (2) good,
 (3) fair,
 (4) bad,
 (5) or very bad?

FREQUENCY: First contact on IHS module surveys

COVERAGE: Applies to all respondents over 15, DVAGE>15 (Although it is asked to all ages in EHS and GLF surveys).

NOTES: This variable is available on the ONS research, GSS client, Special License and End User License datasets. This question was introduced to the APS in July 2009.

We can see that the question is only asked to respondents aged over 15, whereas previously, all household members were recorded. We can also see that health is recorded on a scale ranging from 1 to 5. There are different ways of dealing with such data. One option would be to consider QHEALTH1 as a continuous variable and examine its mean values across ethnic groups.

The other one, which we are going to explore below, consists of recoding the original QHEALTH1 into a new variable with a smaller number of categories. This way, we can keep some degree of the detail included in the original variable while minimizing the risk of having cells with numbers of observations that are too small to provide meaningful analysis.

Recoding a variable in Stata is straightforward, and is done in three stages. first, we do the actual recoding. For the purpose of this example, we will merge together the 'Good' with the 'Very good', and the 'Bad' with the 'Very bad' categories. We will then define new labels for the categories of the new variable. Finally, we 'stick' the new label to the values of the new variable:

```
recode qhealth1 (1/2=1) (3=2) (4/5=3), gen (n3qhealth1)
```

The new variable has three categories (1,2,3), for which we now need to create labels:

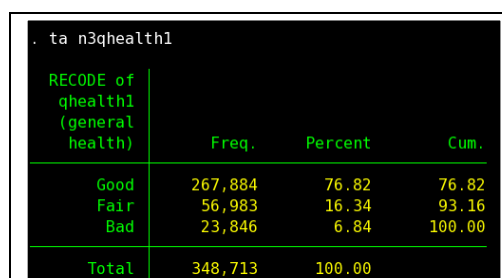
```
label define lab1 1 Good 2 Fair 3 Bad
```

We can finally 'stick' the 'lab1' label to the new variable we have just created:

```
label values n3qhealth1 lab1
```

We can inspect the results by producing a simple table of the new variable:

```
tab n3qhealth1
```



RECODE of qhealth1 (general health)	Freq.	Percent	Cum.
Good	267,884	76.82	76.82
Fair	56,983	16.34	93.16
Bad	23,846	6.84	100.00
Total	348,713	100.00	

We are now ready to produce a two-way table of our recoded health variable by ethnicity. The Stata syntax to do this is almost identical to the one we used in the previous example:

```
tab ethcen15 n3qhealth1 if ethcen15>0
```

. tab ethcen15 n3qhealth1 if ethcen15>0				
15 level ethnicity coding	RECODE of qhealth1 (general health)			Total
	Good	Fair	Bad	
british	225,488	49,459	20,617	295,564
other white	14,763	2,553	1,045	18,361
white and black carib	574	98	27	699
white and black afric	221	36	6	263
white and asian	467	73	27	567
other mixed	501	73	23	597
indian	5,420	972	404	6,796
pakistani	3,412	766	360	4,538
bangladeshi	1,200	222	139	1,561
other asian	2,075	321	117	2,513
black caribbean	1,920	553	236	2,709
black african	2,783	314	124	3,221
other black	217	38	14	269
chinese	1,154	149	30	1,333
other	3,548	473	267	4,288
Total	263,743	56,100	23,436	343,279

We notice that for most ethnic groups, sample sizes are fairly large. However, for all the mixed ethnicity groups and the Other Black group, the sample sizes are very small and likely to produce imprecise results with large standard errors and confidence intervals.

We could either produce our own recoded variable of ethnicity (grouping all from the mixed ethnicity groups into one category, and grouping the Other Black group into the 'other' category). Though a simpler option is to use the ETHCEN6 variable, which has only 6 ethnic groups.

However, since we want to keep the rich ethnic differentiation allowed by the IHS, we will adopt another solution. We will use a more rudimentary measure of health, by recoding the 3 category variable into a dichotomic version. Respondents considering themselves in 'fair' or 'bad' health are gathered into one category (fair/bad), and all those who declared being in 'good' or 'very good' health into the other (good):

```
recode qhealth1 (1/2=1) (3/5=2), gen (n2qhealth1)
```

We also want to clear up the table by only keeping percentages, not frequencies any more, since we now know that we have enough observations in all of the cells of the table. Since we are interested in comparing differences in self-reported health between ethnic groups, we also want row percentages, that will allow us to compare proportions of respondents who declared being in good health across all ethnic groups. Finally, we also need to use weights in order to prevent the results from being biased. The Stata syntax we need to type in order to achieve this is simple:

```
tab ethcen15 n2qhealth1 if ethcen15>0, row nofreq
```

```
. ta ethcen15 n2qhealth1 if ethcen15>0 [aw=hhwt094], row nofreq
```

15 level ethnicity coding	RECODE of qhealth1 (general health)		Total
	Good	Fair/bad	
british	77.98	22.02	100.00
other white	83.00	17.00	100.00
white and black carib	81.74	18.26	100.00
white and black afric	82.24	17.76	100.00
white and asian	84.36	15.64	100.00
other mixed	84.77	15.23	100.00
indian	81.61	18.39	100.00
pakistani	75.98	24.02	100.00
bangladeshi	78.56	21.44	100.00
other asian	83.14	16.86	100.00
black caribbean	72.41	27.59	100.00
black african	86.44	13.56	100.00
other black	83.58	16.42	100.00
chinese	87.81	12.19	100.00
other	83.96	16.04	100.00
Total	78.61	21.39	100.00

Black Caribbean respondents are those least likely to declare themselves in good health, by contrast with Black African and Chinese respondents. People from a mixed ethnic background tend to be in better health than White British and Asian respondents, especially those from a Pakistani background.

4.6 Plotting categorical variables: health, gender and ethnicity

Instead of exploring detailed ethnic differences in self-reported health in such great detail, we may be interested instead in combining ethnicity with gender, and look at whether we can observe additional contrasts. This time however, we want to do it with a graph rather than a table.

Since we are adding a new variable, gender, we need to produce a three way table (health by gender by ethnic group). In order to avoid the small cell problem we encountered earlier, we will need to use a less detailed version of ethnicity. The IHS provides a simplified version of ethnicity with the ETHCEN6 variable:

Integrated Household Survey User Guide – Volume 3: Details of IHS variables 2009 / 2010	
ETHCEN6 - Ethnicity revised	
(1)	White
(2)	Mixed
(3)	Asian or Asian British
(4)	Black or Black British
(5)	Chinese
(6)	Other ethnic group
FREQUENCY:	First contact on IHS module surveys
COVERAGE:	Applies to all respondents.
NOTES:	This variable is available on the ONS research, GSS client, Special License and End User License datasets.
ETHCEN6 and ETHCEN15 cover Ethnic origin. They are fully in line with the Census definitions of ethnicity. The classification has two levels. Level 1 (ETHCEN6) is a broad classification into 5 main ethnic groups. Level 2 (ETHCEN15) nests within Level 1 and provides a finer classification. ETHCEN6 does include all Northern Ireland cases.	

There are different ways to produce a graph with categorical variables in Stata. The easiest way is to use `catplot` which is a user-defined command that can be downloaded from the internet. In order to install it on your version of Stata, you need to type:

```
ssc install catplot
```

Stata will check whether it is already present on your computer, download and install it.

```
. ssc install catplot
checking catplot consistency and verifying not already installed...
installing into /home/piet/ado/plus/...
installation complete.
```

In order to check if it is now available on your version of Stata, you can type:

```
help catplot
```

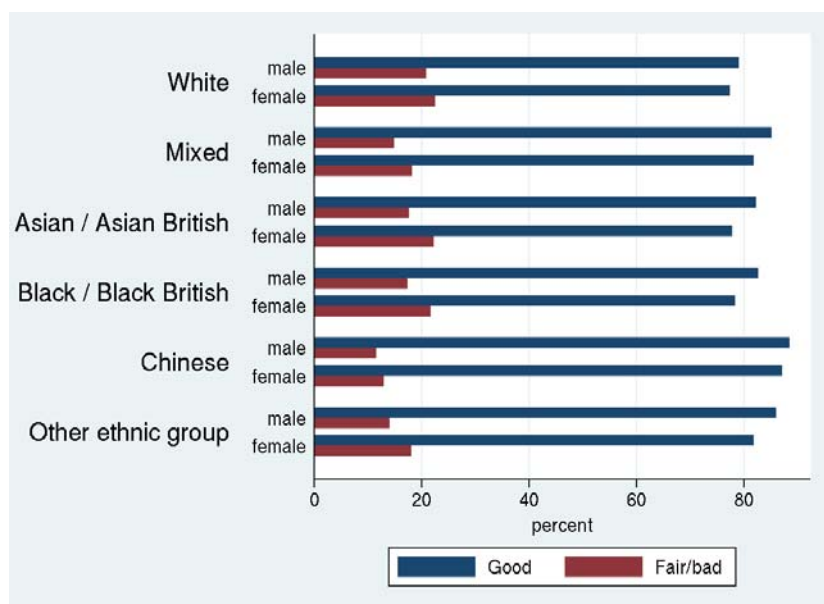
which will display the online help for the command.

We want to compare at the proportion of women and men who declared themselves in good health within each ethnic group. This is equivalent to splitting the sample according to the ethnic groups of respondents, then by gender, and finally obtain the proportion of respondents in good health within each one of these subgroups.

The syntax for using `catplot` is not very different from the one we saw before with `tab`. One needs to type the command name, followed by the name of the three variables we are interested in. Please note that the order matters, and variables needs to be entered in the reverse order they are going to appear in the graph: health then gender then ethnicity. We also want `catplot` to display the result as percentages, and these needs to the expressed as row proportions within each gender and each ethnic group, as we saw above. Finally, we want the bars to be stacked

```
catplot n2qhealth1 sex ethcen6 [iw=hhwt094] if ethcen6>0, percent(sex ethcen6)
```

This will give us the following output:

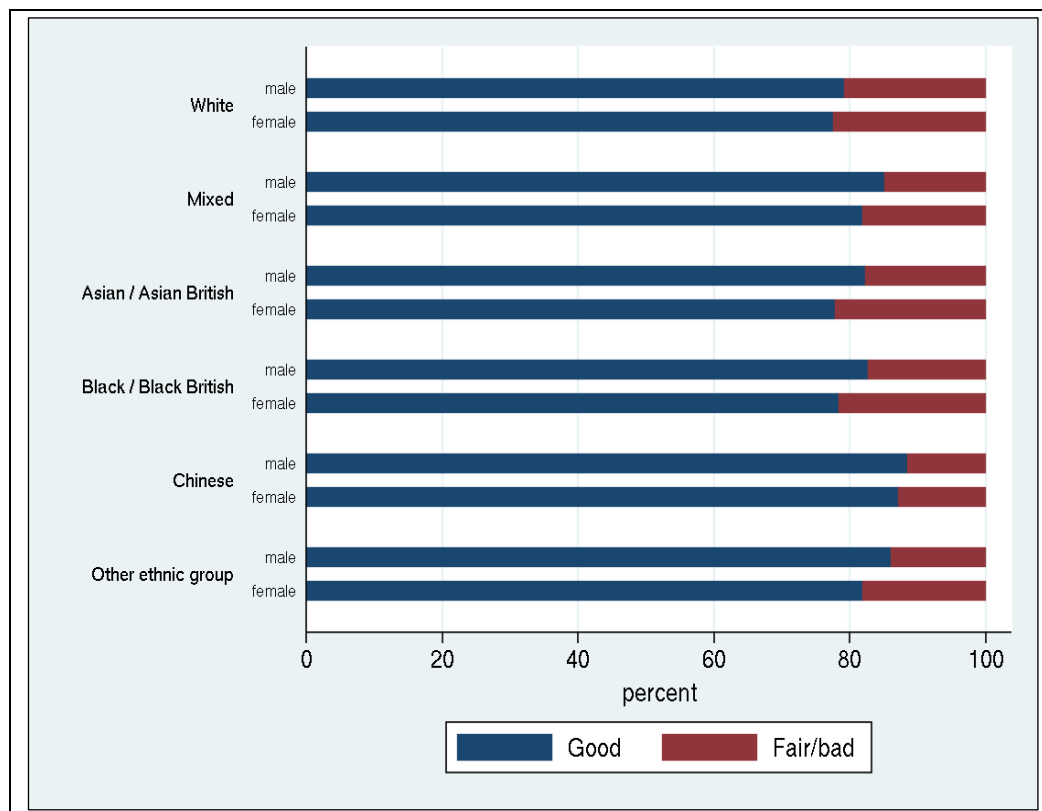


We can improve the layout of the graph by stacking the bars, reducing the size of the labels and of the legend. More information about the possibilities of `catplot` is available in the online help (`help catplot`):

```
catplot n2qhealth1 sex ethcen6 [iw=hhwt094] if ethcen6>0, ///4
percent(sex ethcen6) stack ///
var2opts(label(labsize(vsmall))) ///
var3opts(label(labsize(small))) ///
legend(size(medsmall))
```

⁴ One can use three forward slash in any Stata command (in a do file) in order to make the display of complex command easier to read.

Our new graph looks like this



One can see gender differences within most ethnic groups: women tend to be less likely to declare being in good health than men.

Finally, we may also want to save the graphs so that they can be easily imported into a Word document⁵.

```
graph export "My Documents/do_files/plot_gend_ethn_health.png", replace
```

⁵ Other commands, such as `graph bar` offer additional customization options.

5. A more advanced example (Special License)

In this section we present another example of the possibilities offered by the IHS, this time using the Special License version of the September 2009 - December 2010 data. The goal of this example is the same as in the previous section: to carry out exploratory descriptive analysis with graphical illustration.

In this example we will begin by further investigating the relationship between ethnicity and health at the Local Authority level. We will then carry out further descriptive analysis. Please note that we will be using more advanced Stata syntax.

Ethnicity and health across LADs

We want to examine whether there is a relationship between self-declared health and the presence of ethnic minorities across Local Authorities of England, Wales and Scotland (at the moment the IHS does not allow this type of analysis for Northern Ireland). In other words, we want to see whether there is a pattern of association between the proportion of respondents in good health and the proportion of non white residents in each Local Authority District (LAD).

A way to do this is to produce weighted LAD-level proportions of respondents in good health, and of white (or non-white) residents, then create a scatter plot with these values. This means that we need to create our own variables containing the percentages of respondents in self-declared good health within each LAD. We can do this by using temporary variables containing the total number of valid observations by LAD, then similarly, the number of respondents in good health. However, this is made more complicated by the fact that we need to use weights, and we have to do it manually: Stata has commands to create summary variable with percentages, but does not allow using weights with them.

We now know how to open a file in Stata (the command necessary to open the Special License dataset are the same as in the previous section). The name of the October 2009 - December 2010 Stata SL IHS file is `ihs_o09s_trusted_protect.dta`.

First we need to create a variable with the total number of working-age respondents who have provided a valid answer to the health question. `UALAGDB` is the name of the variable containing the Local Authority codes for Great Britain, and we will be using the dichotomic health indicator we previously used.

```
by ualadgb: egen uqhl_t_w=total(hhwt094) if qhealth1!=.
```

The `by` statement before the column tells Stata to repeat the command that comes after the column for each value of `UALAGDB`. `egen` is a variable creating command (the `gen` stands for generate). `total` sums the value of the weighting variable (`hhwt094`) between brackets: the value created thus reflects the unequal importance of the observations, rather than giving them equal importance (which would have

implicitly been the case had we simply given each observation the value of '1'). Finally, the condition statement at the end tells Stata to do this summation for valid cases of the health variable only.

Let's now repeat the command to create a variable with the number of respondents in good health. In order to avoid problems when we create a dataset with the summary variables only we will do it in two steps:

1. We create a variable where observations of respondents in good health are weighted. Since the health variable it is a dichotomic variable, we can do this by multiplying it by the weights:

```
gen qhlt12_w=n2qhealth1*hhwt094 if n2qhealth1!=.
```

As we did before, we can now sum the weighted number of respondents in good health in each LAD:

```
by ualadgb:egen uqhlt12_w = total (qhlt12w)
```

Finally, we can create the percentage variable:

```
gen p_uqhlt12w = 100*(uqhlt12_w/uqhlt12)
```

Note that we do not need to add the `by` statement any more since the value of the summary variables are identical within each LAD.

Finally, we want to make sure that it all went smoothly by comparing the values we have just computed with the output of a normal `tab` command for randomly chosen LADs. Unfortunately, at the time of writing this guide, the Local Authority identifier UALADBG only contains the UA/LA codes, not their names in plain English. We need to have a look at the documentation in order to be able to match them:

Integrated Household Survey User Guide – Volume 3: Details of IHS variables 2009 / 2010					
UALADGB - Unitary Authorities and LADs of Great Britain					
00AA	City of London	00CX	Bradford	00NJ	Flintshire
00AB	Barking and Dagenham	00CY	Calderdale	00NL	Wrexham
00AC	Barnet	00CZ	Kirklees	00NN	Powys
00AD	Bexley	00DA	Leeds	00NQ	Ceredigion
00AE	Brent	00DB	Wakefield	00NS	Pembrokeshire
00AF	Bromley	00EB	Hartlepool	00NU	Cardiganshire
00AG	Camden	00EC	Middlesbrough	00NX	Swansea
00AH	Croydon	00EE	Redcar and Cleveland	00NZ	Neath Port Talbot
00AJ	Ealing	00EF	Stockton-on-Tees	00PB	Bridgend
00AK	Enfield	00EH	Darlington	00PD	The Vale of Glamorgan
00AL	Greenwich	00EJ	County Durham	00PF	Rhondda, Cynon, Taff
00AM	Hackney	00EM	Northumberland	00PH	Merthyr Tydfil
00AN	Hammersmith and Fulham	00EQ	Cheshire East	00PK	Caerphilly
00AP	Haringey	00ET	Halton	00PL	Blaenau Gwent

Let's take Leeds for example, which has the ONS code '00DA'. We can obtain the percentage of respondents who declared themselves in good health by using the `tab` command as shown below. We need to put the code between inverted commas, since it is an alphanumeric variable:

```
tab ualadgb n2qhealth1 if ualadgb=="00DA" [aw=hhwt094], row nofreq
```

```
. tab ualadgb n2qhealth1 if ualadgb=="00DA" [aw=hhwt094], row nofreq
```

gb local authority 2009 codes	RECODE of qhealth1 (general health)		Total
	0	1	
00DA	20.19	79.81	100.00
Total	20.19	79.81	100.00

We can compare it with the value we computed previously. A quick way to do this is to tabulate the percentage variable only for respondents in Leeds. There should be only one value, identical to the above:

```
tab p_uqhlt12w if ualadgb=="00DA"
```

```
. tab p_uqhlt12w if ualadgb=="00DA"
```

p_uqhlt12w	Freq.	Percent	Cum.
79.80608	3,001	100.00	100.00
Total	3,001	100.00	

We are now reassured that we can safely continue computing percentages the way we did. Let's now do the same for the proportion of respondents in the White ethnic group.

We first need to create a dummy variable for White/non White respondent:

```
recode ethcen6 (min/-1=.) (1=1) (2/6=0), gen (n2ethcen6)
```

Note that the `(1=1)` can be omitted (in the original variable, white respondents are already coded 1), but it is helpful to keep it in order not to get confused. We now have a variable with white respondents coded as '1', the other ones as '0', and we just need to repeat what we have done for the health variable:

First, the weighted ethnicity variable:

```
gen eth_w=n2ethcen6*hhwt094.
```

Then, the weighted total number of valid observations of the health variable by LAD:

```
by ualadgb: egen ueth_w =total(hhwt094) if n2ethcen6!=.
```

We do the same with the number of white respondents:

```
by ualadgb: egen ueth1_w =total(eth_w) if n2ethcen6!=.
```

Finally, we can create the percentage variable:

```
gen p_ueth1w = 100*(ueth1_w/ueth_w)
```

Again, we can check that the result is identical to the one we would get with the `tab` command:

```
tab ualadgb n2ethcen6 if ualadgb=="00DA" [aw=hhwt094], row nofreq
```

```
. tab ualadgb n2ethcen6 if ualadgb=="00DA" [aw=hhwt094], row nofreq
```

gb local authority 2009 codes	RECODE of ethcen6 (6 level ethnicity coding)		Total
	0	1	
00DA	10.44	89.56	100.00
Total	10.44	89.56	100.00

and

```
tab p_ueth1w if ualadgb=="00DA"
```

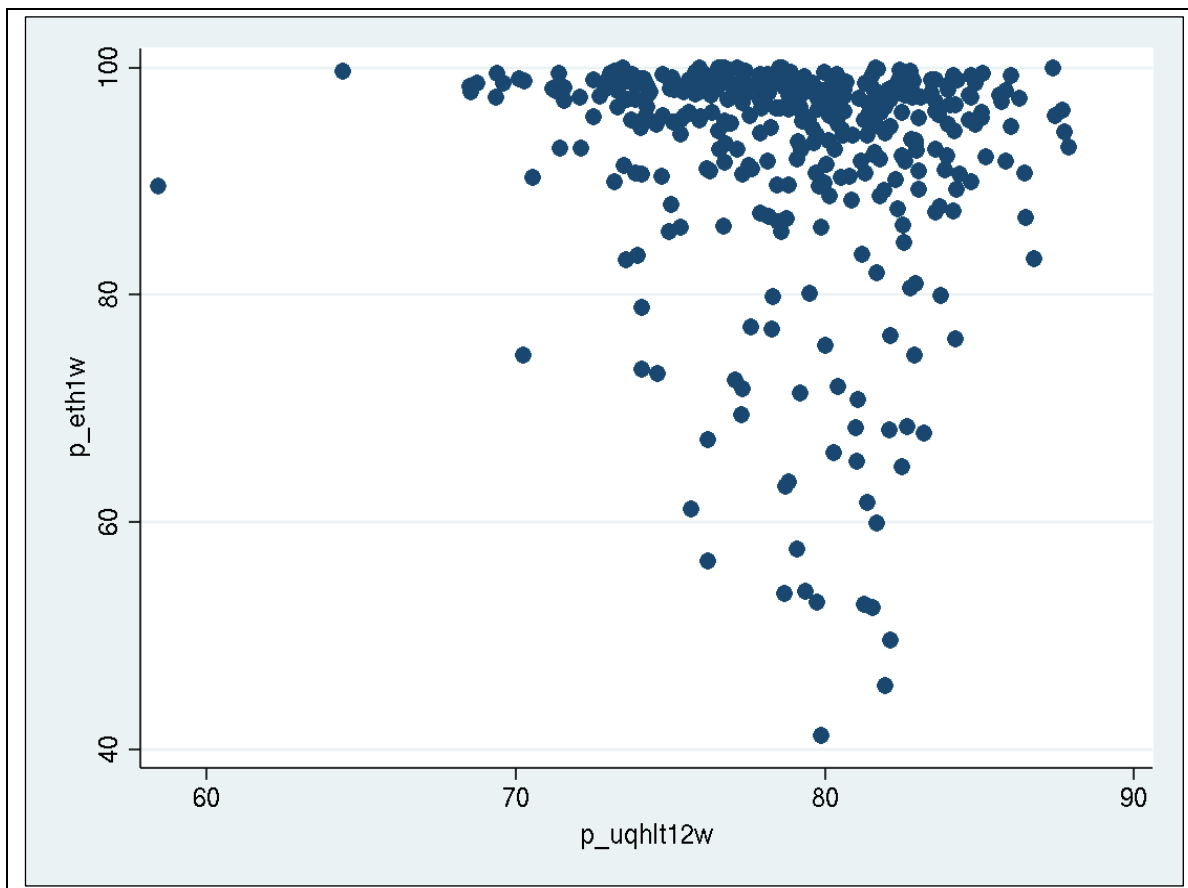
```
. tab p_ueth1w if ualadgb=="00DA"
```

p_ueth1w	Freq.	Percent	Cum.
89.5631	3,001	100.00	100.00
Total	3,001	100.00	

We are now ready to produce a scatter plot with the two variables we created.

```
scatter p_eth1w p_uqh1t12w
```

The output looks like this:



Each dot on the graph represents a local authority. We can see that no clear pattern is emerging - there is a wide dispersion of the proportion of respondents in good health (between 70 and 90%) among LADs with close to 100% white respondents. There seems to be some additional dispersion of the ethnic composition among LADs with about 80% of respondents who stated they were in good health.

Stata may have taken some time to produce the graph. This is due to the fact that it computed the dots on the graphs for each individual respondent, even if they are identical within LADs. We can use a trick in order to speed it up. We may also want to make the graph easier to interpret by adding titles to the axis.

```

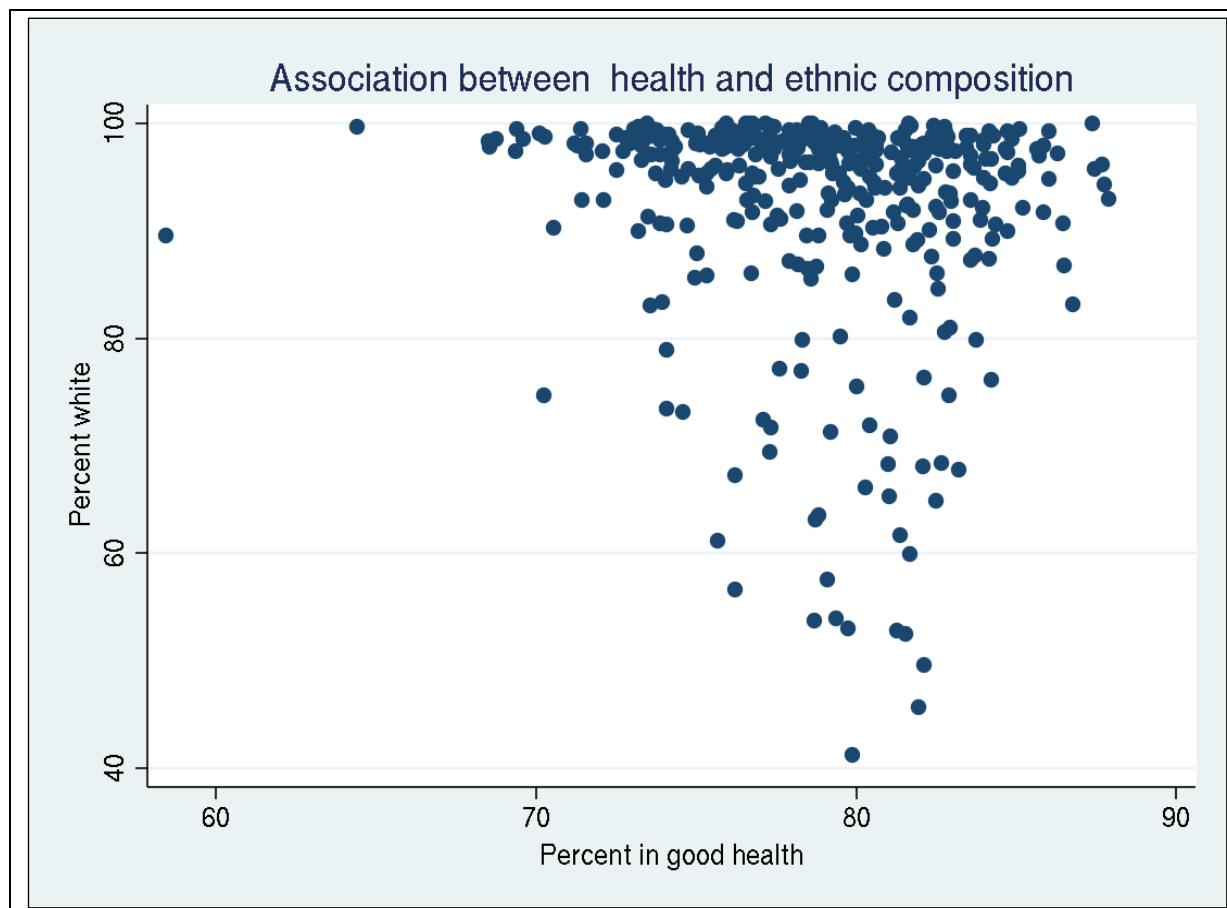
by ualadgb: gen nr=_n
preserve
keep if nr==1
scatter      p_eth1w p_uqhlt12w, ///
yttitle("Percent white") xtitle("Percent in good health") ///
title(Association between health and ethnic composition)
restore

```

The first command creates a variable, nr, with incremental observation id numbers within each LAD. _n is a system counter variable in which observations are allocated a unique number. Within each LAD thus, observations are numbered from 1 to N (N

being the total number of observations in the LAD).

We then temporarily drop all but one observation per LAD (which is not a problem since the variables we created are identical within each one of them), and use it to draw a scatter plot, which will be the same to the previous one but much easier to compute for your computer. We also add titles for the y and x axis on the graph. When this is all done, we revert to the original dataset.



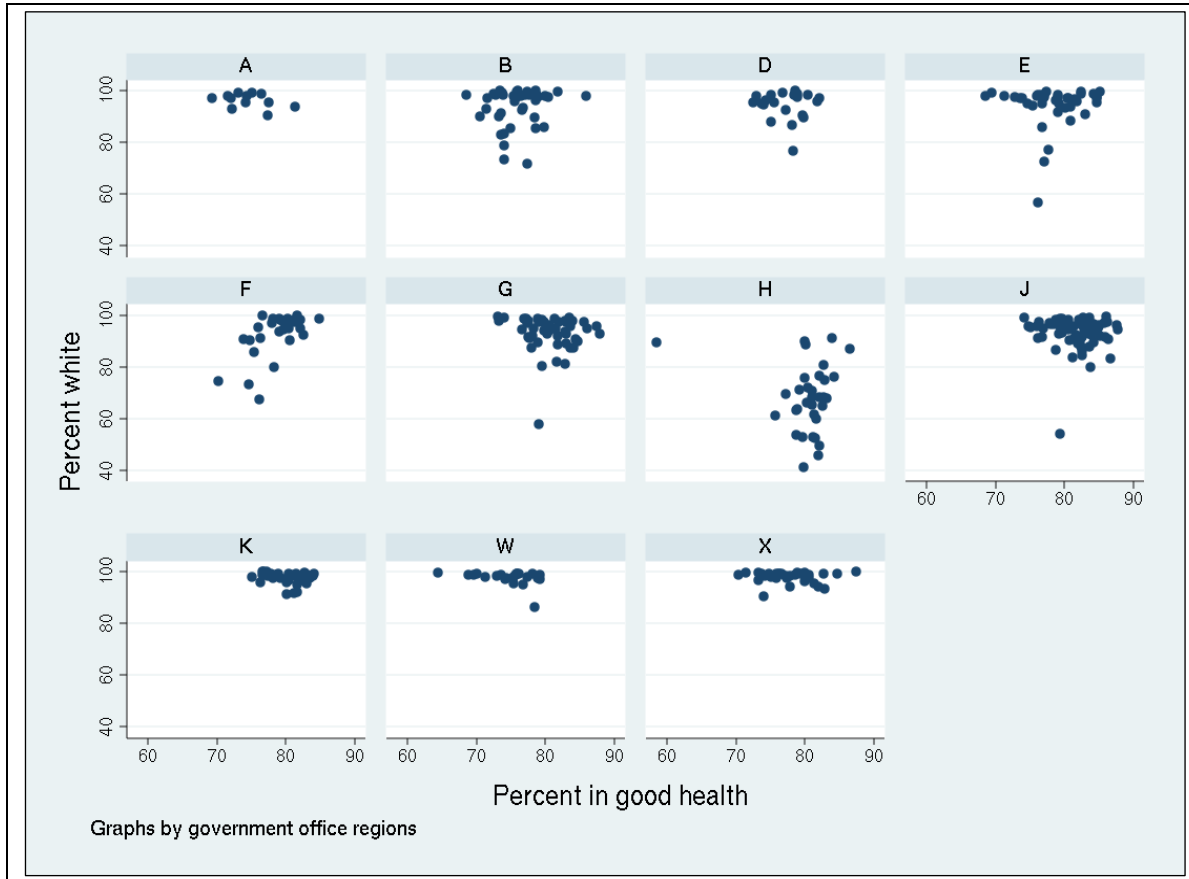
One might find these results a bit disappointing. In order to make sure that we are not missing relationships that might be hidden by the heterogeneity of LADs within regions, we can split the above graph by Government Office Region, and see whether results are identical. This is a straightforward thing to do in Stata.

As above, we create a temporary dataset made of only one observation by LAD. We then ask Stata to produce a graph, with one scatter plot by Government Office Region and country of the UK (the GORA variable). We just need to add a `by` statement, similar to the one we used earlier:

```

preserve
keep if n==1
scatter      p_ethlw p_uqhl12w, by(gora) ///
yttitle("Percent white") xtitle("Percent in good health")
restore

```



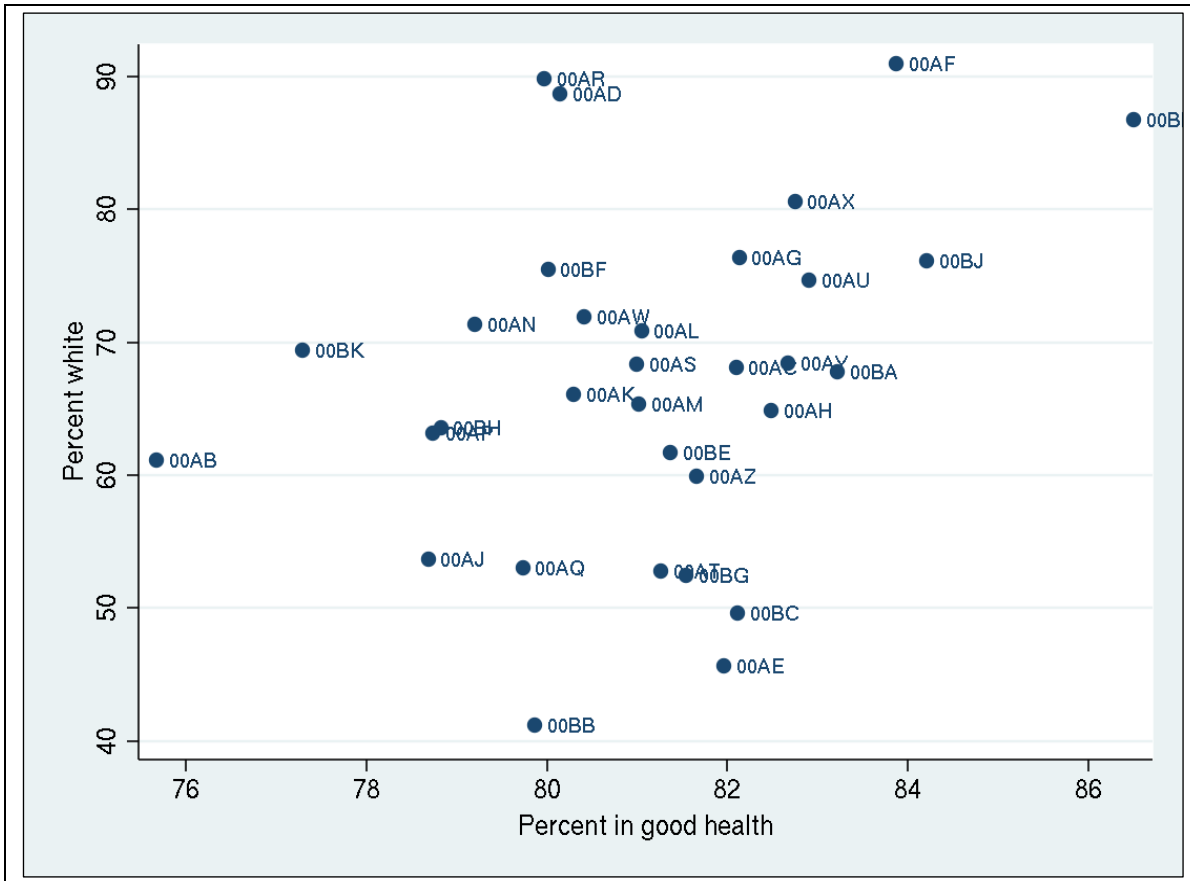
Each letter represents a Government Office Region (the codebook can be found in Volume 3 of the documentation). We can now see that whereas in most of the the Government Office Regions (GORs) the absence of a relationship is confirmed. However in two GORs (the West Midlands and London), a pattern of association has emerged i.e. the higher the proportion of white respondents, the higher the proportion of respondents in good health.

We could even go one step further and produce a scatter plot for only one of these regions, for example London. Given the smaller number of LADs, this time we are able to have them identified on the graph (the `mLabel` option, followed by the name of the variable that contains the label). We will also use a shortcut and remove the outlier that we can notice on the right hand side of the graph (which is the city of London). We do this by specifying two conditions separated by a '&'. H is the value of GORA corresponding to London and '00AA' is the ONS code for the City of London.

```

preserve
keep if n==1
scatter p_ethlw p_uqhl12w if gora=="H" & ualadgb!="00AA", ///
mlabel(ualadgb)///
ytile("Percent white") xtile("Percent in good health")
restore

```



We have now gained a clearer picture of the pattern of association between health and ethnicity in London, at the aggregate level. This is of course the beginning of the analysis, since we need to control for other factors (such as for instance deprivation), to see what lies behind this relationship.