\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* A Practical Guide to Using Panel Data

\* Simonetta Longhi and Alita Nandi

\* ISER, University of Essex

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Appendix A**

**Understanding Society: the UK Household Longitudinal Study (UKHLS)**

This is a short introduction to ‘*Understanding Society*: The UK Household Longitudinal Study’ (UKHLS) that summarises the main characteristics of the study discussed in Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’. For further details see the user guide:

McFall, Stephanie L. (ed.) (2013) *Understanding Society* – UK Household Longitudinal Study: Wave 1-3, 2009-2012, User Manual. Colchester: University of Essex

You can find the user guide, interactive online documentation, questionnaires, fieldwork and technical documents at <https://www.understandingsociety.ac.uk/documentation>.

*The Survey*

The UKHLS is a multi-purpose household panel survey of a sample drawn from the non-institutionalised residential population of the UK in 2009. It is similar to the British Household Panel Survey (BHPS, discussed in a separate section of this online Appendix) in a number of ways. Some of the key differences of the UKHLS from the BHPS are the much larger sample size, an over-sample of ethnic minorities, health and bio-markers component and a larger geographical spread.

The sample of approximately 30,000 households and 77,000 individuals was drawn from the Postcode Address Small Users File list of domestic addresses (for the Great Britain samples) and the Land and Property Services Agency list of domestic addresses (for the Northern Ireland sample). The sample has several components. One of these is the General Population Sample (GPS): the sub-sample for Great Britain has a clustered and stratified sample design while the sub-sample for Northern Ireland (NI) has a simple random sample design and a selection probability that is twice that of the Great Britain sub-sample. An additional over-sample of around 4,000 ethnic minority households consisting of 13,000 individuals, the Ethnic Minority Boost (EMB) sample, has been selected from high ethnic minority concentration areas of Great Britain. This too has a clustered and stratified design. From the second wave onwards, the sample of households that had not dropped out after the 18th wave of the BHPS becomes eligible for inclusion into the UKHLS, resulting in an additional 6,600 interviewed households consisting of about 16,500 individuals. The different sub-samples are summarised in Table 1 at the end of this document.

Compared to the BHPS, the UKHLS is much more geographically dispersed and the cluster sizes are smaller: 18 addresses selected from each of the 2640 primary sampling units (PSUs). For comparison, in the BHPS sample, on an average, 33 addresses were selected from 400 primary sampling units (250 in the original sample and 75 each in the Scottish and Welsh boost samples).

The UKHLS is an indefinite life panel survey, with no refreshment sample planned as yet. Sample members are interviewed every year across a 24 month fieldwork period resulting in an overlapping fieldwork design. However, given their small sample sizes the GPS-NI and the BHPS sub-samples are interviewed over the first 12 months of the fieldwork period.

*The Survey*

All household members of responding households in wave 1 (except for non-ethnic minority members of EMB households) are considered to be Original Sample Members (OSM). OSMs are followed wherever they go as long as they reside in the UK. Children of OSM mothers are also considered to be OSMs. Others who join OSM households and non-ethnic minority members of EMB households are considered to be Temporary Sample Members (TSM). TSMs are only interviewed as long as they are co-resident with at least one OSM. Any TSM who becomes the father of an OSM child becomes a Permanent Sample Member (PSM). PSMs have the same following rules as the OSMs.

In the UKHLS, data are collected using a similar set of survey instruments as in the BHPS:

* Household and enumeration grid: interviewers collect information about who lives in the household, their relationships to each other and some basic information about them such as marital status, age, sex.
* Household questionnaire: information about the residential property, household expenditures, assets and so on is collected from a knowledgeable adult in the household.
* Adult face-to-face interview questionnaire: all adults (aged 16 years old or over) are asked detailed factual and subjective information about themselves.
* Proxy interview questionnaire: basic factual information is collected about non-responding adults from their spouse, partner or adult child who knows them well.
* Adult self-completion questionnaire: adult respondents are also asked to fill in a self-completion questionnaire which may include questions on sensitive topics.
* Youth self-completion questionnaire: young persons in the household between the ages of 10 and 15 are asked to complete a self-completion questionnaire which includes questions that are particularly important to understand the lives and experiences of young persons such as eating habits, experiences of bullying, computer usage.
* In the UKHLS currently interviews are not conducted by telephone but the BHPS telephone sample members continue to be interviewed by telephone when they join the UKHLS sample. Unlike in the BHPS, the telephone questionnaire in UKHLS is the same as the face-to-face questionnaire.

To allow comparison of ethnic minority groups with each other and the white majority group, an extra five minutes is set aside for questions of particular importance to ethnicity related research. These “extra five minutes” questions are asked of adults in the following sub-samples: EMB sample, a random sub-sample of 500 households in the GPS also referred to as the General Population Comparison (GPC) sample, all ethnic minority adult respondents in the GPS who were living in low ethnic minority concentration areas at wave 1.

*Data files*

Understanding Society data are provided in Stata, SPSS and TAB formats. Information collected in each wave is made available in a separate set of data files: each file corresponding to one data source. Across waves, file names have the same root name with a letter prefix identifying the wave, followed by an underscore: “a\_” for the first wave, “b\_” for the second wave, and so on. For example, all information collected in the household questionnaire in the first wave is provided in the file a\_hhresp while all information collected from household members aged 16 or above in the third wave is provided in c\_indresp. Information about all individuals in responding households, including children and non-respondents are available in w\_indall (here we use w as a placeholder for any wave prefix). To make it easier for users to access fixed information collected at different waves (such as year and country of birth, and highest qualification obtained), this information is stored in an individual level cross-wave file called xwavedat (note the absence of a wave prefix in the file name). A list of file names and their content is provided in Table 2 at the end of this document.

Variable names follow a similar naming convention as file names. For example, the variable identifying the main activity status is called a\_jbstat in the first wave, b\_jbstat in the second wave and so on. All derived or generated variables in the UKHLS have the suffix \_dv. For example, the generated variable for usual monthly pay is called w\_paygu\_dv.

The only variables not to have wave prefixes are the cross-wave unique person identifier pidp, the BHPS cross-wave identifier pid, and the fixed information variables in xwavedat. The unique within-wave household identifier is called w\_hidp. There is no concept of a longitudinal household in UKHLS and so there is no unique cross-wave household identifier.

Within each wave a person can also be uniquely identified by the household identifier and person number (w\_hidp and w\_pno). These variables can be used to merge individual and household level files as well as individual or household level files across waves. For further details see Chapter 5 of Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’.

Missing values in the UKHLS are: -1, -2, -7, -8, -9, -10 and -11. These values represent don’t know (-1), refusal (-2), not asked in proxy interview (-7), valid skips or not applicable (-8), inconsistent or implausible values (-9), no data from BHPS 1-18 (-10) and no data from UKHLS (-11).

Data for all sub-samples (EMB, GPC, BHPS and so on) are provided in the same files and cases from the different sub-samples can be identified by the variable hhorig in cross-wave files and w\_hhorig in any of the wave specific files. See Table 18 of McFall (2013) for a list of the key variables in the dataset.

*BHPS respondents in UKHLS*

From wave 2 onwards all individual level datasets contain the variable pid, the individual cross-wave identifier for BHPS sample members. This can be used to link the UKHLS data of the BHPS sample members to their BHPS data for the previous 18 waves. Hence, BHPS respondents have valid data for both pid and pidp (the same person has two different identification numbers, one for BHPS and one for UKHLS), while UKHLS respondents who were not part of BHPS have missing values for the variable pid. Individuals joining BHPS households for the first time after it became part of the UKHLS also have missing values for pid.

The 18th wave of the BHPS was fielded in September to December 2008 and the subsequent time the sample members were interviewed was as part of the UKHLS from January to December 2010. While some were interviewed after approximately one year, others were interviewed after two years. Panel data methods assume equal time interval between observations; if you are building a panel data with the BHPS that includes the UKHLS components, it may be preferable to treat the 2nd wave of the UKHLS as 20th wave of the BHPS instead of the 19th.

*Identifying other household members*

The UKHLS also provides indicator variables for identifying the spouse, partner, parents and grandparents. Other family members can be identified using the file w\_egoalt. For details see Chapter 6 of Longhi and Nandi (2014) ‘A Practical Guide to Using Panel Data’.

*Sample Weights*

As different sub-samples of the UKHLS were selected with different selection probabilities and there is non-response and attrition, weights are needed to produce unbiased population estimates based on sample statistics. Different weights are provided for different types of analyses. The data provide cross-sectional and longitudinal weights for each wave which correct for unequal selection probability and non-response or attrition, but also weights that only correct for the sample design. The sampling design variables which represent the strata and primary sampling units are also available. Some are to be used for individual level analysis while others for household level analysis. See McFall (2013) for details on computation of these weights and guidance on which weights to use for specific types of analyses.

*History Files*

In the UKHLS most of the data are collected by prospective methods. Exceptions are the marital, fertility and employment histories asked when respondents are interviewed for the first time, which are collected by retrospective method.

The UKHLS collects information on the marital, fertility and employment histories of respondents before they joined the panel, as well as changes in these domains between interviews. The initial employment history was collected for only a quarter of the sample in wave 1 and is collected for the rest of the sample in wave 5. These are multi-level files where each row is identified by the individual and the specific spell and include information on the type of spell, start and end dates and whether the spell is still ongoing (see Table 2 at the end of this document, and the UKHLS website for details).

*Health and Biomarkers*

Health and bio-markers were collected by nurses five months after the wave 2 interview from a sub-sample of the wave 2 GP sample, and after the wave 3 interview from a sub-sample of the BHPS sample. The sub-samples that were eligible were people residents in Great Britain who gave a full-interview in English at these waves. During the nurse visits, subject to consent, a number of health information was collected from adult respondents such as height, weight, grip strength, lung function test, blood spot, and so on. This resulted in nurse health assessments for around 20,000 individuals and health assessments as well as blood samples for around 13,000 individuals. These data are available as a separate set of data files which can be combined with the UKHLS main data using the variable pidp. The data structure and naming conventions of the data files and variables are similar to the main survey files. These data files have a suffix \_ns. For more information see the Health Assessment user guide and online documentation at <https://www.understandingsociety.ac.uk/documentation>.

*Innovation Panel*

A household panel survey called the Innovation Panel (IP) with a sample of approximately 1,500 households drawn from Great Britain is fielded one year prior to the UKHLS main survey. For the first few waves the main purpose of this survey was to inform UKHLS on survey methodological issues. Nowadays it is increasingly used for research on more general survey methodology issues related to longitudinal surveys. The innovation panel data are available as a separate set of data files. The data structure and naming conventions of the data files and variables are similar to the main survey files. These data files have a suffix \_ip. The innovation panel should not be used together with the main sample for analysis. For further information on the IP see the user guide and online documentation available at <https://www.understandingsociety.ac.uk/documentation>.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 1: Description of Understanding Society samples** | | | | | |
|  | General Population Sample | | Ethnic Minority Boost Sample | British Household Panel Survey  (wave 2) | Total |
|  | Great Britain Component | Northern Ireland Component |
| Issued Households | 48,144 | 2,395 | 44,769 | 8,992 | 104,300 |
| Responding Households | 24,797 | 1,292 | 4,080 | 6,692 | 36,861 |
| Enumerated Individuals in Responding Households | 60,597 | 3,351 | 13,361 | 16,562 | 93,871 |
| Adult Respondents (excluding proxy respondents) | 39,050 | 1,997 | 6,685 | 11,260 | 58,992 |
| Proxy Respondents | 2,536 | 91 | 635 | 450 | 3,712 |
| Telephone Respondents | - | - | - | 326 | 326 |
| Youth Respondents | 3,783 | 212 | 904 | 1,117 | 6,016 |
|  | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 2: Description of Understanding Society data files** | | | |
| **Contains** | **File Name** | **Responses provided by** | **Each row of observation is uniquely identified by** |
| *Substantive information* |  |  |  |
| Information on the household and enumeration grid | w\_indall | Any adult in the household | pidp or  w\_hidp w\_pno |
| Responses from the household questionnaire | w\_hhresp | Knowledgeable adult in the householda | w\_hidp |
| Responses from adult individual interviews (face-to-face, telephone, self-completion and proxy). Also includes interviewer remarks about the interview process | w\_indresp | Adults (aged 16 and over) AND interviewers | pidp or  w\_hidp w\_pno |
| Responses from adult (face-to-face) individual interviews. |  | Adults (aged 16 and over) |  |
| Information of all income sources since last interview, one row for each income source of each individual | w\_income |  | pidp w\_fiseq or  w\_hidp w\_pno w\_fiseq |
| History of marriages, cohabitation and employment statuses before the start of the survey. One row for each marriage, cohabitation and employment spell of each individual | w\_marriage(a)  w\_cohab(a)  w\_empstat(a) |  | w\_hidp w\_pno w\_marno  w\_hidp w\_pno w\_cohabno  w\_hidp w\_pno w\_spellno |

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 2: Description of Understanding Society data files (continued)** | | | |
| **Contains** | **File Name** | **Responses provided by** | **Each row of observation is uniquely identified by** |
| Information about natural, adopted/step children (including non-resident children) of the adult respondent | w\_child(a)  w\_natchild(a)  w\_newborn(b,c)  w\_parstyle(c)  w\_chmain(c) |  | pidp or  w\_hidp w\_pno  w\_hidp w\_pno w\_childno  pidp w\_newchno or  w\_hidp w\_pno w\_newchno  pidp w\_childpno or  w\_hidp w\_pno w\_childpno  pidp w\_ absparno or  w\_hidp w\_pno w\_absparno |
| Responses from the youth questionnaire | w\_youth | 10-15 year olds in the household | pidp or  w\_hidp w\_pno |
| *Sampling information and Paradata* | |  |  |
| Sampling information and information from the ARF (includes information on non-responding households) | w\_hhsamp | Survey organisation and interviewer | w\_hidp |
| Household location and interview outcome information about every person in fielded households | w\_indsamp | Interviewer | pidp w\_finloc |
| Interview outcome at each issue | w\_issue | Interviewer | w\_hidp w\_issueno |
| Interview outcome at each call | w\_callrec | Interviewer | w\_hidp w\_issueno w\_callno |
| Interviewer information | xivdata | Survey organisation | intnum |
| *Derived files* |  |  |  |
| Relationship between every pair of household members | w\_egoalt |  | pidp w\_apidp or  w\_hidp w\_pno w\_apno |
| Fixed information about everyone in every enumerated households | xwavedat |  | pidp |
| Interview outcome in every wave | xwaveid |  | pidp |