DEFINING A STUDY POPULATION

Introduction

After reading this document you should be able to:

- Specify the type of study you are undertaking.
- Define your study population.
- Explain the difference between matching, tracing and presence at census, and the implications for using census and events data.
- Exclude from your study population any subjects that will cause data analysis problems.

The ONS Longitudinal Study (LS) is a rich source of data that can be used for several kinds of study. Users need to select parts of the LS appropriate to their study design. This document will help you to work out what selection you need for your study.

Structure of this document
This document will look at three examples of studies using the LS that illustrate the different ways in which a study population can be constrained by the criteria for inclusion of the study subjects.
The studies all focus on women aged 15-19 at the time of each census, and the examples are kept as simple as possible in order to demonstrate the underlying principles.

First however, we will discuss the common issues surrounding matching, presence at census and tracing, which will largely determine the study population. Then we will look at those groups of people who may need to be excluded from a study. Finally, we will work through the example research questions to illustrate these principles in practice.

Matching and tracing

LS members can be identified as being present at a census using the HISCEN variables in the CORE1 table. HISCEN71 indicates if a LS member was present at the 1971 census, HISCEN81 at the 1981 census, HISCEN91 at the 1991 census, HISCEN01 at the 2001 census, and HISCEN11 at the 2011 census (see the Data Dictionary for these variables).

Matching and tracing are key issues to be considered when defining a study population:

- Matching is the process carried out by ONS in the creation of the LS. It consists of taking a new census record for a person born on one of the four LS birthdays and finding a corresponding record for that person in the LS. If a record has been successfully matched, we can study that person longitudinally.
- Tracing is the process of finding a LS member's record in the NHS Central Register (NHSCR). Once a person has been traced, we can study the events[1] that they experience.

Matching
Matching is the process of linking a new census record for a LS member to their existing record in the LS.

Let's consider where the LS record came from in the first place. The initial sample was taken from the 1971 census. By 1981, the sample had been modified by new births and immigrants entering the study, and emigrants and deaths leaving the study. Therefore, it was this modified sample to which the 1981 census returns were linked.

[1] Events in the lives of LS sample members that are added to the LS dataset include their birth and death, live and still births (to women), cancer registrations and widow(er)hoods.

All people at the 1981 census born on one of the four LS birthdays were eligible to join the LS. The matching process compared these people with the existing LS records to see if they were already in the LS. Those that were already present were called "matched", and the 1981 census data was added to their existing LS record. Unmatched individuals were added as new members of the LS (after a manual matching process to double-check for existing records).

The key issue for a user of the LS is whether or not a LS member was present at a given census. Where a study compares the same characteristics between two censuses, the LS member must be present at both censuses in order to be included.

Linkage rates
After a new set of census records is added to the LS, there will be a number of LS members who were present at the previous census but are no longer present in the new census. Most of these people will have either died or left England and Wales ("embarked") in the inter-censal period. However, there will be a remnant whose absence is unexplained.

In the same way, there will be several unmatched records that should have entered the LS via immigration or birth registration. Such records highlight deficiencies in these registrations or deficient information on the census form. We refer to such cases as linkage failures.

The quality of linkage is measured by the linkage rate. For a given group (e.g. everyone present at a given census, or all new births during one inter-censal period), the linkage rate to the following census is given by the proportion of members of that group who are successfully matched at the next census.

Linkage rates for the LS in general are high (e.g. ~90% between successive censuses). For 1971-1981, the linkage rate was 91.3%, for 1981-1991 it was 90.1%, for 1991-2001 it was 88.0%, and for 2001-2011 it was 87.7%. However, for certain subgroups they are lower, and therefore you may find that some groups will be under-represented in your study. Rates for sub-groups including age, marital status, and ethnic group have been produced by ONS (see the ONS Longitudinal Study website for links to documents describing the linkage rates for the successive inter-censal periods). Rates for some minority ethnic groups are poor, with the linkage rates for Black Caribbeans and Black Africans less than 70%.

Tracing
A 'traced' LS member is one whose record has been found in the NHSCR, meaning that they registered with an NHS general practitioner (GP), or have accessed the NHS, and have been given a NHS number. Tracing allows events (such as births, deaths, cancer registrations) occurring to LS members to be linked to their LS record.

There are two main ways in which events are added to the LS. First, through routine notifications of events to NHSCR, and second, by a search through the annual ONS files for events occurring to people with a LS birthday. In the second case, the events detected are sent to NHSCR to be linked to LS members. Hence a LS member must have an NHSCR record in order to have events linked to their LS record. Therefore, it may be advisable to remove untraced LS members for the analysis of event(s) of interest to avoid denominator bias; only LS members for whom we know for certain that the event has not occurred will be included in the denominator. Although uncommon, there may be instances where a LS member has experienced an event but not been traced, such as when he or she has sought private medical treatment when an event was experienced.

This places an additional constraint on a study population when you are doing a study involving events, for example a study of the association between housing tenure and mortality. In order to determine the LS member's housing tenure you need their census details, and to detect the death of that member they must also be traced by the time of the census.

Tracing history
Tracing is an ongoing procedure carried out in the development of the LS. With a few exceptions, it is only those events that occur after the member is traced that are added to their LS record, so the date at which tracing occurred is also important.

Tracing can occur at the time of the census, when LS members' records are found in the NHSCR, but also between censuses, when an event occurs to a LS member who had not been traced until that point. An event is added to the LS when the NHSCR is notified of it via a registration process (e.g. birth registration, death registration, a new immigrant registering with a GP, etc.)

In the data dictionary the best variable to use to find out the tracing history for LS members is TRACE (see the CORE1 table in the Data Dictionary). This variable tells you whether the LS member has been traced, and when the tracing occurred. The variable takes an odd-numbered code when the member has been traced at a census, and an even-numbered code when tracing occurred between censuses (i.e. when an event was recorded for that member).
NB: the tracing history indicators do not give you exact dates of tracing, they just tell you that a person was traced at census or between censuses.

Tracing rates
There are various ways in which tracing can fail. Some of the main ones are as follows:
- LS members who are found at census but have not registered with a GP.
- When the date of birth on the LS record is inconsistent with that on the NHSCR record (i.e. one or none of the component parts match [day, month, year]), because one of the dates has been incorrectly recorded.
- When the surname is inconsistent (especially for recently-married women who are still registered in the NHSCR under their maiden name).
- LS members who have recently moved to a new house may not have re-registered with a GP at their new address.

The extent of these tracing failures in a given census year is measured by the 'not traced' rate. This is the proportion of LS members in a given subgroup not traced at NHSCR and for whom no event data can be linked. For the LS as a whole, the not traced rate is rather low: 3.2% in 1971, 1.2% in 1981, 1.7% in 1991, 0.7% in 2001 and 1.2% in 2011 (see the ONS Longitudinal Study website for links to documents describing the tracing rates for the successive inter-censal periods). However, in certain subgroups it can be much higher. If you're studying events, you may find that such groups are under-represented. For example:
- In 1971 people born in the New Commonwealth had a 'not traced' rate of 17%. Possible reasons for this were immigrants who had not yet registered with a GP, inconsistent recording of names (e.g. forename/ surname confusion), or inconsistencies in reporting dates of birth.

Tracing rates for sub-groups, including age groups, marital status groups and ethnic groups, have been produced by the ONS (see the ONS Longitudinal Study website).

Special cases & exclusions

Certain categories of LS member present problems when included in a study. This section looks at each of these categories, and the issues that they raise (the number and proportion of LS members present at census in each category are shown in parentheses):
- Visitors and absent residents (only relevant for 1991; at 2001 and 2011 individuals were enumerated at their usual address) [1991: 12,090 (2.22%)].
- People in communal establishments [1991: 7,793 (1.43%); 2001: 10,002 (1.85%); 2011: 10,156 (1.73%)].
- Students [1991: 18,630 (3.4%); 2001: 25,682 (4.76%); 2011: 36,976 (6.31%)].

- Records that contain data discrepancies or other data problems (number depends on the study).

In each case, you may have to decide whether to include the group in question. This will depend on the nature of your study. In some cases, a group will be unavoidably excluded, because of the research question being asked.

If a group is excluded, you will have to consider the characteristics of that group and consider the effect of their exclusion on your results. For example, the age distribution of students is concentrated in the 18-21 years age band, so if you exclude students from your study the age distribution of your sample population will differ from that of the overall population.

Visitors and absent residents
A visitor is a person enumerated at the census on a household census form at an address other than their usual one. This causes problems because we do not know the characteristics of the visitor's usual address. This is less of a problem for the 2001 and 2011 censuses because the form instructed the person completing the form only to include those individuals who usually live at the address, including anyone who is temporarily away and schoolchildren and students if they are away during the school or university term.

An absent resident is a person who usually lives at a given address but was away on the night of the census. The form filler is also instructed to complete details of absent residents on the census form (note though that this data was not added to the LS in 1971).

Sometimes a visitor at one address is also recorded as an absent member at another address. The LS member appears in the main members' file as a visitor and can be linked to their household data in the "absent, usually resident" members' file. In 1981, 3,994 of the 10,781 visitors enumerated (37%) had corresponding "absent resident" records, and in 1991 the number was 7,402 out of 12,090 visitors (61%). However, this "absent members" file is seldom used in practice because the extra work required to link in the records and account for duplicates does not usually justify the small increase in sample size that results.

Since the household characteristics of visitors in the LS apply to the address they are visiting, rather than their usual address, you might prefer to exclude them from your study. For example, if you are interested in the subject's car access, or whether they live with their parents, the data recorded for the address they are visiting would not be appropriate. This is also an issue if you are undertaking a geographical analysis – in this case using the ‘absent members’ records might be worthwhile.

If visitors are excluded, you should consider the characteristics of the excluded group. For example, many students will appear as visitors when enumerated at their term-time address (see "Students" below).

In 2001 and 2011, people who were temporarily away from home and those who were working away from home but usually resident were (supposed to be) enumerated at their usual address. There was therefore no information on visitors, with one exception: those who were present on Census night but who had no usual address were enumerated at the address where they were present.

Communal establishments
The inclusion of people living in communal establishments (and others not in private households, such as the homeless, travellers and those on boats) is one of the strengths of the LS, since most comparable longitudinal studies are restricted to those in private households.

However, people in communal establishments do not complete any census questions concerning their living arrangements. Instead they complete an 'individual' (or 'personal', in 1971) census form which

only includes questions about themselves. The individual forms used at each census can be accessed through the CALLS-HUB.

This means that, as for visitors, there is no household information for people in communal establishments. There is a significant number of such people; 1991: 1.43% of LS members present; 2001: 1.85% of LS members present; and 2011: 1.73% of LS members present. They include those living in:

- hospitals
- care homes
- prisons
- defence establishments
- educational establishments
- hotels

Note that some people enumerated in communal establishments may also be enumerated elsewhere. For example, some elderly people in care homes may have a partner living in their private home who will record them as an absent resident.

The lack of household data means that for many studies, persons living in communal establishments will be excluded. As usual, the characteristics of the excluded population should be considered. Exclusion may cause biases if the outcome under investigation is associated with age, sex, long-term illness, poverty, etc.

In 2001 and 2011, the definition of a communal establishment changed from 1991 and was "an establishment providing managed (supervised full or part-time) residential accommodation". This definition included some medical facilities. The variables CETC0 and CETC11 indicate the type of communal establishment in 2001 and 2011, and the variables CEMTYPE0 and CEMTYPE11 indicate how the establishment is managed (e.g. by the NHS, by a charity/ voluntary organisation) (see tables ME01 and ME11 in the Data Dictionary).

Students

Students are a special case because they constitute the largest group of people who are away from their usual address for a substantial part of the year. The effect on the census varies: in 1971 the census took place during some student holidays, in 1981 it was during term-time, in 1991 during the holidays, and in 2001 and 2011 during term-time.

There are two main effects of this to consider: firstly, whether we can obtain household information for students at a particular census; secondly, how to distinguish a student's home and term-time addresses.

1971

The census was conducted during term time for some institutions, but out of term-time for others. Therefore, some students were present at their home address whilst others were at their term-time addresses. For students who were at their term time address, the 1971 census form specified that they should put their home address as their usual address. They were defined as visitors and we have the household information for their term time address, so long as it was not a communal establishment (e.g. halls of residence).

Back at home, the 1971 census form recorded basic information on absent usually resident people, but these data were not included in the LS (owing to the limitations of computer storage and processing capacity at that time). Therefore, there is no household information for the usual address of these students. Furthermore, a student living at their term-time address might have regarded it as their usual address. Therefore, they will be usually resident there and those in private households will

have household information for that address. In 1971, students aged ≥15 can be identified in the LS by the variable STUJOB7 (see the Data Dictionary for details).

1981

The 1981 census was held during term-time, so most students would have been at their term-time address. The census form specified that students living away from home during term-time should put their home address as their usual address. In such cases, students enumerated at their term-time addresses were classed as visitors at that address. However, the data in the LS show that in fact 93% of students were enumerated at their "usual" address. This suggests that either they regarded their term-time address as their usual address, or else that they were in fact at home at the time of the census. Therefore, in 1981 it is not possible to determine whether a student's household information applies to their home address or term-time address.

Students in private households took the household information of that address, but those in communal establishments (e.g. halls of residence) could not. The proportion of all students in the LS at 1981 who were enumerated in communal establishments was 631 out of the 16,773 (4%).

Students away from home during term time were also entered as usually resident on a census form at their home address. This information, including household data, is on the "absent, usually resident members" file, though as stated above, this file is not often used in practice.

Students aged 16 or over can be identified by the variable ECONACT8 (see the Data Dictionary for details).

1991:

The 1991 census took place during the holidays, so most students were at their home address. Therefore, the household and relationship information recorded was usually that of their home address.

For any students who were still at their term-time address, the 1991 census form specified that they should put their home address as their usual address. In such cases, they were defined as visitors and we have the household information for this term time address if it was not a communal establishment (e.g. halls of residence). Back at home, the 1991 census form recorded individual and household information on absent usually resident people. These data can be found on the "absent, usually resident members file", though as stated above, this file is not often used in practice.
A student living at their term-time address might have seen it as their usual address. Thus, they will be recorded as usually resident at their term-time address and, if in a private household, they will have household information for that address.

This census was also the first to ask for the term-time address of students as a separate question on the form. Therefore, for students enumerated at their home address, we can identify those who lived elsewhere during term-time. See the variable WERASTU9 in the Data Dictionary for details.
Another consideration relevant for students is that in the 1991 census they could also be classified as employed. The economic activity variable ECONPO9 only shows those students who did not work but the variable ECONPO89 identifies all students whether economically active or not (see the Data Dictionary for details).

2001 and 2011:

Information at the 2001 and 2011 censuses was collected on the basis of usual residence; it was not collected for visitors. Students and schoolchildren at their home address on the day of the census were asked to complete all the questions if they lived there during term-time. If they lived elsewhere during term-time, they were only asked to complete the first five questions (sex, date of birth, marital status, relationships). They were not included in families, household size and household composition

variables. A student living away from home during term-time was fully enumerated at their term-time address.

The variables STUP0 and STUP11 identify LS members who are full-time students or schoolchildren, and TTIND0 and TTIND11 indicate whether they were at their term-time address or not (see the Data Dictionary for details).

The economic variables ECOP80 (2001) and ECOP81 (2011) place full-time students in a separate category regardless of whether they were economically active or not (and so are consistent with ECONACT8 and ECONPO89), and NS-SEC (National Statistics Socio-Economic Classification) also has a category for full-time students regardless of economic activity. The variables ECOP0 (2001) and ECOP11 (2011) distinguish between economically active and inactive students (see the Data Dictionary for more information on these variables).

In summary:

- 1971 Census: taken during holidays for some, so students enumerated at both home and term addresses.
- 1981 Census: taken during term-time, but most students enumerated at "usual" address.
- 1991 Census: taken during holidays, so most students enumerated at home address. Students who live elsewhere during term-time can be identified.
- 2001 and 2011 Censuses: taken during term-time, so students enumerated at their term-time address. Only basic information collected at their home address if studying away from home.

Other exclusions

Where there are inconsistencies in the LS or other data problems, these records might need to be excluded. The numbers involved can be small, but they vary by study. The CeLSIUS support staff will advise you on the exclusions that should be considered for different research questions.

Some types of data problem are described below.

Discrepancies

Certain records in the LS contain discrepancies where the same item of data is held in two different files, and the values do not correspond. For example, if a female LS member registers a birth, her own date of birth as recorded on the birth registration might differ from the date of birth recorded in the CORE1 table when she entered the LS. Discrepancies in the sex of LS members are also sometimes observed. It is sometimes hard to determine whether the discrepancy is due to different recording for the same LS member, or whether it indicates a matching or linking error, i.e. there are two different people's data combined here. For this reason many researchers prefer not to make assumptions about what has caused a discrepancy unless there is good reason to trust such an assumption.

Researchers might decide to exclude discrepant records because they cannot be confident which is the correct value. Alternatively, the record could be retained if it is possible to verify it against another field, such as a different census point. In many cases, the discrepancy will not matter because the data that is affected by the discrepancy is not being used.

The LS provides six indicator variables for date of birth and sex discrepancies in the CORE1 dataset. The variables DOBD001 to 6 and SEXD001 to 6 allow these records to be identified. The variables indicate the type of LS record containing the discrepancy (or -9 if there is no discrepancy). See the Data Dictionary for more details.

Complex multiple enumerations

This is where a LS member is enumerated at more than one address at the same census. Such records are identified in the variables HISCEN91 (1991), HISCEN01 (2001) and HISCEN11 (2011) in the CORE1 dataset. In HISCEN91, 3 denotes a complex enumeration. In HISCEN01 and HISCEN11, complex enumerations are coded 2 (see the Data Dictionary for more details). In 2001, MEIND also identifies LS members with multiple enumerations. For these individuals only one record is retained on the members file (ME01). In 2001 around 5,700 LS members had multiple enumerations, and in 2011, it was around 17,000.

Duplicate family relationships

This is where a LS member appears to be living with more than one father, mother, or spouse. This is most likely to appear in complex households, where the relationships of the members cannot be fully established from the census questions. For example, some households consist of seven or more young adults together with an infant (or two), and the algorithms that allocate the parents of that infant may ascribe parent status to all the adult household members. However, if the young adults have similar socio-economic characteristics, it may not matter for the research study which ones are the biological mother and father of the baby. Similarly, if an adult LS member has both parents and parents-in-law in the household and these cannot be distinguished, it may not matter.

The next three sections look at the main types of study that can be undertaken using the LS. By working through an example research question for each type of study, we will identify issues that may arise when defining the study population.

Cross Sectional Studies

Research question: *How has the proportion of 15-19 year-old women living in social housing changed between 1971 and 2011?*

This is a cross-sectional study because it is looking at a separate sample of women at each census, rather than following up the same women through all the censuses.
NB: since this is a cross-sectional study, there are data sources other than the LS that could be used. Normally for a cross-sectional study you would use the Census Microdata, formerly known as Samples of Anonymised Records; see https://census.ukdataservice.ac.uk/use-data/guides/microdata.aspx[2]. Alternatively, you could use one of the ONS surveys, for a broader range of questions but with a much smaller sample size – the Labour Force Survey would be one possibility, or the Annual Population Survey and (earlier) the Integrated Household Survey.

Housing tenure in the census
You will need to identify people who were living in social housing at each census. Social housing is used to mean properties rented from a local authority or from a housing association. By looking at the census forms you can see how the housing tenure question was asked at each census, and therefore how social housing will be represented in the LS.

Study subjects
Since all the data you are using comes from the census, all that is required of the study subjects is that they are present at the census for the census year that you are looking at. However, some study members will be excluded because household information is inappropriate, such as people who completed an individual form, e.g. those in communal establishments, and visitors (assuming you have chosen not to use any absent members' records).

[2] These are extracts from Census records which are designed to enable researchers to carry out detailed analyses using Census data for individuals or households. They have been produced by ONS for each census year since 1991. Some reductions to the detail for certain highly visible or disclosive variables have been made.

Table 1 (see the Cross-Sectional Study tab in the associated Excel file) shows the number of female LS members aged 15-19 who were present at each of the 1971, 1981, 1991, 2001 and 2011 censuses. The second column shows the effect of the exclusions. The table shows that excluding these groups does not seriously reduce the sample size in this case.

Results
Table 2 (see the Cross-Sectional Study tab in the associated Excel file) shows the numbers of women aged 15-19 living in social housing in each of the census years examined (1971, 1981, 1991, 2001 and 2011) and row percentages, the proportion of women aged 15-19 in each census year who are living in social housing. The proportion is seen to fall with each successive census to 1991, remaining stable since then.

Longitudinal studies

Research question: *Among women aged 15-19 and still living with their parents in 1971, were those living in social housing more likely than those with other types of tenure to have left the parental home by 1981? What about by 1991?*

For longitudinal studies, the emphasis is on how the study subjects' individual circumstances change over time.

Study sample
Such questions are where the LS comes into its own, because it allows researchers to follow up individual LS members over many years. This example also makes use of information in the LS on the co-residents (i.e. other household members) of LS members (commonly known as "non-members"). The question requires census information and therefore the important issue in defining the study population is presence at census. However, this time in addition to being present in one census year, the subjects need to be present in other census years.

Living with parents
To determine whether a person is living with their parents, you need to find out how they are related to the people they live with.

The LS not only has census information for the LS members themselves, but also for the people that they live with. This information in the "non-members" file. You can use this information to find out which members are living with their parents. The non-members file contains the same census information that is collected for LS members but for each of the people that they live with. The variable names for non-members are usually the same as for the LS members but prefixed by the letter 'n'.

Each LS non-member's record contains a corresponding LS core number that identifies the LS member that they live with. So, to determine a parent's presence, you can just list all the non-members that are linked to the study subject and see if any of them is a parent of the study subject (see the Data Dictionary for the variables in non-members files that tell you the relationship of the non-member to the LS member).

For this example, do not include parents-in-law (they are not the actual parents of study members). There is no code for step-parents, so you will need to assume that these are grouped with parents. Also, there is no way of using census data to find those living with a foster-parent or other guardian acting in loco parentis. These people will be excluded. In addition, you will need to exclude both visitors and those in communal establishments. Exclusion is necessary for visitors, because the non-members file records their relationship to the other members of the household they are visiting, not where they usually live. It is also necessary for people in communal establishments because there is no non-members data, and therefore relationships cannot be established.

Study subjects
You are following individuals from one census to the next, and therefore the study subjects must be present in both the censuses that you are using in your study (in this case 1971 and 1981, and 1981 and 1991) in order to be useful. This longitudinal requirement means that the sample sizes you obtain will be smaller than they were for the cross-sectional study (certain members will enter or leave the study between censuses due to death or migration).

The initial study cohort is taken from women aged 15-19 and living with their parents in 1971, but the actual study sample will only include those from that group who are still present in 1981. To illustrate this, table 3 (see the Longitudinal tab in the associated Excel file) shows the numbers of women aged 15-19 and living with their parents at the 1971 census, and the numbers still present at the 1981 and 1991 censuses.

You need to know whether the subjects are living with their parents at the end of each study period, therefore you need household data for their usual address at end of the study period as well (i.e. 1981 or 1991). Therefore, women who are enumerated as visitors or in communal establishments will not be included (the exclusions for 1971 have already been made in effect because you are looking at women living with their parents, and this can only be determined where valid household data is available). The table also shows the numbers without these excluded groups.

Results
Table 4 (see the Longitudinal tab in the associated Excel file) shows numbers of women living with their parents in 1981, for women aged 15-19 and living with parents in 1971. They are categorised according to their housing tenure in 1971. Tenure is divided into two categories: social housing and other tenure. Table 5 shows the same information for women living with their parents in 1991.In the first case (table 4), the row percentages indicate that those living in social housing at the start of the follow-up period are less likely than those with other forms of tenure to be still living in the parental home at the end of follow-up period; in the second case (table 5) this difference has disappeared.

Adding a second cohort
You could rewrite the previous research question in a more general way as follows:
Among women aged 15-19 and still living with their parents at a given census, are those living in social housing more likely than those with other types of tenure to have moved out of the parental home by the time of the following census?

If the census in question is 1971 then this is identical to the first part of the previous question. However, you can ask the same question for the 1981 census. In this way you will need to define a completely new cohort of women (i.e. those who are aged 15-19 in 1981. Your original cohort will be aged 25-29 by the time of the 1981 census.) Doing this allows you to compare results over two separate periods of follow-up. This means that you can see how the association between housing tenure and living with parents has changed over time. Note also that although the two cohorts together cover the period 1971-1991, this is very different to following up a single cohort from 1971 to 1991.

Study subjects for the two-cohort study
For this study it is important to note that you are dealing with two distinct cohorts. For the first part of the study (above) you are looking at women aged 15-19 in 1971, and because you are following these women up until 1981, you also need them to be present in the 1981 census. For the second part of the study, you will have a completely different sample taken from those women aged 15-19 in 1981. Remember that your first cohort are now aged 25-29, so none of them will be in this new cohort. The sample is followed up to 1991, so you also need them to be present at the 1991 census.

Contrast this two-part study with a single longitudinal analysis of women aged 15-19 in 1971, followed up until 1991. Such a cohort would not answer our new research question to determine how the association between housing tenure and living with parents has changed over time.

Table 6 shows the numbers of women in the target age range at each census. As before, visitors and residents of communal establishments have been excluded, to give an idea of the sample sizes you can expect in practice.

Results for the two-cohort study

Tables 4 and 6 show the results for the 1971-1981 cohort and 1981-1991 cohort respectively. In table 4, the subjects are categorised according to their housing tenure in 1971, and in table 6 they are categorised according to their housing tenure in 1981. The row percentages for each table show that while for the period 1971-1981 there is an there is an indication of an association between social housing and remaining in the parental home ten years later (those living in social housing at the start are less likely to be living in the parental home at the end), for the later period (1981-1991) there is no indication of any association.

Tables 7-8 and tables 9-10, add two further cohorts to the study, for the periods 1991-2001 and 2001-2011 respectively. The row percentages for each table show that for both periods there is an indication of an association between social housing at the start of follow-up and living in the parental home at the end of follow-up: those living in social housing at the start are less likely to be living in the parental home at the end, and the effect is stronger than it was in 1971-81.

Adding events

Introducing events

The final study type that is considered here is a study that uses events data (see the Events module for more information on Events data in the LS). The events recorded for LS members include new births[3], births to sample members[4], deaths[5], cancer registrations[6] and migrations out of (or into) England and Wales. This example research question looks at births.

Research question: *Amongst nulliparous[7] women aged 15-19 at the 1981 census, do those in social housing have a greater probability of giving birth within the next five years than those with other types of tenure?*

This question is again looking at housing tenure, but this time it is also introducing birth event data, which is collected between censuses. In order to use a subject's event data, that subject must be traced to the NHSCR. This represents a new constraint on the study population.

Birth data

Live births in the LS are recorded in the LBSM (Live Births to Sample Mothers) file, and still births in the SBSM (Still Births to Sample Mothers) file. You need to know the date of each birth, which is recorded in the variable EDATEBM in both tables (see the Data Dictionary). Using this variable, you can determine whether a birth occurred within the 5-year period of follow-up – in this case from census day 1981 (5th April) to 5 April 1986.

Determining parity (number of births to a woman)

The research question also specifies that the women must be nulliparous. Fertility information was not recorded in the 1981 census and therefore you need to look back through the event records before

[3] Babies born on one of the four LS birthdays, who therefore become LS sample members
[4] Live births that have been registered for sample members who have been traced to the NHSCR.
[5] Deaths have to be registered. The Deaths information in the LS contains the information that is given in the death registration form.
[6] Cancer registrations are collected by the National Cancer Registration and Analysis Service in England and the Welsh Cancer Intelligence and Surveillance Unit. Similarly to births to sample members and deaths of sample members these are linked to LS members through being traced to the NHSCR.
[7] Women who have not previously given birth (live or still births).

1981 to make sure that there have been no births to sample members during the period that they are at risk of giving birth (using the LBSM and SBSM tables). Since there will be very few births to women under 13 years old, you need to look back six years before the 1981 census, when the sample population was 9-13 years old. In the same way that you extracted those women who had given birth during the follow-up period, here you exclude any women that had given birth during the previous six years (i.e. back to 5 April 1975). Again, you can use the variable EDATEBM in the LBSM and SBSM tables and exclude any women whose LS record links to a birth event record where EDATEBM lies between 5 April 1975 (19750405) and 4 April 1981 (19810405).

Study subjects
The sample for this study is taken from all women aged 15-19 who were present at the 1981 census. The sample is then further constrained by the need to know whether each woman gave birth during the next five years.
NB: subjects could die or emigrate during the five years of follow-up after 1981, in which case they would not be at risk of giving birth throughout the whole period. In this example, you will therefore need to exclude these people, although a more sophisticated treatment would be to calculate total person-time at risk. In addition to this, in order to test that these women are nulliparous, you will have to check that there have been no births to sample women in the six years before the 1981 census.

To link a birth to a LS record requires that the LS member is traced at the NHSCR at the time of birth. Therefore, our study subjects need to be traced before 1981. The TRACE variable in the CORE1 table will enable you to see whether a study subject has been traced before 1981 (see the Data Dictionary for more information on this variable).

Table 11 shows the number of women in the required age band at the 1981 census, and how the sample size is successively reduced by the various constraints imposed by the research question (i.e. traced to the NHSCR, nulliparous and not dying or embarking in the following 5 years).
NB: This study will be particularly sensitive to differences in the age distribution between the social housing group and the population as a whole. You would need to account for such differences before drawing any conclusions.

Tables 13, 15 and 17 show the same information for women aged 15-19 in 1991, 2001 and 2011 respectively.

Results
Table 12 shows the results for the study. It shows column percentages, indicating that nulliparous women aged 15-19 and living in social housing in 1981 were twice as likely as those with other forms of tenure to give birth during the following five years.

Tables 14, 16 and 18 show the results for similar analyses for women aged 15-19 and living in social housing in 1991, 2001 and 2011 respectively. Similarly to women aged 15-19 and living in social housing in 1981, the results indicate that nulliparous women aged 15-19 and living in social housing in 1991, 2001 or 2011 were more than twice as likely as those with other forms of tenure to give birth during the following five years.