Office for National Statistics

Labour Force Survey

User Guide

Volume 11 - LONGITUDINAL DATASETS

LONGITUDINAL USER GUIDE
LFS TWO-QUARTER, LFS FIVE-QUARTER AND APS TWO-YEAR LONGITUDINAL DATASETS

Contents

Introduction

The Labour Force Survey (LFS) and the Annual Population Survey (APS) are household surveys, gathering information on a wide range of labour force characteristics and related topics. They both use the same core questionnaire, therefore many of the variables are available on both the LFS and APS, although the APS has a much larger sample than the LFS. More information about both these surveys can be found in the other LFS user guides (in particular volumes, 1, 6 and 10).

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

Since 1992 the LFS has been conducted on a quarterly basis, with each sample household retained for five consecutive quarters, and a fifth of the sample replaced each quarter see volume 1 of the LFS user guide for more information. The survey was designed to produce cross-sectional data but it has been recognised that linking together data on each individual across quarters would produce a rich source of longitudinal data therefore a 2 quarter and 5 quarter longitudinal datasets have been produced. The APS was introduced in 2004 to allow for analysis to be carried out on detailed subgroups and below regional level. In recent years (particularly with the sample size of the LFS 5 quarter dataset reducing) there has been some interest in producing a two year APS longitudinal dataset to look at any trends that may occur over a year.

There are however methodological problems which could distort the data resulting from linking. They fall into two main groups: biases arising from non-response and the sample attrition arising from it; and biases arising from response errors, particularly their effects in producing spurious flows between economic activity states. ONS undertook a joint research project with Southampton University to address these methodological issues, which produced a satisfactory methodology for compensating for the biasing effects of non-response, and a procedure has been developed for applying it in longitudinal datasets.

This guide describes the three different types of longitudinal datasets that are produced and how to use them, it does not give details of the methodological development - this is covered in detail in paper 17 of the GSS Methods and Quality series, entitled “Methodological Issues in the Production and Analysis of Longitudinal Data from the Labour Force Survey” by Paul Clarke and Pam Tate.

Datasets

Based on the LFS

The quarterly LFS started in spring 1992, but the rotational pattern of the sample was not established until winter 92/93, therefore this is the first quarter available for longitudinal linking. From May 2006, the LFS moved from seasonal quarters (e.g. Spring: March to May) to calendar quarters (April-June: Q2), the LFS user guide volume 1 provides more details about this.

Two-quarter longitudinal datasets have been produced for all pairs of adjacent quarters from winter 1992/93 onwards - for example, the winter 1992/93 dataset was linked with the spring 1993 dataset. Below (figure 1.1) is an illustration of the structure of the two quarter dataset

	LFS two quarter longitudinal dataset (Q3 to Q4 2018)				
	OD17	JM18	AJ18	JS18	OD18
LFS cohort 1 *(first sampled OD16)*	Wave 5				
LFS cohort 2 *(first sampled JM17)*	Wave 4	Wave 5			
LFS cohort 3 *(first sampled AJ17)*	Wave 3	Wave 4	Wave 5		
LFS cohort 4 *(first sampled JS17)*	Wave 2	Wave 3	Wave 4	Wave 5	
LFS cohort 5 *(first sampled OD17)*	Wave 1	Wave 2	Wave 3	**Wave 4**	**Wave 5**
LFS cohort 6 *(first sampled JM18)*		Wave 1	Wave 2	**Wave 3**	**Wave 4**
LFS cohort 7 *(first sampled AJ18)*			Wave 1	**Wave 2**	**Wave 3**
LFS cohort 8 *(first sampled JS18)*				**Wave 1**	**Wave 2**
LFS cohort 9 *(first sampled OD18)*					Wave 1

Figure 1.1: Structure of the two quarter longitudinal dataset.
The highlighted waves form part of the LFS 2Q longitudinal dataset.

Five-quarter longitudinal datasets have also been produced for the same periods, for example linking spring 1993 with spring 1994 and containing data from all five waves of the survey. Figure 1.2 illustrates the structure of the five quarter dataset

	LFS five quarter longitudinal dataset (Q4 2017 to Q4 2018)				
	OD17	JM18	AJ18	JS18	OD18
LFS cohort 1 *(first sampled OD16)*	Wave 5				
LFS cohort 2 *(first sampled JM17)*	Wave 4	Wave 5			
LFS cohort 3 *(first sampled AJ17)*	Wave 3	Wave 4	Wave 5		
LFS cohort 4 *(first sampled JS17)*	Wave 2	Wave 3	Wave 4	Wave 5	
LFS cohort 5 *(first sampled OD17)*	**Wave 1**	**Wave 2**	**Wave 3**	**Wave 4**	**Wave 5**
LFS cohort 6 *(first sampled JM18)*		Wave 1	Wave 2	Wave 3	Wave 4
LFS cohort 7 *(first sampled AJ18)*			Wave 1	Wave 2	Wave 3
LFS cohort 8 *(first sampled JS18)*				Wave 1	Wave 2
LFS cohort 9 *(first sampled OD18)*					Wave 1

Figure 1.2: Structure of the five quarter longitudinal dataset
The highlighted waves form part of the LFS 5Q longitudinal dataset.

The two-year APS longitudinal datasets has been produced for JD (January-December) periods from 2013, with the JD12 dataset being linked with the JD13 one. Figure 1.3 illustrates the structure of the two-year longitudinal dataset

	APS two year longitudinal dataset (Jan-Dec 2017 to Jan-Dec 2018)							
LFS cases	JM17	AJ17	JS17	OD17	JM18	AJ18	JS18	OD18
LFS cohort 2 *(first sampled JM17)*	**Wave 1**	Wave 2	Wave 3	Wave 4	**Wave 5**			
LFS cohort 3 *(first sampled AJ17)*		**Wave 1**	Wave 2	Wave 3	Wave 4	**Wave 5**		
LFS cohort 4 *(first sampled JS17)*			**Wave 1**	Wave 2	Wave 3	Wave 4	**Wave 5**	
LFS cohort 5 *(first sampled OD17)*				**Wave 1**	Wave 2	Wave 3	Wave 4	**Wave 5**
LFS cohort 6 *(first sampled JM18)*					Wave 1	Wave 2	Wave 3	Wave 4
LFS cohort 7 *(first sampled AJ18)*						Wave 1	Wave 2	Wave 3
LFS cohort 8 *(first sampled JS18)*							Wave 1	Wave 2
LFS cohort 9 *(first sampled OD18)*								Wave 1
Boost cases	JD17				JD18			
LLFS cohort 1 *(first sampled JD14)*	Wave 4							
LLFS cohort 2 *(first sampled JD15)*	**Wave 3**				**Wave 4**			
LLFS cohort 3 *(first sampled JD16)*	**Wave 2**				**Wave 3**			
LLFS cohort 4 *(first sampled JD17)*	**Wave 1**				**Wave 2**			
LLFS cohort 5 *(first sampled JD18)*					Wave 1			

Figure 1.3: Structure of the two year longitudinal dataset
The highlighted waves form part of the APS 2Yr longitudinal dataset.

Coverage

The focus of analyses of these datasets is on the population of working age. The dataset is restricted to those aged 15-69 in the first quarter. Before the 2010 reweighting, the working age was 15-59 for women and 15- 64 for men (which will be the case for datasets before OD01).

The small proportion of people in the sample whose data, at any of the linked quarters, had been imputed by rolling forward from the previous interview, were excluded from the longitudinal datasets.

From spring 1996 onwards, with the introduction of the household matrix approach to gathering data on the people present in the household, a small proportion of people in the sample have no data available on economic activity. People with no data on economic activity at one or more of the linked quarters have been excluded from the longitudinal datasets.

For the period from winter 1995/6, the datasets cover the UK. The Northern Ireland survey did not change from an annual to a quarterly survey until winter 1994/95, and the rotation pattern of the sample was not fully established until winter 1995/96, therefore the longitudinal datasets which include quarters before winter 1995/96 cover just Great Britain.

Linking procedure

The individual-level LFS dataset for the first of the period to be linked (quarter for the LFS or yearly for the APS) is used to produce a reduced cross-sectional dataset confined to the age range and variables to be used for the longitudinal dataset. A unique identification variable (PERSID) is created. A similar procedure is followed on the datasets for the other periods that are to be linked. The datasets for the two quarter, five quarter or two year to be linked are matched by the unique identification variable and checked to ensure that the cases linked match also on sex and date of birth. All unmatched cases are dropped, as are all cases where the data were rolled forward, or where there are no data on economic activity, at any of the quarters.

Variables

Because of the resources involved in production and the size of the resultant datasets, the longitudinal datasets include only a subset of the full LFS/APS variable set. This subset has been agreed in consultation with users and represents the most important and commonly used variables covering the main areas of the survey.

When the linked datasets are created, all the variables relating to the first of the linked quarters/year are renamed, with a suffix of 1 added to the original variable name, and all the variables relating to the second of the linked quarters/year have a suffix of 2 added to the original variable name, and so on. For example, if we link together the OD17 and JM18 quarters, then the variable TEN1 from the first quarter, OD17, becomes TEN11 in the linked dataset, and TEN1 from the second quarter,

JM18, becomes TEN12, and so on until the OD18 variable becomes TEN15. This is true for all the variables in all the datasets, except for the unique identification variable and the variables for sex and date of birth which are used for checking that the match between the quarters are correct. These must be identical for each of the quarters being linked and therefore have no suffix.

Some of the variables are not available for all the periods (e.g quarters) and are therefore not available in one or more of the quarters/years of some of the linked datasets. For example, the variable TRNLEN (Length of training course) is only available for JM and AJ quarters from 1997. Therefore on the AJ18 5 quarter longitudinal dataset it will be available for the first quarter (AJ17) as TRNLEN1, the fourth quarter (JM18) as TRNLEN4 and the fifth quarter (AJ18) as TRNLEN5, TRNLEN2 and TRLEN3 won't be available as they represent the JS and OD quarters where the TRNLEN variable isn't available.

As SPSS only allows variable names to be eight characters long, a few variable names which are already eight characters long have to be amended when the suffix is added. These are as follows:

	1st Qtr	**2nd Qtr (3rd, 4th, 5th)**
IOUTCOME	IOTCOME1	IOTCOME2/3/4/5
SHFTWK99	SHTWK991	SHTWK992/3/4/5
HIQUAL15	HIUAL151	HIUAL152/3/4/5
HITQUA15	HIQUA151	HIQUA152/3/4/5

A variable FLOW has been added to the datasets. It gives in a convenient form the categories relating to labour force gross flows, distinguishing between states in and outside working age. The codes and 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	

For the two-quarter datasets this variable shows the flow over a 3 month period, while for the five-quarter datasets it shows the flow over a 12 month period. For the two-year datasets this variable shows the flow over 12 months (the quarter in the label categories is replaced by year)

In addition, a variable ANFLOW has been added to the five-quarter datasets. It gives categories relating to labour force gross flows across all five of the linked quarters. There are 243 possible sequences over five quarters, many of which will have very

small frequencies, particularly the ones involving 3 or 4 moves. For this reason a simplified categorisation is presented which combines together those sequences where only the timing differs - for example, all cases which start in employment and end in unemployment (with no other transitions) are in category 4 below, regardless of the wave in which they became unemployed. The codes and 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	

Sample size

Because of sampling variability, the smaller the group being estimated the poorer the precision of the estimate becomes, until eventually the estimate is not reliable enough to be used. (See Volume 1, in particular section 8 of the LFS User Guide for a detailed discussion.) For the two-quarter longitudinal datasets, the number of sample cases available for linkage is around 28,000. For the five-quarter dataset it is around 4,000, therefore the results are subject to greater variability due to higher attrition. The two-year APS longitudinal dataset has around 69,000, so could be used as an alternative to the five quarter.

Reliability threshold levels

For the regular quarterly cross-sectional LFS datasets, a publication threshold is set at 20,000 (i.e. estimates below 20,000 are not published), at which level the standard error is about 20% of the estimate, and the 95% confidence interval for the estimate is about +/-7,500. This corresponds to having about 25 cases in the group. For the two-quarter longitudinal datasets, the same principle applies, but the number of

sample cases that are linked is smaller (usually around 28,000), so the threshold level for these datasets is higher; it is about 40,000. For the five-quarter dataset, the threshold is about 280,000.

For estimates below these figures, the standard error is likely to be greater than 20% of the estimate and therefore the estimate should not be used. The figures are higher for the unemployed category because of the design effect and higher attrition within this group. For some of the other categories, particularly those involving more than one transition, there may be very few cases present in each dataset. Therefore, it may be necessary to combine categories or use several datasets to reach the required threshold and get a reliable result.

Weighting

LFS longitudinal weights are produced for the following datasets: two-quarter, five-quarter and two-year. The weighting system is applied on the linked samples of respondents; to reduce the effect of non-response bias, we adjust the pre-calibration weights to maintain the distribution of tenure and calibrate on estimates of economic activity from the previous and current period, in addition to sex, age and region.
The weight (LGWT**) (where **denotes the year that the weight was published) for these datasets serves two purposes: it compensates for non-response bias; and produces estimates of economic activity at the level of the longitudinal target population, composed of people in the age group 15 to 69 in the first period.

The calculation of the weights for the two-quarter datasets involves the following stages:

1. Calculating the initial prior weights which reproduce the distribution of the cross-sectional sample from the first quarter according to the tenure/landlord categories: owned; rented from local authority/housing association; privately rented.

2. Adjusting the calculated initial prior weights by multiplying with a single scaling factor, (except for Northern Ireland where this factor is again multiplied by an adjustment factor to compensate for the different sampling fraction), such that the weighted sample cases sum to the overall population control total. For multi-occupancy households, the initial prior weights are adjusted by multiplying them by the number of households at the address, or 4, whichever is the smallest, prior to applying the scaling factor.

3. Applying the calibration method (also known as generalised raking) using CALMAR software (see Elliot 1997). This process minimises the distance between the prior and final weights, while constraining the final weights simultaneously to several marginal distributions or control totals. Four sets of control totals are used:

    a) the population estimates used for weighting the second quarter's cross-sectional LFS dataset, for the selected age range, classified by sex and

age (15-17, 18-20 and then five-year age groups) - this produces estimates as close as possible to the population available for sampling in both the linked quarters;

b) the population estimates used for weighting the second quarter's cross-sectional LFS dataset, for the selected age range, classified by region;

c) the weighted cross-sectional estimates from the second linked quarter for the selected age range classified by broad economic activity categories: in employment; self-employed; unemployed; economically inactive;

d) the weighted cross-sectional estimates from the first linked quarter for the selected age range classified by broad economic activity categories: in employment; self-employed; unemployed; economically inactive, adjusted to the same total as (a) to (c) by reducing the economically inactive category as necessary.

CALMAR software (developed by French statistics offices) is run using the log distance function. This is an iterative method which converges to a solution when the sum of the final weights is very close to the calibration totals in each group.

The extension of the two-quarter method to create five-quarter weighted datasets consists of constraining to the cross-sectional economic activity distribution at each of the five quarters. This involves repeating the constraint in (iii)d for the second, third and fourth quarters as well as the first, adjusting in the same way to achieve a total consistent with the fifth quarter. When running CALMAR to create five-quarter datasets, wider limits have to be set for the ratio of final to prior weights, typically 0.1 to 3.1.

For the two-year dataset, the process is the same as the two-quarter dataset, except years are used instead of quarters.

Some points on longitudinal analysis, including the implications of response error bias

The longitudinal data only contains individuals who responded in consecutive waves, and due to attrition, the sample size is smaller than the cross-sectional datasets and the sample composition is different. Depending on both the response rate and who respond, there will sometimes be differences between the two-period longitudinal and cross-sectional datasets.

The longitudinal weighting is constrained so that both periods have identical population totals reflecting the population total in the cross-sectional for the first period. An additional constraint exists on economic activity which calibrates employment and unemployment numbers in the consecutive periods to the same numbers in the corresponding first cross-sectional period. Combined, this means that there will likely be some underestimation of inactivity numbers in the longitudinal data.

Consideration for longitudinal analysis:

- Given the constraint on population totals for the consecutive period on a longitudinal dataset and the calibration to employment and unemployment levels, inactivity flows using these datasets should be treated with caution.

- Comparisons between longitudinal and cross-sectional datasets are not recommended. Due to differences in the sample, some differences will be present between longitudinal datasets and cross-sectional datasets of the same year. Users should pick one type of dataset for an analysis, rather than comparing the two datasets.

- Each longitudinal dataset should be used as a standalone dataset; analyses should be completed within that single dataset. Calculating changes over long periods of time should be avoided, as due to cohort effects, there will be discrepancies between the end of one dataset and the beginning of the subsequent dataset.

- All analyses should be run weighted by LGWT**, otherwise the results will be distorted by non-response bias, and possibly misleading.

Careful thought is needed about the precise coverage of any analysis – is it the population of working age in the first period, the last period, or all periods? The variable FLOW can be used to select any of these groups: codes 3 to 12 give working age at the first period, 2 to 11 at the last period, and 3 to 11 in all periods.

Most analyses of interest are likely to be cross-tabulations of a characteristic at the first period with a characteristic a later period (quarter or year), often restricted to a subgroup. Some examples are: lone parents of working age at both quarters by sex and age of youngest child and by economic activity at both quarters; young people aged 18 to 24 unemployed at the first quarter by educational qualification and economic activity at the last quarter; people reaching retirement age by the last quarter by economic activity at both quarters and by reason for inactivity if inactive. Doing analyses of this kind, the numbers of cases in some cells can very quickly decrease.

Research so far on response error has been based on empirical analysis of differences in levels of transitions between different economic activity categories and of apparent internal inconsistencies. The initial investigations have provided evidence supporting the suggestion that response error is likely to affect the longitudinal datasets, probably in the direction of an upward bias in estimates of gross flows between different broad economic activity categories. It has also provided some tentative indications of transitions and subgroups particularly likely to be affected. These are transitions between unemployment and inactivity, transitions between part-time employment and either unemployment or inactivity, for women any transitions involving unemployment, and for students transitions between employment and unemployment. However, some of the apparent inconsistencies may be caused by genuine volatility (repeated movements back and forth between different economic activity states) rather than by response error.

Contact details

All enquiries about the longitudinal datasets should be directed to socialsurveys@ons.gov.uk.

Reference

Elliot, D (1997) *Software to weight and gross survey data.* GSS Methodology Series No 1.