**Understanding Society Teaching Datasets**

Stephen McKay, Michael Adkins and Helen Williams.

**Abstract**

This note describes datasets produced for teaching purposes, and based on the new UK household panel study, *Understanding Society*. We set out the background to the study, and describe the new datasets which are both cross-sectional and longitudinal.

**Background**

Between 2012 and 2014 a research project was run at the University of Birmingham to help enhance the capacity of social science undergraduates to understand and use numeric data in their studies. Part of the aim of the underlying research project was to make it easier to handle data from *Understanding Society*, the relatively new longitudinal household study in the UK[[1]](#footnote-1). It is possible for registered users to download the full versions of these datasets from the UK Data Service (the relevant data link is <http://discover.ukdataservice.ac.uk/catalogue/?sn=6614>) but there are important challenges to consider. Would-be users of these datasets are confronted by 44 data files, around 30 files of documentation and 44 data dictionaries (the last of these in SPSS format no matter the format of the download), and that merely covers the first three waves of a developing study. Moreover the data are typically only made available in two commercially-linked formats (Stata and SPSS) or as tab-delimited files (with limited data labelling). Against this, the free software programme R[[2]](#footnote-2) seems to be increasing as a means of analysis. For these reasons we developed new data files[[3]](#footnote-3) that permit learning of statistical approaches using either individual datasets for each of the three waves, or longitudinal analysis that combines those datasets in two different ways – the so-called wide and long formats.

**The datasets**

In deriving these teaching datasets there was designed to be a core of common variables from the first three waves, plus some extra variables of intrinsic interest that only appear in particular waves. The research term was mostly focused on political science and public policy. Whilst this may have influenced the selection of particular variables, these datasets should be sufficiently diverse to be of interest across a wider range of social sciences. The dataset for the first wave is the most extensive, and may be best suited to introductory courses at least as a starting point. In the Appendix, Table 4 gives a complete list of the variables included, with their variable labels and pattern of inclusion within the datasets in the first three waves. The first wave is A, the second is B, and the third is C. Generally speaking these cover the years 2009-10, 2010-11 and 2011-12.

The richness of the data, and in particular the combination of ratio/interval data as well as categorical/ordinal data, makes for an effective dataset from which to teach quantitative methods of all kinds. A distinctive feature is the provision of datasets in native R format in addition to those in Stata or SPSS statistical format.

**Data structure and linking**The 'parent' dataset (SN 6614 at the UK Data Service) is the definitive guide to the variables, sampling, etc.. A key feature is that variables are prefixed a\_ when pertaining to the first wave, b\_ for the second wave, and c\_ for the third wave. When translated into R format, the underscore character is not permitted and so these are replaced by full stops, so that a\_sex becomes a.sex, for instance. There is an identifier for each person that stays constant through the panel, ‘pidp’, and which appears in each dataset. This linking variable is necessary for linking information on the same people over time. It could also be used to bring in further data if required. Individuals *within* each wave are also identified by a household identifier and a person number (e.g. a\_hidp and a\_pno in the first wave), which can be used to look at within-household analysis or to aggregate data to a household level.

An important part of analysis is learning about generalising from the sample to the population. Often the sample differs in important ways from the population because certain groups are less likely to agree to be interviewed or even located (those in urban settings compared to rural areas, for instance), or because some groups are sample in greater numbers because they are of particular interest (such as members of ethnic minority groups). The topic of weights is an important one, not always covered in detail in introductory statistical courses. For each data we have included the main adult weight variable, which has been renamed as weight\_xs (weight for the cross-section).

**Dataset structures for longitudinal analysis**

There is one dataset for each wave of the dataset, containing around 150-200 variables in each case, rather than the several thousand of the original datasets. For simplicity we also only select fully responding households with full individual adult interviews. When merging together information from different waves we use two alternative approaches. The first is the ‘wide format’, where each line represents a respondent, and the variables for each wave appear on that line. This makes it easier to compare, say, employment status over time. The alternative is the ‘long format’, where each row represents a particular year’s data for a particular respondent. An example should help. Let us assume that respondent #1 is married at each of the first two waves, and then separated for the third wave; respondent #2 is divorced at wave 1, and then fails to participate in the second two waves. We can represent these two individuals’ data in the following two ways (see Table 1 and Table 2). The variable indicating marital status is marstat, with the relevant prefixes for each wave and data format.

Table 1 Panel data in wide format

|  |  |  |  |
| --- | --- | --- | --- |
| **Respondent** | **a\_marstat, in Stata or SPSS (a.marstat, in R)** | **b\_marstat (b.marstat)** | **c\_marstat (c.marstat)** |
| 1 | Married | Married | Separated |
| 2 | Divorced | (missing) | (missing) |

Table 2 Panel data in long format

|  |  |  |
| --- | --- | --- |
| **Respondent** | **Wave** | **Marstat** |
| 1 | 1 | Married |
| 1 | 2 | Married |
| 1 | 3 | Separated |
| 2 | 1 | Divorced |

Those relatively new to panel data often find it easier to work with the wide data version first. This permits some simple analysis of transitions between particular waves – such as the first and last observations. A simple cross-tabulation may show how many workers in wave 1 are still employed by wave 3, for instance, or how many single people now live with someone. Going beyond that simple approach, such as to try to capture *all* relevant transitions for each new wave, presents greater problems with such data. Indeed in due course many analysts become more comfortable with the long format. This happens to be the most suitable arrangement for more advanced statistical models of various kinds, and is also a more efficient means of storing the data, particularly when there are complex patterns of non-trivial attrition. By using functions that harvest data in the previous data row (or the one before the previous, or the next) it is straightforward to measure the extent of transitions of various kinds on an annual basis.

Overall, there are advantages and disadvantages to each mode of storing panel data. Fortunately if an analyst needs data needs to be in the ‘other’ format then software packages often permit relatively easy ways to transform data from one to the other (such as Stata’s reshape command; SPSS’s data restructure wizard or syntax commands varstocases and casestovars; R’s reshape library with commands melt and cast).

The wide datasets contain the variable ‘partpatt’ – meaning participation pattern – that sets out the specific waves in which people took part (see Table 3). Overall there are over 63,000 different individuals in the dataset, with around one-third taking part in each and every wave. Hence those becoming used to looking at longitudinal data immediately face issues of which groups to analyse, owing to the extent of ‘missing’ data. Sometimes data is missing for particular reasons – those aged 15 in wave 1 would not have been eligible for interview at that stage, but probably would have been by wave 2. Some respondents will die between waves, and this is more likely for older respondents. It is not just a matter of people making a decision to stop participating in particular waves, or moving without a robust means of locating them, although these are often the reasons for losing people from panel studies,

Table 3 Panel data pattern of participation

|  |  |  |
| --- | --- | --- |
| Participation pattern | Number of respondents | Per cent of respondents |
| 111 | 21,442 | 33.98 |
| 11- | 3,920 | 6.21 |
| 1-1 | 3,399 | 5.39 |
| 1-- | 7,008 | 11.1 |
| -11 | 13,072 | 20.71 |
| -1- | 2,441 | 3.87 |
| --1 | 11,826 | 18.74 |
| Total | 63,108 | 100 |

The appendix now sets out the variables included within the reduced datasets, and indicates in which wave the variables appear.

**Appendix 1: List of variables included in the datasets**

Table 4 List of variables and their appearance by survey wave.

| **Variable** | **Variable label** | **Pattern of inclusion** |
| --- | --- | --- |
| ***Common core of variables in each wave*** | |  |
| Hidp | household identifier (public release) | ABC |
| pno | person number in household grid | ABC |
| pidp | cross-wave person identifier (public release) | ABC |
| sex | Sex | ABC |
| dvage | age for whole sample, from birth or ageif | ABC |
| mvever | lived at address whole life | ABC |
| mvyr | year moved to current address | ABC |
| mlstat | present legal marital status | ABC |
| ukborn | born in uk | ABC |
| plbornc | country of birth | ABC |
| yr2uk4 | year came to Britain | ABC |
| citzn1 | uk citizen | ABC |
| citzn2 | citizen of country of birth | ABC |
| citzn3 | citizen of other country | ABC |
| qfhigh | highest qualification | ABC |
| pacob | country father born in | ABC |
| macob | country mother born in | ABC |
| natid1 | English | ABC |
| natid2 | Welsh | ABC |
| natid3 | Scottish | ABC |
| natid4 | northern irish | ABC |
| natid5 | British | ABC |
| natid6 | Irish | ABC |
| natid97 | Other | ABC |
| racel | ethnic group | ABC |
| oprlg | whether belong to a religion | ABC |
| oprlg0 | religion brought up in: e/s/w | ABC |
| oprlg0ni | religion brought up in: ni | ABC |
| oprlg1 | religion: e/s/w | ABC |
| nirel | religion: ni | ABC |
| niact | religion active: ni | ABC |
| jbsect | private company | ABC |
| jbsectpub | non-private organisation | ABC |
| jbhrs | no. of hours normally worked per week | ABC |
| jbttwt | minutes spent travelling to work | ABC |
| basrest | estimated amount - hourly basic pay rate | ABC |
| finnow | subjective financial situation – current | ABC |
| finfut | subjective financial situation – future | ABC |
| vote1 | supports a particular political party | ABC |
| vote2 | closer to one political party than others | ABC |
| vote3 | party would vote for tomorrow | ABC |
| vote4 | which political party closest to | ABC |
| vote5 | strength of support for stated party | ABC |
| vote6 | level of interest in politics | ABC |
| drive | respondent has driving license | ABC |
| mobuse | has mobile phone | ABC |
| netuse | frequency of using the internet | ABC |
| nch14resp | number of children under 15 resp is responsible for | ABC |
| nnatch | number of biological children in household | ABC |
| nadoptch | number of adoptive children in household | ABC |
| vote3\_all | party would vote for | ABC |
| vote4\_all | party supported | ABC |
| prfitb | total personal income | ABC |
| prfitbw | total personal income weekly | ABC |
| prfitba | total personal income annually | ABC |
| marstat | legal marital status | ABC |
| livesp | living with spouse | ABC |
| livewith | living as part of a couple in household | ABC |
| employ | in paid employment | ABC |
| respf16 | whether father of child under age 16 in hh | ABC |
| respm16 | whether mother of child under age 16 in hh | ABC |
| ioutcome | final outcome code | ABC |
| ivfio | individual response outcome | ABC |
| mastat\_dv | De facto marital status | ABC |
| agegr10\_dv | Age group: 10 year intervals | ABC |
| hiqual\_dv | Highest educational qualification | ABC |
| jbft\_dv | Full or part-time employee | ABC |
| jbseg\_dv | Current job: Socio-economic Group | ABC |
| jbrgsc\_dv | Current job: Registrar General's Social Class | ABC |
| jbnssec5\_dv | Current job: Five Class NS-SEC | ABC |
| hhresp\_dv | Household response status | ABC |
| country | Country of residence | ABC |
| gor\_dv | government office region | ABC |
| urban\_dv | Urban or rural area, derived | ABC |
| fimngrs\_dv | personal income – gross | ABC |
| fimnlabgrs~v | labour income – gross | ABC |
| weight\_xs | Cross-sectional adult main interview weight | ABC |
| wave | wave 1, 2 or 3 | ABC |
|  |  |  |
| ***Wave-specific questions*** | |  |
| payruk | father lived in uk | A-- |
| payruk1 | year father moved to the uk | A-- |
| mayruk | mother lived in uk | A-- |
| mayruk1 | year mother moved to the uk | A-- |
| pgprob | country father's father born in | A-- |
| pgmrob | country father's mother born in | A-- |
| paid | father's ethnic group | A-- |
| spaid | strength of identification with father's ethnicity | A-- |
| maid | mother's ethnic group | A-- |
| smaid | strength of identification with mother's ethnicity | A-- |
| britid | importance of being british | A-- |
| englang | english is first language | A-- |
| engspk | difficulty speaking day to day English | A-- |
| spkdif | degree of difficulty speaking day-to-day English | A-- |
| engtel | difficulty speaking english on the phone | A-- |
| teldif | degree of difficulty speaking english on phone | A-- |
| engread | difficulty reading English | A-- |
| readdif | degree of difficulty reading English | A-- |
| engform | difficulty completing forms in English | A-- |
| formdif | degree of difficulty completing forms in english | A-- |
| oprlg2 | attendance at religious services | A-- |
| oprlg3 | religion makes a difference to life | A-- |
| mabroad | has lived abroad | A-- |
| mnotherc | number of countries lived in | A-- |
| moveage | age respondent moved to uk | A-- |
| mlivedist | current home: distance from first home/age 14 | A-- |
| lcmarm | month of current marriage | A-- |
| lcmary4 | year of current marriage | A-- |
| mpno | person number of spouse | A-- |
| lcmcoh | cohabited before current marriage | A-- |
| lcmcbm | month began cohabiting before current marriage | A-- |
| lcmcby4 | year began cohabiting before current marriage | A-- |
| lcmspm | month separated | A-- |
| lcmspy4 | year separated | A-- |
| nmar | number of marriages | A-- |
| lcoh | ever cohabited | A-- |
| lncoh | number cohabiting partners | A-- |
| sf1 | general health | A-- |
| lvrel1 | Mother | A-- |
| lvrel2 | Father | A-- |
| lvrel3 | son(s)/daughter(s) | A-- |
| lvrel4 | brothers/sisters | A-- |
| lvrel5 | Grandchildren | A-- |
| lvrel6 | Grandparents | A-- |
| lvrel7 | great grandchildren | A-- |
| lvrel8 | great grandparents | A-- |
| lvrel96 | none of these | A-- |
| maage | mother's age | A-- |
| paage | father's age | A-- |
| parmar | parents live together in same household | A-- |
| ohch16 | children under 16 not living in hh | A-- |
| seekid | how often contact child outside hh | A-- |
| wekid | child outside hh stays with r regularly | A-- |
| envhabit1 | environmental habits: tv | A-- |
| envhabit2 | environmental habits: lights | A-- |
| envhabit3 | environmental habits: water | A-- |
| envhabit4 | environmental habits: heating | A-- |
| envhabit5 | environmental habits: packaging | A-- |
| envhabit6 | environmental habits: recycled paper | A-- |
| envhabit7 | environmental habits: shopping bags | A-- |
| envhabit8 | environmental habit: public transport | A-- |
| envhabit9 | environmental habit: short journeys | A-- |
| envhabit10 | environmental habit: car share | A-- |
| envhabit11 | environmental habit: fewer flights | A-- |
| swemwbs\_dv | Short Warwick-Edinburgh Mental Well-being Scale | A-- |
|  |  |  |
| volun | volunteer in last 12 months | -B- |
| volfreq | frequency of volunteering | -B- |
| volhrs | hours spent volunteering in last 4 weeks | -B- |
| chargv | donated money to charity | -B- |
| charfreq | frequency donated to charity | -B- |
| charam | amount given to charity last 12 months | -B- |
| hubuys | who does the grocery shopping (couples) | -B- |
| hufrys | who does the cooking (couples) | -B- |
| humops | who does the cleaning (couples) | -B- |
| huiron | who does the washing/ironing (couples) | -B- |
| hupots | who does the gardening (couples) | -B- |
| hudiy | who does the diy jobs (couples) | -B- |
| husits | who is responsible for childcare | -B- |
| huboss | household financial decisions | -B- |
| howlng | hours per week on housework | -B- |
| vote7 | voted in last general election | -B- |
| vote8 | party voted for in last general election | -B- |
|  |  |  |
| poleff1 | qualified to participate in politics | --C |
| poleff2 | better informed about politics | --C |
| poleff3 | public officials don t care | --C |
| poleff4 | don t have a say in what government does | --C |
| newsmain | main source of news | --C |
| paperm2 | most frequent newspaper | --C |
| tvm2 | most frequent tv channel | --C |
| netm2 | most frequent news website | --C |
| tvhours | hours of tv per weekday | --C |

1. University of Essex. Institute for Social and Economic Research and NatCen Social Research, *Understanding Society: Waves 1-3, 2009-2012* [computer file]. 5th Edition. Colchester, Essex: UK Data Archive [distributor], November 2013. SN: 6614. [↑](#footnote-ref-1)
2. The home page, with documentation and the facility to download, is at <http://www.r-project.org/>. [↑](#footnote-ref-2)
3. R can read data that is in Stata or SPSS format, using the library called ‘foreign’, but it is generally slow when reading larger datasets and a range of technical problems can occur. Hence the addition of specific R data workspaces within the data uploaded here. [↑](#footnote-ref-3)