Analysis of data collected by the Labour Force Survey

Which dataset should I use?

Contents

1.What are the Labour Force Survey and the Annual Population Survey ?

What is the Labour Force Survey?

The UK Labour Force Survey (LFS) collects a wealth of information from approximately 39,000 households (or approximately 95,000 individuals) every quarter. Its size and design means that the data collected on the LFS can be analysed in many different ways. This document is a guide to the type of analysis that can be carried out on data collected by the LFS, and what datasets, weights and methods should be used for the different types of analysis available.

This guide should be used alongside existing guidance on the survey. LFS user guides can be found on the following web page:

https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/labourforcesurveyuserguidance

Volume 1 – methodology; a background to the LFS and APS methodology.
Volume 2 - a paper version of the questionnaire used to collect LFS and APS data. Contains information concerning question wording, exact routing to a question and location of a question in the context of the questionnaire.
Volume 3 - contains detail about questionnaire **and** derived variables. Volume 3 should let you know what years and quarters the variables are available and broadly what area they cover.
Volume 4 - shows you how derived variables are calculated using flow charts.
Volume 5 – Classifications such as industry and occupation codes.
Volume 6 – APS user guide (Local Area Data).
Volume 7 - LFS variables 1979-1991; useful to reference when using older datasets.
Volume 8 – A detailed guide on how to conduct household and family analysis.
Volume 9 – LFS variables delivered to Eurostat.
Volume 9a-Eurostat Ad hoc module user guide (*being published soon*)
Volume 10 –Analysis of data collected by the Labour Force Survey: Which dataset should I use?
Volume 11– Longitudinal guide

What is the Annual Population Survey?

The Annual Population Survey (APS) has a larger sample size than the LFS with approximately 320,000 individuals and 150,000 households on its annual datasets. The LFS and APS use the same core questionnaire; therefore a lot of variables are available on both LFS and APS datasets. It is then down to the analyst to decide which dataset best suits their analysis needs. This document aims to help analysts make this decision. However, some questions are only suitable for APS analysis, whilst some are only available for LFS analysis.

2.Types of analysis that can be carried out on data collected on the LFS – an overview

Person level analysis

The LFS is primarily used for conducting person level analysis. For example, the Office for National Statistics (ONS) uses the LFS to publish the number of unemployed people on a monthly basis.

There are two datasets available for person analysis that all serve different purposes.

(i) LFS Person Datasets

PURPOSE: To carry out person-level analysis where you do not want to carry out analysis below regional level. Regional level includes geographical areas such as Wales, Yorkshire and London etc.

FREQUENCY OF PUBLICATION: Quarterly

TIME PERIODS PUBLISHED: January-March (JM), April-June (AJ), July-September (JS) and October-December (OD)

AVAILABILITY: AJ92 to AJ01 (2007 weight)

JS01 to AJ12 (2014 weight)

JS12 onwards (2017 weight)

WEIGHT: PWT**,

PIWT** (earnings weight)

Please note ** denotes the year that the weight was published. For example, the 2017 weight is pwt17.

GENERAL ADVICE: When analysing earnings variables, the earnings weight must be used

LIMITATIONS: Some variables are quarterly specific and therefore only available on certain quarterly datasets. Earnings information is not available for the self employed.

(ii) APS Person Datasets

PURPOSE: To carry out person analysis below regional level (i.e. Local Authority level) or analysis on detailed subgroups

FREQUENCY OF PUBLICATION: Quarterly

TIME PERIODS PUBLISHED: January-December (JD), April-March (AM), July – June (JJ), October – September (OS)

AVAILABILITY: JD04 to JD12 (2014 weight)

A12M onwards (2017 weight)

WEIGHT: PWTA** from A05M. Previous periods have PWAPS** and PWLFS**. (See the LFS user guide volume 6 for more details on how to use these). Please note ** denotes the year that the weight was published. For example, the 2017 weight is pwta17.

GENERAL ADVICE: On the Government versions of the JD files (from 2009), there are some additional variables added to the dataset, including Eurostat ad- hoc modules and wave 1 variables in addition to the weight (EWEIGH**) that should be used for these variables. More information can be found in the Volume 6 LFS user guide.

From April 2011 the APS also contains the personal Well-Being questions (satis, worth, happy, anxious), along with the Well-Being non-proxy weight (NPWT**), which should be used when analysing these variables. Previously (from 2011) a specific 'APS Well-Being micro dataset' was created, however the production of this separate dataset ceased from A14M. The APS person datasets are now the source for the Well-Being variables previously released as part of the 'APS Well-Being micro dataset'

From January 2011 the APS person datasets contains a Sexual Identity variable (SIDV), along with the Sexual Identity weight (SIDWT**), which should be used when analysing this variable. Previously Sexual Identity variables were released as part of the Integrated Household Survey (IHS).

It is important to note that the size of the achieved sample for the Well-Being and Sexual Identity variables within the APS dataset is much smaller than the full APS file. This reduction is due to the Well-Being and Sexual Identity questions only being asked of persons aged 16 and over, who gave a personal interview; proxy answers are not accepted. As a result some caution should be used when analysing responses to Well-Being and Sexual Identity questions at detailed geography areas, or other variables, where unweighted respondent numbers maybe relatively small

LIMITATIONS: Certain LFS variables are not available on APS datasets e.g. LFS quarterly specific variables. Earnings analysis can be carried out on the APS, although care should be taken. See section 'LFS and APS person datasets – the basics' for more detail.

(iii) APS 3 Year Pooled Person Datasets

PURPOSE: To carry out person analysis at lower level geographies and for certain topics whose achieved sample size is smaller

FREQUENCY OF PUBLICATION: Annual

TIME PERIODS PUBLISHED: January to December (covering three years)

AVAILABILITY: JD13 to JD15 (2016 weight)

JD14 to JD16 (2017 weight)

WEIGHT: PWTA**C SIDWT**C and NPWT****C**

GENERAL ADVICE: A 3 year pooled dataset has been produced (with the first period being January13-December15), which will allow more robust analysis at lower level geographies, that isn't always possible when using the single year APS dataset, especially for certain topics whose achieved sample size is smaller. The pooled dataset should be treated solely as point-in-time estimates and not for any time series analysis. More information on the 3 pooled dataset can be found in user guide volume 6.

Household analysis

The LFS is a household survey that tries to collect information for all eligible people in a sampled household. The LFS can therefore be used to conduct household and family level analysis. For example, the ONS uses the LFS to publish analysis on workless households.

There are two datasets available for household/family analysis that all serve different purposes.

(i) LFS Household Datasets

PURPOSE: To carry out household or family analysis where you **do not** want to carry out analysis below regional level. Regional level includes geographical areas such as Wales, Yorkshire and London etc.

FREQUENCY OF PUBLICATION: Biannual from 2004 to 2011

Quarterly AJ12 onwards

TIME PERIODS PUBLISHED: Quarterly from AJ12, Between 2004 and 2011 April-June (AJ) and October-December (OD). Pre 2004, AJ only.

AVAILABILITY: AJ96 to AJ01 (2007 weight)

AJ02 to AJ12 (2014 weight)

JS12 onwards (2017 weight)

WEIGHT: PHHWT** Please note ** denotes the year that the weight was published. For example, the 2017 weight is phhwt17.

GENERAL ADVICE: Generally, when you want a count of the number of households, you need to apply the filter RELHRP6=0.

LIMITATIONS: Please note that household earnings analysis cannot be conducted on LFS household datasets as there is no appropriate weight.

(ii) APS Household Datasets

PURPOSE: To carry out household or family analysis below regional level (e.g. local authority) or to carry out analysis at a detailed level that will yield a small sample size.

FREQUENCY OF PUBLICATION: Annual

TIME PERIODS PUBLISHED: January-December (JD)

AVAILABILITY: JD04 to JD05 (2010 weight)

JD06 to JD11 (2014 weight)

JD12 onwards (2017 weight)

WEIGHT: PHHWTA**Please note ** denotes the year that the weight was published. For example, the 2017 weight is phhwta17.

LIMITATIONS: Please note that some LFS variables are not available on APS datasets, particularly LFS quarterly specific variables. Also earnings analysis can't be done, as there is no appropriate weight.

Note: Up until JD14 (released in 2015), there was an annual dataset produced using APS information, called the Integrated Household Survey (IHS) dataset. The IHS was a composite survey combining questions asked on a number of ONS social surveys, referred to as IHS modules. This resulted in an increased sample size, ideal for primary analysis of sexual identity, perceived general health and smoking prevalence. Due to a variety of factors, all of the additional modules ceased to be included by the end of 2013. This left the concept of the IHS dataset redundant. As a result the APS person data is now the data source for this analysis.

Longitudinal analysis

LFS respondents are interviewed 5 times in total – once every quarter for a year. This allows certain longitudinal analysis to be performed on the data, particularly relating to employment information.

(i) LFS Longitudinal Datasets – 2Q

PURPOSE: For conducting analysis on flows over 3 months (for example, the number of people who have moved from employment to unemployment during a 3 month period).

FREQUENCY OF PUBLICATION: Quarterly

TIME PERIODS PUBLISHED: Q1-Q2, Q2-Q3, Q3-Q4, Q4-Q1

AVAILABILITY: Q197-Q297 to Q101 –Q201 (2007 weight)

Q201-Q301 to Q112-Q212 (2014 weight)

Q212-Q312 (JS12) onwards (2017 weight)

WEIGHT: LGWT**Please note ** denotes the year that the weight was published. For example, the 2017 weight is LGWT17. Previous dataset might just have LGWT.

GENERAL ADVICE: Longitudinal datasets are for person level longitudinal analysis – household longitudinal analysis cannot be carried out using LFS data. The main longitudinal variable is FLOW (labour force gross flow over the 3 month period).

LIMITATIONS: Please note that a **subset** of LFS variables is available on longitudinal datasets. Longitudinal datasets only include respondents of working age who have responded in both quarters; attrition between waves means that the sample size is smaller compared with a single quarterly dataset.

(ii) LFS Longitudinal Datasets – 5Q

PURPOSE: For conducting analysis on flows over a year (for example, the number of people who have moved from employment to unemployment during a 12 month period).

FREQUENCY OF PUBLICATION: Quarterly

TIME PERIODS PUBLISHED: Q1-Q1, Q2-Q2, Q3-Q3, Q4-Q4

AVAILABILITY: Q197-Q198 to Q201-Q202 (2007 weight)

Q301-Q302 to Q210-Q211 (AJ11)(2010 weight)

Q310-Q311(JS11) to Q211 to Q212 (AJ12) (2014 weight)

Q311 –Q312 (JS12) onwards (2017 weight)

WEIGHT: LGWT**

Please note ** denotes the year that the weight was published. For example, the 2017 weight is LGWT17. Previous dataset might just have LGWT.

GENERAL ADVICE: Longitudinal datasets are for person level longitudinal analysis – household longitudinal analysis cannot be carried out using LFS data. The variables FLOW (labour force gross flow over a 12 month period) and ANFLOW (labour force gross flows across all five quarters) are available on the datasets. They give the categories relating to labour force gross flows in a convenient form.

LIMITATIONS: Please note that a **subset** of LFS variables is available on longitudinal datasets. Longitudinal datasets only include respondents of working age. Only those respondents who respond in every one of the five waves are included in this dataset; attrition between wave's means that the sample size is very small compared with a single quarterly dataset.

3.The differences between the Labour Force Survey andthe Annual Population Survey

There are two types of LFS cases:

(a) LFS main cases. These cases are interviewed quarterly for 5 consecutive quarters.

Wave 1	Wave 2	Wave 3	Wave 4	**Wave 5**
January 1st 2016	April 1st 2016	July 1st 2016	October 1st 2016	**January 1st 2017**
First interview	Second interview	Third interview	Fourth interview	**Final interview**

Only LFS main cases are used for the LFS datasets (household, person and longitudinal).

(b) LFS boost cases. These cases are interviewed annually for 4 consecutive years.

Wave 1	**Wave 2**	**Wave 3**	**Wave 4**
January 1st 2014	**January 1st 2015**	**January 1st 2016**	**January 1st 2017**
First interview	Second interview	Third interview	Fourth interview

On the LFS dataset for a given quarter, you will have cases from the 5 different waves, as illustrated in the columns in the table below:

	Quarter			
	JM17	AJ17	JS17	OD17
LFS cohort 1 (first sampled JM16)	**Wave 5** **(fifth interview)**			
LFS cohort 2 (first sampled AJ16)	Wave 4 (fourth interview)	**Wave 5**		
LFS cohort 3 (first sampled JS16)	Wave 3 (third interview)	Wave 4	**Wave 5**	

LFS cohort 4 (first sampled OD16)	Wave 2 (second interview)	Wave 3	Wave 4	**Wave 5**
LFS cohort 5 (first sampled JM17)	**Wave 1** **(first interview)**	Wave 2	Wave 3	Wave 4
LFS cohort 6 (first sampled AJ17)		**Wave 1**	Wave 2	Wave 3
LFS cohort 7 (first sampled JS17)			**Wave 1**	Wave 2
LFS cohort 8 (first sampled OD17)				**Wave 1**

The wave 1 and wave 5 cases, highlighted are known as the wave 1 and 5 main LFS cases, which will feed into the Annnual Population Survey (APS).

The APS is constructed by bringing together waves 1 and 5 **main** LFS cases from each quarter in the year and all the boost respondents from the **same** adjacent quarters into one data set.

For JD16 the boost cases would be the following:

	In JD16
LLFS cohort 1 (first sampled Jan-Dec 2013)	Wave 4 (fourth interview)
LLFS cohort 2 (first sampled Jan-Dec 2014)	Wave 3 (third interview)
LLFS cohort 3 (first sampled Jan-Dec 2015)	Wave 2 (second interview)
LLFS cohort 4 (first sampled Jan-Dec 2016)	Wave 1 (first interview)

Therefore the APS annual dataset will look like this:

Waves 1 and 5 of LFS Main	+	All waves of LLFS	=	APS Dataset
LFS wave 1 LFS wave 2 LFS wave 3 LFS wave 4 LFS wave 5		LLFS wave 1 LLFS wave 2 LLFS wave 3 LLFS wave 4		LFS wave 1 LLFS wave 1 LLFS wave 2 LLFS wave 3 LLFS wave 4 LFS wave 5

This creates a large dataset, which allows for Local Authority level analysis to be carried out.

4.LFS and APS Person Level Datasets

The LFS and APS person level datasets are an important source of information for employment related person level statistics. They provide an analyst with the opportunity to analyse employment information alongside many other topics such as education and health.

The main difference between the LFS and APS person datasets is the sample size. The APS has a much bigger sample due to its design (see previous section-section 3) which allows for local authority analysis or more detailed analysis to be carried out on the APS. As mentioned above, the APS has roughly three times the number of individuals on a dataset when compared to a LFS quarterly dataset. Also the LFS is used for analysis within individual calendar quarters; whereas APS is used for analysis within a calendar year

However, the APS has its limitations. Not all LFS questions are available on APS datasets. If a question is not asked of wave 1 and 5 main LFS responders and all LFS boost responders, they are not available on APS datasets. This is due to the design of the APS (see section 3).

For example, if a LFS variable is asked in AJ quarters of main LFS responders only (i.e. not asked of boost responders and not asked in the other quarters) this variable would not be on the APS dataset.

The APS well being datasets should only be used to analyse the personal well being questions. They have their own separate datatset due to the way the weight is derived. Any other APS analysis should be carried out on the normal APS person datasets.

LFS and APS Person Datasets – the basics

Availability

LFS Person datasets are available on a **calendar quarter** basis:

LFS Dataset	**Release month**
January – March (JM)	May

April – June (AJ)	August
July – September (JS)	November
October – December (OD)	February

APS Person datasets are made available every quarter;but they are **rolling annual** datasets:

APS Dataset	**Release month**
January – December (JD)	March
April – March (AM)	June
July – June (JJ)	September
October - September (OS)	December

Missing Values

Both LFS and APS **person** datasets have two types of missing values.

(a) -9s denote that the respondent is not applicable for the variable
(b) -8s denote that the respondent doesn't know or has refused information for a particular variable

Non-responders and imputation

- LFS Person Datasets and imputation

For LFS person datasets, the only form of imputation is rolling forward data.

If a responder responds in one wave but in the consecutive next wave chooses not to respond, their data is rolled forward from the previous wave. When a responder has rolled forward data, their IOUTCOME value is generally set equal to 6. Responders who **had an interview last wave** and were coded as **economically inactive** (IOUTCOME=7) but in the **current wave decide not to respond** also have their data rolled forward, but their IOUTCOME value is 7 (economically inactive).

Note that data is only rolled forward for **one wave** and it is only rolled forward to the next **consecutive** wave.

If a responder doesn't respond for two consecutive waves they become a non responder (IOUTCOME=3) and are **not included** on LFS person datasets.

Consider the below case study.

Respondent A and Respondent B live together in a household that is chosen to take part in the main LFS. Respondent A says they are a baker and Respondent B is a gardener.

In wave one both Respondent A and Respondent B are responders. They both appear on the person dataset.

In wave two, Respondent A does not respond whilst Respondent B does. Respondent A's information is brought forward from the wave before and therefore remains on the person dataset alongside Respondent B i.e. we assume that Respondent A is still a baker.

In wave three, Respondent A does not respond for the second time whilst Respondent B does. Respondent A's information **is not brought forward for a second consecutive time** and **Respondent A will not appear on the person dataset.** Respondent B will remain on the LFS person dataset.

In wave four, neither Respondent A nor Respondent B responds. This time, Respondent B's information is brought forward from the wave before and remains on the dataset (i.e. we assume Respondent B is still a gardener). Respondent A remains off the dataset.

In wave five, Respondent B does not respond again - their information is not brought forward for a second consecutive time and they do not appear on the dataset. However, Respondent A does respond and tells the interviewer they are no longer a baker, they are now a teacher. Responder A will therefore appear on a dataset with this new information.

	Respondent A	**Respondent B**
Wave 1	Responder – on person dataset	Responder – on person dataset
Wave 2	Non responder for first time – data rolled forward	Responder – on person dataset
Wave 3	Non responder in 2nd consecutive wave – not included on person dataset	Responder – on person dataset
Wave 4	Non responder in 3rd consecutive wave – not included on person dataset	Non responder for first time – data rolled forward
Wave 5	Responder – on person dataset	Non responder in 2nd consecutive wave – not included on person dataset

- Imputation on APS Person Datasets

There is very little emphasis on imputation on APS person datasets. This is because data is not rolled forward on LFS boost cases. The only imputed cases on an APS dataset will be certain wave 5 main LFS cases (where their data has been rolled forward from the wave 4 interview).

Weights

Dataset	Weight
LFS	PWT** PIWT** (earnings weight)
APS	PWTA** (from A05M for JD04 to JD05 PWAPS** and PWLFS** - more details can be found in user guide volume 6)

Earnings Analysis

When carrying out LFS analysis on earnings variables, the weight PIWT must be used. This weight makes up for the fact that the earnings questions are only asked to two-fifths of the sample and also compensates for the higher levels of non response.

You can carry out earnings analysis on the APS although there is no earnings weight. It is recommended that when using the APS, calculating median earnings is the most reliable measure as it is similar to those obtained from the LFS using income weights. If you want to carry out analysis on levels of earnings using the APS datasets, it is recommended that you apply a scaling factor to your analysis to compensate for the lack of non-response adjustment in the person weights.

The scaling factor is calculated by weighting the 4 quarter average LFS by the income weight and then dividing this by the weighted total of the APS (answering the earning questions). The scaling factor can be produced overall or at an individual level, depending on what scaling adjustment you choice will have an impact on your estimate, especially for groups where the sample size is small.

Note that the LFS/APS measures **employees' earnings.** The LFS/APS does not ask any earnings questions to the self-employed and does not collect information on money coming from any other source apart from wages; therefore the LFS/APS cannot be used to measure **income.**

See Example 1 in the annex for an example of basic earnings analysis on the LFS

LFS and APS weighting method

The LFS and APS reflect only a sample of the total population. All cases are therefore weighted on the basis of sub-national population totals by age and sex to give estimates for the entire UK household population.

This weight shows how many people in the population they are representing on that dataset. When carrying out analysis, you should always apply the weight.

A “calibration” weighting method is used. This is an iterative algorithm designed to produce individuals’ weights that are consistent with three sets of population totals, or “partitions”. These partitions are:

• Local authority totals for people aged 16+, by gender;

• Great Britain and Northern Ireland totals by gender and by single year of age for 16-24s and totals for 0-15 and 25+; and

• Regional totals by quinary age bands and by gender.

A more detailed explanation of the weighting method can be found in the LFS User Guide Volume 1.

When to use person level datasets

Person level data should be used for person level analysis. Person level datasets should not be used for household or family analysis, because some household/family members may be missing on the person datasets (see non-responders and imputation above). This is not an issue on the household datasets as non-responders are included.

The main advantage of the APS over the LFS is the sample size. If you are looking at a very specific sub-group of the population that will yield a small sample size, an APS dataset would be a better alternative to the LFS. As mentioned above, only the APS datasets should be used for analysis below regional level (such as Local Authority analysis).

Certain variables are only available to analyse on the LFS datasets, in particular variables that are only asked in certain quarters. The LFS user guides give more detail.

Person variables

There are two main types of person variables – questionnaire variables and derived variables.

Questionnaire variables are what their name suggests – they are the questions that are asked to respondents during their questionnaire interview.

Derived variables are calculated using a variety of questionnaire variables after the data has been collected.

Details of all LFS variables can be found in the LFS User Guides:

Volume 2 is a paper version of the questionnaire and is useful if you want to see how a question is worded, or where the question is placed in the questionnaire.

Volume 3 contains both questionnaire and derived variables. It provides a broad description of the variable and information such as when it was introduced to the LFS and whether it is a derived variable.

Volume 4 shows you how the derived variables are calculated

Here are a few of the key person level variables:

VARIABLE	DESCRIPTION
CASENO	Unique person identifier Not available on the EUL Can be used to link people across periods
CASENOP	Pseudonymous unique person identifier Available on EUL Cannot be used to link people across periods
IOUTCOME	Person outcome code – shows whether the interview was a proxy, data was brought forward etc.
INECAC05	Economic Activity
ILODEFR	Economic Activity (high level)

Every person will have a unique identifier (CASENO) or a pseudonymous equivalent (CASENOP).

4 quarter average analysis on LFS person datasets

In general, it is not recommended to calculate annual averages of a variable by calculating a 4 quarter average. This is because of the wave structure of the LFS. Consider the below table:

	Q1	Q2	Q3	Q4
Responder A	Wave 1	Wave 2	Wave 3	Wave 4
Responder B	Wave 2	Wave 3	Wave 4	Wave 5
Responder C	Wave 3	Wave 4	Wave 5	
Responder D	Wave 4	Wave 5		
Responder E	Wave 5			
Responder F		Wave 1	Wave 2	Wave 3
Responder G			Wave 1	Wave 2
Responder H				Wave 1

When calculating 4 quarter averages, you are double, triple or quadruple counting some responders over others and therefore giving their responses a disproportionate weight.

You also have to consider the variance of your estimates. As the answers from one wave to a next are correlated (it's quite likely respondents answers are similar or related) the calculation of the variance becomes very complicated and you cannot use the method of adding up the individual quarterly variances and dividing by 4. This is also the case with the confidence intervals.

It is therefore recommended to use APS datasets to calculate the annual average of a variable whenever possible.

5.LFS and APS Household analysis

The LFS is a household survey. This can initially seem confusing as the Labour Market Division within the Office for National Statistics (ONS) uses the LFS to publish the number of unemployed people (not households!) on a monthly basis and ONS also publishes person datasets on a quarterly basis. So what makes the LFS a household survey?

It's all down to the way we sample. We do not randomly choose people to take part in the survey – we randomly choose **households** (more specifically private residential households).

We return to **households** every wave – we do not necessarily go back to people.

For example, say an interviewer visits the household at address Z in wave one. They interview everyone in the household – let's call them the Mr and Mrs X.

When the interviewer returns to address Z three months later, Mr and Mrs X have moved house – and now Miss Y lives at this address. If the interviewer does not find out where Mr and Mrs X have moved and then go and interview them in their new address – the X family are no longer of interest to the interviewer. Instead, the interviewer interviews Miss Y instead.

It is this that makes the LFS a household survey.

As we sample households, and try to collect information about everyone who is eligible for the survey in that household, we are able to conduct household and family level analysis.

LFS and APS Household datasets – the basics

Availability

LFS Household datasets are available for all quarters from JS12. AJ household datasets are available from 1996, whilst AJ and OD datasets are available from 2004. In addition to LFS household datasets, the ONS releases Annual Population Survey (APS) household datasets.

The main difference between the LFS and APS household datasets is the sample size – APS household datasets are approximately three times bigger than LFS household datasets.

APS household datasets should be used if you want to carry out household analysis at a geographical level **lower** than Government Office Region (e.g. local area analysis) or more detailed household analysis that would yield a small sample size.

They are available once a year since 2004 and cover a January-December time frame (e.g. JD12, JD13, JD14, JD15, JD16 etc).

APS datasets will not contain all LFS variables – this is because not all LFS variables are APS variables. See section 3 'The differences between the LFS, APS and IHS' for more detail.

LFS Household Datasets	Release Month
January –March (JM)	May
April-June (AJ)	August
July-September (JS)	November
October-December (OD)	February

APS Household Datasets	Release Month
January-December (JD)	Currently September (but under review).

Missing Values

Before AJ12, all missing values on household datasets are denoted by a -10.

-10s do not distinguish between missing data and don't knows/ refusals. This is different to person level datasets where there are -9s (not applicable) and -8s (don't know/refusal).

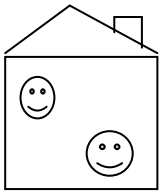
For AJ12 onwards, -10s on household datasets are replaced by -9s and -8s.

For most of the person-level variables on the household datasets, the values are not imputed, so there will be more missing values observed than in the person-level dataset. Therefore users need to take care when dealing with these in their analysis.

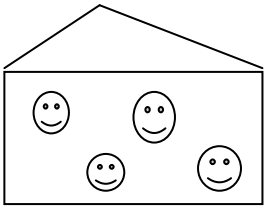
Weights

Household datasets have household weights. The household weight variable on the LFS always starts with PHHWT followed by two numbers that indicate the year this weight was created. For the APS the household weight starts with PHHWT but then has an 'A' before the last two numbers.. Every person in a household has the same household weight.

For example, consider the below households.



Household A



Household B

The couple in household A will both have the same value for their household weight. However, their person weights (which are on the LFS and APS person datasets) will probably be different from each other.

The couple and their children in Household B will all have the same value for their household weight. However, their person weights will probably be different.

This applies to the LFS and APS.

Non-responders

It is important to know that **non-responders (IOUTCOME=3) are included on LFS and APS households datasets and will have a weight.**

For example, let's consider Household A above. Member A of that household did not respond, whilst Member B did. Member A will still have a household weight (the same household weight as Member B). However, Member A will not appear on the **person** dataset as they are a non-responder. The reason for inclusion on household datasets is that it is useful for data analysts to gain a profile of who is in the household e.g. they might want to conduct analysis among all households where a male aged 21-24 is present.

Imputation

The household datasets aim to provide more robust household and family labour market statistics.

To ensure the economic status of all individuals within a household is known, a method of 'donor' imputation takes place for those with a 'missing' economic activity status allowing for analysis of the combined economic status of households.

Following a review of the imputation methods, it was decided that it wasn't appropriate to impute any personal characteristic variables (e.g. Religion, Ethnicity, Country of Birth, Nationality, National Identity etc), from JM15, see user guide 1 section 13.4 for more details.

When to use

All household analysis should be carried out on household datasets (even if some household variables are on person datasets).

All family analysis should be carried out on household datasets.

Similarly, although household files are at person level and contain the majority of person variables, person analysis should be carried out on person datasets. This is because there is not a person weight on the household datasets.

If you want to carry out a set of analysis which is part person level analysis, part family/household analysis and you want consistency between the two sets of analysis, household datasets should be used.

Geographical analysis

LFS household datasets should be used if you want to carry out UK or regional level household analysis (e.g. Wales, Yorkshire, London etc.) or analysis that will not yield a small sample size. If you want to carry out household analysis at a geographical area lower than region, or for very detailed analysis, LFS household datasets should not be used as the sample size becomes too small.

Therefore, if you want to carry out analysis at geographical levels such as Local Authority level, APS Household datasets should be used.

If you plan to carry out household analysis on a sub-group of the population that you know will yield a small sample size, you should once again consider using the APS household dataset.

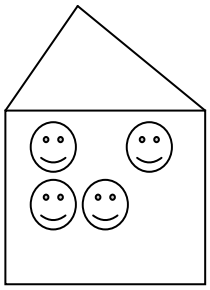
The difference between households and families on LFS and APS datasets

Household analysis and family analysis are two very different things.

This is because there can be more than one family in one household.

The variable TOTFU tells you the number of families in a household.

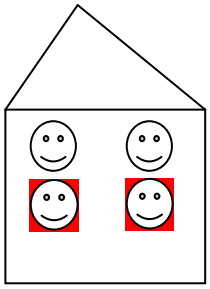
Consider the below Household A; a married couple and their 2 children (never married) live there:



This is one household and one family. TOTFU=1.

Everyone in this household will have the same value for household variables (such as HPNTIME) **and will have the same value for family variables** (such as FUTYPE6).

Now consider the below Household B; a couple and their lodgers – a mother and her child:



This is **one** household but **two** families. TOTFU=2.

Everyone in this household will have the **same value for household variables** (such as HPNTIME) and will have **different values for family variables** (such as FUTYPE6).

This is how these three families within the 2 households would look on a dataset:

HSERIAL	FUSERIAL	CASENO	PHHWT**	RELHRP6	RELHFU	HPNTIME	FUTYPE6	TOTFU	INECAC05
11111	9876	1111101	600	0	1	2	6	1	2
11111	9876	1111102	600	1	2	2	6	1	4
11111	9876	1111103	600	3	3	2	6	1	34
11111	9876	1111104	600	3	3	2	6	1	34
22222	7651	2222201	950	0	1	0	4	2	31
22222	7651	2222202	950	1	1	0	4	2	34
22222	7652	2222203	950	19	1	0	12	2	1
22222	7652	2222204	950	19	3	0	12	2	34

Notice that there are two values for HSERIAL and PHHWT** as there are two households, but there are 3 values for FUSERIAL. This is because within the second household, there are two values for FUSERIAL as there are two families living there.

See Example 2 (annex) for how applying the wrong filter can affect your analysis.

See Example 3 (annex) for how choosing the wrong type of variable can affect your analysis.

Household variables

The majority of household variables are derived (calculated using questionnaire variables after the data has been collected).

The volume 8 LFS user guide is useful if you want more information about household and family analysis.

As the majority of household variables are derived, volumes 3 and 4 of the LFS User Guides should be used if you are interested in household variables.

User Guide 4 will show you how the variable is calculated, whilst volume 3 gives you a broad description. If the information in User Guide 3 is not what you require, it is worth referencing User Guide 4.

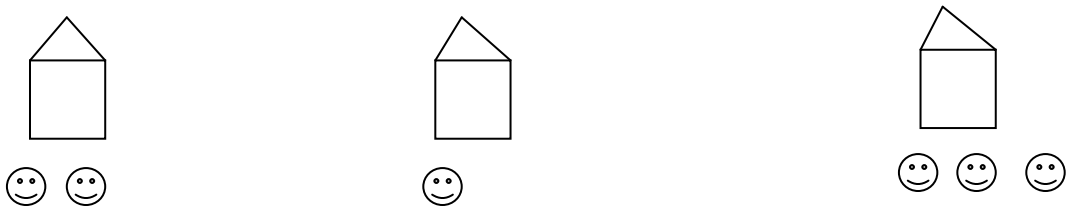
Here are a few of the key **LFS and APS** household variables:

VARIABLE	DESCRIPTION
HSERIAL	Unique household identifier Not available on the EUL Can be used to link households across periods
HSERIALP	Pseudonymous unique household identifier Available on EUL Cannot be used to link households across periods
HHTYPE6	Type of household Can tell you things such as how many married couple households there are etc.
HRP	Household reference person.
RELHRP6	Relationship to household reference person
TOTNUM	Total number of eligible people in a household

Every household will have a unique identifier (HSERIAL) or a pseudonymous equivalent (HSERIALP).

Imagine a row of houses with their residents:

Household 1 Household 2 Household 3



HSERIAL: 1234 HSERIAL: 5678 HSERIAL: 9101

Every one of these households has a different HSERIAL.

Everyone in Household 1 will have the same HSERIAL (1234) – it doesn't matter how the residents are related to each other.

Everyone in Household 2 will have the same HSERIAL (5678).

And…everyone in Household 3 will have the same HSERIAL (9101).

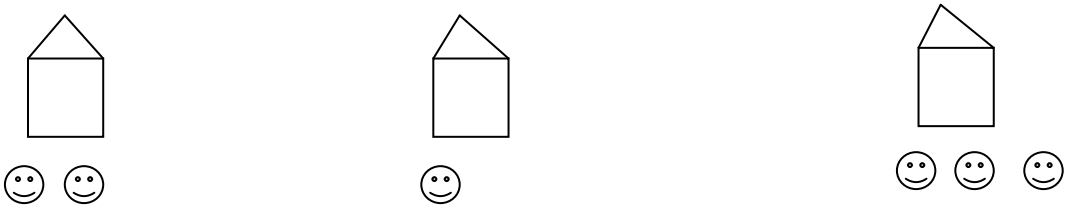
Conducting household analysis on APS and LFS household datasets

All household analysis needs to be performed on household datasets.

Everyone in a household will have the same value for most household variables (excluding RELHRP6, relationship to the HRP).

If we consider the same row of houses again:

Household 1 Household 2 Household 3



HSERIAL: 1234 HSERIAL: 5678 HSERIAL: 9101

Everyone in Household 1 will have the same value for household derived variables such as HHTYPE6.

Everyone in Household 2 will have the same value for household derived variables such as HHTYPE6.

And…everyone in household 3 will have the same value for household derived variables such as HHTYPE6.

Household files are at person level (each row represents a person) therefore if you want to carry out household analysis, you have to be very careful.

Generally, when you want to count the number of households you need to apply the filter **RELHRP6=0.**

This leaves only the household reference people (HRP) on the datasets which is the same as reducing the dataset to household level. After this filter has been applied, the HRP is representing the entire household. This is fine for conducting analysis on household variables.

For example, say I have 3 households on a dataset with 5 people in each household. You want to find out how many how many **households** have part time workers living there.

HNPTIME is a household variable that tells you how many people in the household are part time.

CASENO	**RELHRP6**	**HNPTIME**
1234561	0	5
1234562	1	5
1234563	2	5
1234564	3	5
1234565	4	5
6543211	0	0
6543212	1	0
6543213	2	0
6543214	3	0
6543215	4	0

1357911	0	1
1357912	1	1
1357913	2	1
1357914	3	1
1357915	4	1

In household one, every person in that household is part time i.e. there are 5 part time workers in that household. Therefore, everyone in that household gets a value of 5 for HNPTIME.

If you run a frequency on HNPTIME with no filter, you will get the following result:

HNPTIME

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	5	33.3	33.3	33.3
	1	5	33.3	33.3	66.7
	5	5	33.3	33.3	100.0
	Total	15	100.0	100.0	

This does not mean that 5 households have no part time workers (after all we only had 3 households to begin with)!

You need to apply the filter RELHRP6=0 first, and then run your frequency.

The filter reduces the dataset to the following:

CASENO	**RELHRP6**	**HNPTIME**
1234561	0	5
6543211	0	0
1357911	0	1

You will then get the following result. This correctly tells you that there is one household that has no part time workers.

HNPTIME

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1	33.3	33.3	33.3
	1	1	33.3	33.3	66.7
	5	1	33.3	33.3	100.0
	Total	3	100.0	100.0	

Family variables

Household datasets also contain family variables.

Everyone in a family will have the same value for all family variables.

The majority of family variables are derived variables, so User Guide volume 3 and 4 will be of most use.

Here are a few of the high profile family variables:

VARIABLE	DESCRIPTION
FUSERIAL	Unique family identifier
FUTYPE6	Type of family unit
HEAHEAD	Economic activity of head of family unit
RELHFU	Relationship to head of family unit
TOTFU	Number of family units in a household

Every family unit has a unique identifier (FUSERIAL).

Consider the below 3 families:

Family 1

FUSERIAL: 4321

Family 2

FUSERIAL: 5432

Family 3

FUSERIAL: 6543

As they are 3 different families, there are 3 different FUSERIALs.

Everyone in Family 1 will have the same FUSERIAL and so on.

Conducting family analysis on datasets

All family analysis needs to be performed on household datasets.

Generally, when you want to count the number of families, you need to apply the filter **RELHFU=1.**

This leaves only the head of the families on the dataset which is the same as reducing the dataset to family level.

After this filter has been applied, the head of the family is representing the entire family.

For example, say I have 5 families on a dataset with 3 people in each family.

You want to find out how many how many lone parent families there are.

If FUTYPE6=10/12 this tells us the family is a lone parent family with dependent children.

CASENO	RELHFU	FUTYPE6
1234561	1	10
1234562	3	10
1234563	3	10
4567891	1	12
4567892	3	12
4567893	3	12
7654321	1	8
7654322	2	8
7654323	3	8
1987651	1	5
1987652	2	5
1987653	3	5
3579131	1	3
3579132	2	3
3579133	3	3

Household one is a lone parent family; therefore the parent and the two children are allocated with the value FUTYPE6=10.

All members of the family will be assigned the same family type – including the children. If you run a frequency on FUTYPE6 with no filter, you will get the following result:

FUTYPE6

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3	3	20.0	20.0	20.0
	5	3	20.0	20.0	40.0
	8	3	20.0	20.0	60.0
	10	**3**	20.0	20.0	80.0
	12	**3**	20.0	20.0	100.0
	Total	**15**	100.0	100.0	

This does **not** tell us that there are 6 lone parent families (3 FUTYPE6=10 and 3 FUTYPE6=12). Remember – we only had 5 families to begin with. This tells us that on our dataset, there are 6 i**ndividuals** who **belong** to lone parent families.

In order to find out the **number of lone parent families**, you need to apply the filter RELHFU=1 first, and then run your frequency.

The filter reduces the dataset to the following:

CASENO	RELHFU	FUTYPE6
1234561	1	10
4567891	1	12
7654321	1	8
1987651	1	5
3579131	1	3

You will then get the following result:

FUTYPE6

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3	1	20.0	20.0	20.0
	5	1	20.0	20.0	40.0
	8	1	20.0	20.0	60.0
	10	**1**	20.0	20.0	80.0
	12	**1**	20.0	20.0	100.0
	Total	**5**	100.0	100.0	

This correctly tells us that there are two lone parent families (1 FUTYPE6=10 and 1 FUTYPE6=12).

Creating your own household variables

You may need to create your own household variable. For example, you may want to know how many households contain a person who has a higher degree.

There isn't a derived variable that tells you this, and it's definitely household analysis, so in cases like this you will have to create your own household variable.

Step one – Choose your dataset and make sure HSERIAL or HSERIALP is on this dataset. As this is household analysis, we will want to choose a household dataset.

Step two – find an appropriate (person level) variable that gives you the type of information you need. In our cases, HIQUAL15=1 tell us a person has a higher level degree.

Step three – convert this variable into a household level variable. This will take a couple of steps:

(a) RECODE the variable you chose in step 2 into a **new variable** (we don't want to alter the existing variable). Let's call our new variable HDEG (higher degree).

Set your new variable to 1 for all people you are interested in. In our case, we are interested in people who have a higher degree (HIQUAL15=1), otherwise set you new variable to 0. This can be done using the RECODE command in SPSS:

RECODE HIQUAL15 (1=1) (ELSE=0) INTO HDEG.

It's worth checking that the above has worked by running a crosstab of your new variable and your original variable.

(b) We need to add up your new variable over everyone in the household which creates a new household variable (let's call it HOUSEHDEG). Consider the below dataset that has three household in it:

HSERIAL	**HIQUAL11**	**HDEG**	**HOUSEHDEG**
123123	1	1	1+0+0=1
123123	5	0	1+0+0=1
123123	9	0	1+0+0=1
234567	24	0	0+0=0
234567	24	0	0+0=0
789012	1	1	1+1=2
789012	1	1	1+1=2

If anyone in the household has a higher degree, then HOUSEHDEG will be greater than 0. If HOUSEHDEG=0, then it tells us that nobody in that household has a higher degree.

We can create this household variable using the AGGREGATE function in SPSS:

```
AGGREGATE
/OUTFILE='AGG DATASET'
/BREAK HSERIAL
/HOUSEHDEG=TOT(HDEG).
```

The way this function works is that it creates your new household variable, alongside with HSERIAL (or HSERIALP), on a new dataset. You now need to merge your new household variable back on to your original dataset. This can be done using the MATCHFILE command in SPSS:

```
GET FILE='ORIGINAL DATASET'.
MATCH FILES /FILE=*
/TABLE='AGG DATASET'
/BY HSERIAL.
EXECUTE.
```

You will need to make sure that both datasets used in the MATCH FILES command are sorted by HSERIAL/HSERIALP before you carry out the command.

In SAS the following code can be run:

```
data originaldataset1;
set originaldataset;
HDEG=0;
if HIQUAL15=1 then HDEG=1;
run;

proc summary data=originaldataset1;
var HDEG;
class HSERIAL;
output out= aggdataset sum=HOUSEHDEG;
run;

proc sort data= originaldataset1; by hserial; run;
proc sort data=aggdataset; by hserial;run;

data mergeddataset;
merge originaldataset1 aggdataset;
by hserial;
run;
```

You should now have a dataset with your new household derived variable attached. Now this has been done, you can carry out household analysis in a normal way.

Step four – Now we have a suitable household variable, we can apply an appropriate filter.

We currently have a dataset like the below:

HSERIAL	HOUSEHDEG
123123	1
123123	1
123123	1
234567	0
234567	0
789012	0
789012	0

Everyone in the household has been assigned a value for HOUSEHDEG.

We therefore need to apply the filter RELHRP6=0. This will reduce the dataset to household level:

HSERIAL	HOUSEHDEG
123123	1
234567	0
789012	0

Step five – run your analysis (and remember to apply an appropriate weight!)

See Example 4 (annex) for another example of creating your own household variable.

See Example 5 (annex) for an example of calculating person level statistics broken down by the characteristics of the household.

Analysis on questionnaire variables that are only asked to one person in the household

Most of the household variables that have been discussed so far are derived variables. However, there are a few questions on the LFS that are only asked about the household reference person (HRP). These are:

TEN1, TIED, LLORD, FURN, NRMS2

These questions focus on the household, not the individual. For example, they ask things such as how many rooms there are in the house, and whether the accommodation is rented. Although they are only asked about the HRP, the answer supplied is then applied to everyone else in the household – including children. More details on these variables can be found in Volume 2 user guide.

For example, say there is a 4 person household. The HRP answers the question TEN1, saying that they rent the household. On the household (and person) datasets, everyone in that household will then have a value of TEN1=4 (although 3 of the respondents have not been asked the question).

There are two types of analysis that can be done on these variables.

(a) Person level analysis

You might want to know the number of people in UK who are living in rented accommodation. You can use the LFS person files for this analysis and run a frequency on TEN1.

For JS17, the result you get is 28,098 people live in rented accommodation (unweighted).

Using the person weight pwt17, this becomes 21,876,557 people.

(b) Household level analysis

However, you might want to know how many houses are rented in the UK. In this case you will need to use a household file – but you can't just run a frequency on TEN1. As mentioned above, the household file is not at household level, so you need to apply a filter before running this type of analysis.

In order to make the file household level, you simply apply the filter RELHRP6=0. You can then run your analysis on TEN1.

In JS17, this gives a result of 12,165 houses are rented in the UK (unweighted).Using the household weight phhwt17, this becomes 9,256,432 households. This is quite different to the number we get in part (a)!

ONS releases that use the LFS/APS household datasets

ONS release information each year and fourth quarter (mid year) about working and workless households. This release contains information about households and the adults and children living in them, by their economic activity status. http://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/workingandworklesshouseholds/previousReleases

There is also an annual families and households release, which presents estimates of families by type, including married and cohabiting couple families and lone parents, as well as tables on household size and household types. http://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/previousReleases

It's always worth looking at these releases to see if they contain some of the information you are interested in.

6. LFS Longitudinal Analysis

Overview

LFS longitudinal datasets allow analysts to look at **change over time.**

There are two types of longitudinal datasets.

- 2Q datasets. These datasets link together two consecutive calendar quarters such as OD16 and JM17. This enables analysts to look at change over 3 months.
- 5Q datasets. These datasets link together five consecutive quarters such as, JM16, AJ16, JS16, OD16, JM17. This enables analysts to look at change over a year.

Cases on the files

Longitudinal datasets focus on employment related variables and therefore only contain a sub-set of LFS variables. Longitudinal files only include respondents of working age who haven't had their information imputed in any quarters. Working age is classed as those aged 15-69 in the first quarter (and 15-70 in the last quarter of the 2Q dataset and 16-70 in the last quarter of the 5Q dataset). Note before OD01 working age is restricted to 15-59 for women and 16-64 for men in the first quarter.

LFS longitudinal files contain LFS main cases only (not boost cases).

Those who have no information for their economic status in any of the periods covered in a longitudinal dataset are excluded from the longitudinal file.

There are no brought forward cases or non-responders on longitudinal files for the 2Q datasets, cases on the file responded in both the quarters, and for the 5Q datasets, cases on the file responded in all five quarters.

For example, consider a LFS main responder whose first interview was in JM16. They respond every quarter apart from AJ16. They are therefore not included in both the 2Q and 5Q longitudinal datasets as their data would have been rolled forward in AJ16.

Longitudinal variables

The main variable on both the longitudinal datasets is FLOW. FLOW looks at labour force gross flows and summarises the main movements in employment between the periods.

Its categories are as follows:

1 Aged 15 at both quarters
2 Entrant to working-age between first and final quarter
3 In employment at first quarter; in employment at final quarter (EE)
4 In employment at first quarter; unemployed at final quarter (EU)
5 In employment at first quarter; inactive at final quarter (EN)
6 Unemployed at first quarter; in employment at final quarter (UE)
7 Unemployed at first quarter; unemployed at final quarter (UU)
8 Unemployed at first quarter; inactive at final quarter (UN)
9 Inactive at first quarter; in employment at final quarter (NE)
10 Inactive at first quarter; unemployed at final quarter (NU)
11 Inactive at first quarter; inactive at final quarter (NN)
12 Reached retirement age by final quarter

Therefore, if you run a frequency on FLOW you will get a summary of the movement in the labour market from one time period to the next.

For 2Q datasets, this shows flows over a 3 month period, whilst for 5Q datasets this shows the flow over a year.

In addition to FLOW, 5Q datasets have an additional variable ANFLOW (annual flow) that looks at the flow across all 5 quarters. Its categories are as follows:

1 In employment in all quarters (E)
2 Unemployed in all quarters (U)
3 Inactive in all quarters (N)
4 In employment at first quarter; unemployed at final quarter (EU)
5 In employment at first quarter; inactive at final quarter (EN)
6 Unemployed at first quarter; inactive at final quarter (UN)
7 Unemployed at first quarter; in employment at final quarter (UE)

8 Inactive at first quarter; in employment at final quarter (NE)
9 Inactive at first quarter; unemployed at final quarter (NU)
10 Employed at first; unemployed; in employment at final quarter (EUE)
11 Employed at first; inactive; in employment at final quarter (ENE)
12 Unemployed at first; inactive; unemployed at final quarter (UNU)
13 Unemployed at first; employed; unemployed at final quarter (UEU)
14 Inactive at first; employed; inactive at final quarter (NEN)
15 Inactive at first; unemployed; inactive at last quarter (NUN)
16 Employed at first; unemployed; inactive at final quarter (EUN)
17 Employed at first; inactive; unemployed at final quarter (ENU)
18 Unemployed at first; employed; inactive at final quarter (UEN)
19 Unemployed at first; inactive; employed at final quarter (UNE)
20 Inactive at first; employed; unemployed at final quarter (NEU)
21 Inactive at first; unemployed; employed at final quarter (NUE)
22 3 or 4 moves between categories

As the longitudinal datasets concentrate on at least two time periods, all the variables relating to the first of the quarters are renamed with a suffix of 1 added to the original variable name. All the variables relating to the second of the linked quarters are renamed with a suffix of 2 added to the original variable name and so on.

For example, consider a longitudinal 2Q dataset covering Q3 (JS) and Q4(OD) of 2016. There are two variables on the file called ILODEFR_1 and ILODEFR_2. ILODEFR_1 is ILODEFR from Q3 2016 and ILODEFR_2 is from Q4 2016.

All datasets have a unique personal identifier – PERSID. This is created during the linking of datasets.

Other Information

All analysis on any longitudinal dataset needs to be weighted by LGWT**. This then avoids non response bias and also produces estimates at the level of the population. Note that there is no attrition element in the weighting, as the datasets only include cases that respond in all waves. A more detailed explanation of the weighting methods can be found in the LFS longitudinal user guide (volume 11).

The sample size on the longitudinal files is not that big when compared to other LFS datasets. If you are analysing a certain sub-group of the population, you may find that the sample size is very small and care should therefore be taken before carrying out detailed analysis on longitudinal files.

See example 6 in the annex for a basic example of using the longitudinal datasets

Further Reading

Please see the longitudinal user guide (volume 11) for more information:

https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/labourforcesurveyuserguidance

7. Conducting APS Well-being Analysis

On the APS, the personal well-being questions (Satis, Worth, Happy, Anxious) are only asked of persons aged 16 and above who gave a personal interview; (this is approximately 54% of the cases on the APS person dataset), proxy answers are not accepted. The non-proxy well-being weight (NPWT**) is therefore calculated for each individual, and is zero for respondents who were under 16 years of age or who were not present in person for the interview. Cases with weights of zero will not be considered in data analyses (when the weight is applied).

Caution should be used when using analysis of responses to personal well-being questions at detailed geography areas and also in relation to any other variables where respondent numbers are relatively small. It is recommended that for lower level geography analysis the variable 'UACNTY09' is used.

It is not possible to combine other single year APS/Personal Well-being datasets together to carry out longitudinal analysis. The Personal Well-being datasets are not designed for longitudinal analysis, e.g. they are not designed to track individuals over time.

The ONS produce a Statistical Bulletin on Personal Well-being in the UK, which is available from the ONS website. It provides an overview of the initial analyses of UK personal well-being data and also includes a information on how personal well-being data can be used: http://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/measuringnationalwellbeing

8. Conducting APS Sexual Identity Analysis

Like with the personal well-being questions on the APS, the sexual identity question (SIDV) is only asked of persons aged 16 and above who gave a personal interview; proxy answers are not accepted, therefore the size of the achieved sample for Sexual Identity is much smaller than the full APS file. As a result any analysis by geographical area below regional level is not recommended, and that caution should be used for analysing Sexual Identity responses by other variables where unweighted respondent numbers are relatively small. The Sexual Identity weight is SIDWT**.

The ONS produce an experimental Statistical Bulletin on Sexual Identity in the UK, which is available from the ONS website. It provides an overview and analysis of UK Sexual Identity data and also includes information on how Sexual Identity data can be used

https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/sexuality

From January 2011 the APS person datasets contain the Sexual Identity variable and weight previously they were released as part of the Integrated Household Survey (IHS).

Annex of Examples (using SPSS and SAS)

Example 1 - An example of applying the wrong weight when conducting earnings analysis on the LFS

You want to find out the number of people who earn £1000 or more a week (gross pay). You choose to look at the variable GRSSWK.

You accidentally apply the wrong weight PWT17 (and not the specific earnings weight PIWT17).

This tells you that there are 657,624 people whose gross pay is more than or equal to £1000 a week.

Now let's apply the correct weight (PIWT17) and compare answers.

You now get the correct result that there are 2,100,626 people whose gross pay is more than or equal to £1000 a week.

The difference between the two values here is over a million. This demonstrates how important it is to apply the correct weight when carrying out earnings analysis.

Example 2 - An example of how applying the wrong household filter can affect your analysis

You want to find out the number **families** that consist of married couples with no children in the UK.

This can be done by looking at the variable FUTYPE6.

You know you need to apply a filter before you run a frequency of FUTYPE6 on a household dataset.

Your syntax looks like this:

In SPSS:

```
WEIGHT PHHWT17.

SELECT IF RELHRP6=0.

FREQUENCIES FUTYPE6.
```

In SAS:

```
proc freq data = js17hhld;
  where relhrp6=0;
  weight phhwt17;
  table futype6;
run;
```

After running this syntax, you look at the frequency and see that the frequency for FUTYPE6=4 is 5,951,493. You interpret this as there are 5,951,493 families that consist of married couples with no children in the UK (according to the JS17 household dataset).

However, this analysis does not tell you the **number of families** that consist of married couples with no children in the UK – it tells you the **number of households** that consist of married couple families with no children in the UK.

The variable, dataset and weight you chose were correct – however, the filter you applied is incorrect and has actually stripped out some families. This is because you reduced the dataset so that it only contained the Household Reference Person (HRP) as you used the wrong variable in the SELECT IF statement.

If a household had more than one family living there, only one of those families would have been taken into account during your above analysis.

This mistake is easily rectifiable. All you have to do is amend one line of your syntax:

In SPSS

```
WEIGHT PHHWT17.
SELECT IF RELHFU=1.
FREQUENCIES FUTYPE6.
```

In SAS

```
proc freq data = js17hhld;
 where relhfu=1;
  weight phhwt17;
  table futype6;
run;
```

You then get the correct result that there are 6,036,373 married couples with no children **families** in the UK.

This is a difference of 84,880 – therefore it shows us that the filter we choose is very important.

Example 3 - An example of how choosing the wrong type of household variable can affect your analysis

Same example as above - You want to find out the number of married couples with no children **families** in the UK.

You choose the variable HHTYPE6 (after all, this is the first variable that contains the phrase "married couple no children" when you search on that phrase in the Volume 3 User Guide).

Your syntax looks like this:

In SPSS

```
WEIGHT PHHWT17.

SELECT IF RELHFU=1.

FREQUENCIES HHTYPE6.
```

In SAS

```
proc freq data = js17hhld;
 where relhfu=1;
 weight phhwt17;
 table hhtype6;
run;
```

You look at the frequency and see that the frequency for HHTYPE6=3 is 5,673,501. You interpret this as there are 5,673,501 married couples with no children **families** in the UK (according to the JS17 household dataset).

However, this analysis is incorrect. The weight and dataset you chose were correct, as was the filter. This time, the problem is the variable.

HHTYPE6 counts the number of married couples with no children in households – not families.

This mistake is also easily rectifiable, you just have to change the variable to FUTYPE6, and then you get the correct result.

Example 4 - Creating your own household variable

You want to find out the number of rented households where somebody drives to work.

(a) You first of all need to decide whether this is person or household analysis - and then choose your dataset.

We want to know the **number of households** therefore let's choose a household dataset.

Make sure HSERIAL/HSERIALP is on your dataset.

(b) Decide on your variables

TRVMTH=1 tells us whether a person travels to work by car

TEN1=4 tells us whether a person lives in a rented house.

(c) Create your own household derived variable

There isn't a household variable that shows you whether anyone in the household goes to work by car – therefore we need to create one.

- Step one - Recode TRVMTH into a new variable CAR, where CAR=1 if TRVMTH=1 and CAR=0 if TRVMTH=2-9
- Step two - Add this variable up over everyone in the household (therefore creating a new household variable - TOTCAR)
- Step three – Match the new household variable back on to your original dataset

Make sure that both your datasets are sorted by HSERIAL/HSERIALP before you match the files.

(d) Decide if you want to apply any filters

We are now in a position to filter the dataset down to household level using the filter RELHRP6=0.

If we don't apply this filter, we would be carrying out person level analysis.

(e) You can then carry out your analysis.

Firstly remember to apply the household weight.

One of the ways to do this is to run a CROSSTAB of the new household variable (TOTCAR) by the variable that shows whether someone is renting (TEN1).

This is how we could do this:

In SPSS

```
GET FILE="D:\LFSH_OD16.sav".

RECODE TRVMTH (1=1)(ELSE=0) INTO CAR.

VARIABLE LABELS CAR 'Travels to work by car'.

EXE.


AGGREGATE

  /OUTFILE="D:\CARSUM.sav"

/BREAK=HSERIAL

/TOTCAR=SUM(CAR).

EXECUTE.
```

GET FILE="D:\CARSUM.sav".

SORT CASES BY HSERIAL.

SAVE OUTFILE="D:\CARSUM.sav".

GET FILE="D:\LFSH_OD16.sav".

SORT CASES BY HSERIAL.

MATCH FILES /FILE=*

/TABLE="D:\CARSUM.sav"

/BY HSERIAL.

EXECUTE.

WEIGHT BY PHHWT17.

TEMP.

SELECT IF RELHRP6=0.

CROSSTAB TEN1 BY TOTCAR.

In SAS

```
data car;
set LFSH_OD16;
car=0;
if trvmth=1 then car=1;
label car="Travels to work by car";
run;

proc summary data=car;
var car;
class hserial;
output out= carsum sum=totcar;
run;

proc sort data= car; by hserial; run;
proc sort data=carsum; by hserial;run;

data totcar;
merge car carsum;
by hserial;
run;
```

```
proc freq data = TOTCAR;
where relhrp6=0;
weight phhwt17;
table ten1*totcar;
run;
```

Using an OD16 household dataset, this is the output generated:

TEN1 Accommodation details * TOTCAR Crosstabulation

Count

		TOTCAR						
		.00	1.00	2.00	3.00	4.00	5.00	Total
TEN1 Accommodation details	1 Owned outright	6508395	1806645	729079	144669	32016	3115	9223919
	2 Being bought with mortgage or loan	2752458	2774665	2299704	287536	54778	5346	8174487
	3 Part rent	84591	61872	28000	1538	0	0	176001
	4 Rented	6332902	**2178696**	**727694**	**85671**	**17364**	**2217**	9344544
	5 Rent free	175242	54242	18760	2171	0	0	250415
	6 Squatting	925	790	0	0	0	0	1715
Total		15854513	6876910	3803237	521585	104158	10678	27171081

The numbers in bold are where at least one person in a rented household owns a car.

You can conclude from this that there are 3,011,642 rented households in the UK where somebody in that household drives to work. (This number was calculated by adding up all the underlined numbers in the output).

You can also see from the output that rented houses (TEN1=4) are the most common type of household where no-one travels to work by car.

Example 5 - Person level statistics broken down by the characteristics of the household

You want to find out the number of dependent children (0-18) living in households where someone is working.

(a) You first of all need to decide whether this is person or household analysis and then choose your dataset accordingly.

You might think that as we are talking about the number of children, we would want person analysis.

However, the **condition** is living in households where someone is working.

This means that we initially need to cut down our dataset so that we only have these types of households.

Therefore, this is household analysis and household datasets should be used.

(b) Decide whether you want to apply an initial filter. As we are carrying out household analysis, let's apply the filter RELHRP6=0 so we reduce the dataset to household level.

(c) Choose your variables. We need to decide on a variable that will tell us if anyone in the household is working. Volumes 3 and 4 may be of use here. I've decided to use HEACOMB:

HEACOMB - Household economic activity

(1) All persons in the household are employed
(2) All persons in the household are either employed or unemployed
(3) All persons in the household are either employed or inactive
(4) All persons in the household are either employed, unemployed or inactive
(5) All persons in the household are unemployed
(6) All persons in the household are either unemployed or inactive
(7) All persons in the household are inactive

We can see that HEACOMB<5 captures everyone in the household where someone is employed.

You also need to choose a variable that tells you how many dependent children (0-18) are in the household. I have chosen HDPCH19.

(d) Reduce your dataset down to the type of households you are interested in.

We are interested in households that contain someone who is employed, therefore run SELECT IF on HEACOMB<5.

We have produced a dataset that only contains households where someone is working.

(e) Now, we want to know how many children are in these households.

Run a frequency on HDPCH19, remembering to apply an appropriate household weight beforehand.

Using the JS17 household dataset, this is the output created:

HDPCH19 No. of dep children in hhld under 19

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	11200603	60.8	60.8	60.8
	1	3229899	17.5	17.5	78.3
	2	2937287	15.9	15.9	94.2
	3	809326	4.4	4.4	98.6
	4	205454	1.1	1.1	99.7
	5	36994	.2	.2	99.9
	6	12683	.1	.1	100.0
	7	2473	.0	.0	100.0
	8	664	.0	.0	100.0
	9	1500	.0	.0	100.0
	Total	18436883	100.0	100.0	

We have to be very careful here and think about what exactly this is telling us.

This tells us that there are 11,200,603 **households** where there are no dependent children in households where someone is employed.

There are 3,229,899 **households** where there is one child in a household where someone is employed.

There are 2,937,287 **households** where there are two children in a household where someone is employed.

It does not tell us the number of children. In order to work out the number of children, we need to add in an **extra step** to this analysis and multiply the frequency by the number of dependent children

HDPCH19 No. of dep children in hhld under

		Frequency	Frequency * Number of dependent children
Valid	0	11200603	0
	1	3229899	3229899
	2	2937287	5874574
	3	809326	2427978
	4	205454	821816
	5	36994	184970
	6	12683	76098
	7	2473	17311
	8	664	5312
	9	1500	13500
	Total	18436883	12651458

This then gives us the number we want – there are 12,651,458 dependent children aged under 19 living in households where someone is employed.

Example 6 - Example of using the longitudinal datasets

You want to find out the number of people who moved from unemployment to employment during the period AJ17 to JS17.

Here, you should use a 2Q dataset as you are interested in the change from one quarter to the next consecutive quarter. There is a Q2 to Q3 2017 2Q dataset available. As you are interested in movement within the labour market, you should use the variable FLOW.

Make sure you have applied the weight LGWT and run a frequency on FLOW.

The 6th category of FLOW "UE" shows the number of people who move from unemployment to employment. This will give you the number you are after (396,378 people in this case).

Glossary of terms

Attrition –The term used when cases drop out from a survey over time

Derived variables – Derived variables are not questions asked of respondents, they are calculated from a variety of variables. For example, say I ask a respondent whether their hair length (HAIRL) is short or long, and in a separate question ask them whether their hair colour (HAIRC) is brown or blond. I then combine this information to create a new variable HAIR that uses the information from HAIRC and HAIRL to create 4 different hair types (short brown, long brown, short blond, long blond). HAIR is therefore a derived variable, which uses information from two questionnaire variables HAIRL and HAIRC.

Family – The LFS definition of a family (or family unit is as follows): "Can comprise of either a single person, or a married/cohabiting couple, or a married/cohabiting couple and their never-married children who have no children of their own living with them, or a lone parent with such children".

Household – The definition used by interviewers to establish what constitutes a household is as follows: "One person living alone or a group of people (not necessarily related living at the same address) who share cooking facilities AND share a living room or sitting room or dining area."

Household Reference Person (HRP) – The household reference person is calculated by looking at the date of birth, age, which respondent has the name on the accommodation, and who is the highest earner of the household. This is for dataset purposes only and is not referenced at all during an interview. Not to be confused with the head of the household (HOHID).

Missing values – Not everyone on a dataset will have a value for every variable. The first reason for this is because the question/variable is not applicable to them. These are generally denoted by -9s on datasets. The other reason is that a person does not know or refuses to answer the question. These are generally denoted by a -8.

Non-responders - This is an individual who doesn't answer the questionnaire but who lives in a household where somebody else has responded. You can find out who the individual non-responders are by looking at the variable IOUTCOME.

Personal Wellbeing - Measured by ONS as part of the Measuring National Well-being programme. The LFS asks respondents for their views on their own well-being. LFS respondents are asked 4 questions in total: SATIS, HAPPY, WORTH and ANXIOUS. Volume 2 of the LFS User guide provides more detail. Please note that personal wellbeing is sometimes referred to as 'subjective wellbeing'.

Quarters – The LFS is a quarterly survey. This is because a slightly different questionnaire is used for the four quarters of the year - January-March (JM), April-June (AJ), July-September (JS) and October-December (OD). The LFS datasets reflect this and there are LFS person datasets for all of the four different datasets. APS datasets include variables common to **all** quarters only. The LFS moved to these calendar quarters in 2008, previously it was run on seasonal quarters. More information about this can be found in User Guide 3.