

# Data Management and Frontiers in Survey Research

*Professor Vernon Gayle  
University of Stirling  
& ISER University of Essex*



## Structure of this talk

- Why bother with data management?
- Complex survey data is not going to disappear
  - BHPS; MCS; UKHLS;
- “Beyond this place there be dragons”

## Why Bother?

The payoff from discipline...

Work can be replicated

Work can be transferred (e.g. personnel changes)

Work ultimately becomes quicker

- Getting into a muddle less often
- Getting out of a muddle quicker

Increasing important in research governance and openness and ethics

3

## Why Bother?

Not just a hobbyhorse...

Complex survey data is not going to disappear we are faced with

- Many variables
- Multiple files
- Complicated data structures

4

Cross-Wave Record Types	Record Description	Variable List
XWAVED	Contains information for matching individuals between Waves	X
XWLSTEN	Contains information on the latest known sample status of individuals	X
XWAVEDAT	Contains substantive data about individuals which is fixed and only measured once in the panel.	X

  

Record Type	Record Description	Waves
WHHSAMP	Contains household-level data for issued households	A B C D E F G H I J K L M N O P Q
WINDSAMP	Contains individual-level data for issued households	- B C D E F G H I J K L M N O P Q
WINDALL	Contains enumerated individuals' data	A B C D E F G H I J K L M N O P Q
WHHRESP	Contains household-level data for respondent households	A B C D E F G H I J K L M N O P Q
WINDRESP	Contains individual-level data for respondents	A B C D E F G H I J K L M N O P Q
WJOBHIST	Contains information from the employment history	A B C D E F G H I J K L M N O P Q
WJOBHSTD	Contains information from the employment history, based on dependent interviewing	- - - - - - - - - - - P Q
WINCOME	Contains income and payment data	A B C D E F G H I J K L M N O P Q
WGOALT	Provides a mechanism for identifying the relationship of each individual in a household to all others	A B C D E F G H I J K L M N O P Q
WYOUTH	Contains the responses to the Young persons questionnaire	- - - D E F G H I J K L M N O P Q
WLIFEJOB	Contains information about jobs held in employment spells	- - C - - - - - - - - - - -
WMARRIAG	Contains one record for each reported legal marriage	- B - - - - - - - K L - - - - -
WCOHABIT	Contains data about each cohabitation spell outside legal marriage	- B - - - - - - - K L - - - - -
WCHILDAD	Contains information about adopted and/or step-children	- B - - - - - - - K L - - - - -
WCHILDNT	Contains information about natural children	- B - - - - - - - K L - - - - -
WLIFEMST	Contains information about employment status spells	- B - - - - - - - K L - - - - -
WCHILD	Contains information about the children of the respondent	- - - - - - - - - L M - - - - Q

5



**Understanding Society**  
THE UK HOUSEHOLD LONGITUDINAL STUDY

# Understanding Society: the UK Household Longitudinal Study



<http://www.understandingsociety.org.uk/>

6

## UK HLS Background

- *Understanding Society* is a longitudinal study based on a household panel design
- Basic design similar to BHPS
- Target sample size of 40,000 households – largest Household Panel Survey
- Main fieldwork started in January 2009

7

## Some key features of Understanding Society

- Very large sample size proposed (40K households)
- Large sub samples(4,000 Scottish households)
- Representative sample of whole population (all ages)
- Multi-purpose multi-topic design to meet a wide range of disciplinary and inter-disciplinary research needs

8

## *Understanding Society* Sample

- Approx. 27,000 households - The fieldwork for this sample commenced in January 2009
- A boost ethnic minority sample, focussed on five main ethnic minority groups, comprising 4,000 households
- Incorporating the BHPS sample of approximately 8,400 households
- An Innovation Panel of 1500 households to enable methodological research (panel began in January 2008)

9

## UK HLS Opportunities

- Starting again, compared with BHPS, an opportunity to review activities and see which are worthwhile to continue, which not
- Focus on new research issues
- Opportunities for mixed methods:
  - Data linkage admin, organisation, spatial
  - Bio-markers and health indicators
  - Qualitative data
  - Other non-standard data: diaries, visual, audio

10



## *Understanding Society: the UK Household Longitudinal Study*



<http://www.understandingsociety.org.uk/>

11

## Millennium Cohort Study (MCS)

- 2000/1 Birth Cohort
- England
  - Stratified
  - Ethnic minority stratum
  - Disadvantage stratum
  - Advantaged stratum(based on geographical wards)

12

## Millennium Cohort Study (MCS)

- 2000/1 Birth Cohort
- Wales, Scotland and N.I.
  - Stratified
  - Disadvantage stratum
  - Advantaged stratum(based on geographical wards)

Plewis, I. (2007) The Millennium Cohort Study: Technical Report on Sampling, IOE

[http://www.cls.ioe.ac.uk/core/documents/download.asp?id=875&log\\_stat=1](http://www.cls.ioe.ac.uk/core/documents/download.asp?id=875&log_stat=1)

13

And now for something relating to  
complexity...

14

## A Comment

- Most statistical software packages do not calculate standard errors for weighted data correctly
- If you require weighted analyses then move from SPSS to Stata
  - *the svy suite is specially designed for survey data*

15

## Comparing Stata & SPSS

An example from the BHPS – where N.I. is a simple random sample

```
use "D:\home\vgayle\BHPS_weights\kind2.dta", clear  
numlabel _all, add
```

```
svyset kpsu [pweight= kxrwtk2], strata(kstrata) ///  
    singleunit(scaled)
```

Without the the 'singleunit (scaled)' stata will not estimate the standard errors and reports the following note

```
"Note: missing standard errors because of stratum with single
```

16

```

.svy:mean kpaygu, over(kregion)
(running mean on estimation sample)

Survey: Mean estimation

Number of strata = 121      Number of obs = 8698
Number of PSUs = 399      Population size = 8063.42
Design df = 278

     _subpop_1: kregion = 17. wales
     _subpop_2: kregion = 18. scotland
     _subpop_3: kregion = 19. northern ireland
     _subpop_4: kregion = 20. england
-----+-----
      Over |      Linearized
           |      Mean Std. Err. [95% Conf. Interval]
-----+-----
kpaygu    |
  _subpop_1 | 1242.017 29.55327 1183.841 1300.194
  _subpop_2 | 1378.118 33.89045 1311.404 1444.833
  _subpop_3 | 1249.98      .      .      .
  _subpop_4 | 1446.1 20.67361 1405.403 1486.797
-----+-----

Note: variance scaled to handle strata with a single sampling unit.

.
end of do-file

* SPSS Comparison

      mean          s.e.          lower          upper
Wales 1242.017472    29.43089575    1184.081753    1299.95319
Scotland 1378.118036    33.75012108    1311.679776    1444.556296
N.I. 1249.980248      0    1249.980248    1249.980248
England 1446.099938    20.58800068    1405.571759    1486.628117

```

17

## General Comments on Survey Weights

- Good practice to use weights
- Appropriate weight depends on your specific analysis
- For certain analyses weights may not be available
  - Think about sub-optimal weights
  - Think about constructing your own weights
- In practice many analysts use weights for descriptive statistics but do not use them in multivariate analyses
  - *BUT ALL GOOD RESEARCHERS CONSIDER THE IMPLICATIONS OF NON-RESPONSE FOR THEIR ANALYSIS*

18

## “Beyond this place there be dragons”

We anticipate...

- Increased interest in longitudinal data analyses
  - A move away from the simple ‘wide’ variable-by- case matrix to ‘long’ formats
  - Handling balanced and unbalanced data

19

## “Beyond this place there be dragons”

We anticipate...

- Increased interest in cross-national and comparative analyses requiring managing multiple data resources
  - Practical problems of documentation etc.
  - Harmonisation of variables
  - Standardisation of measures (reflecting changing distributions)

20

## “Beyond this place there be dragons”

We anticipate...

- A move towards currently less mainstream (and often more advanced techniques)

- Generalized Linear Latent and Mixed Models (GLLAMM)
- [Structural Equation Modelling (SEMs)]

Many of these techniques require “organising” or “preparing” data in a specific way before analysis<sup>21</sup>

## “Beyond this place there be dragons”

An example...

When fitting a latent class model to attitudinal data in GLLAMM, you will be modelling a ‘design matrix’ rather than the conventional ‘raw’ data set

The design matrix is the format of how the data are ‘set up’ for the modelling

Andrew Pickles states that the design matrix is the leap into advanced statistics!

22

## “Beyond this place there be dragons”

We anticipate...

- A move towards currently heterodox techniques
  - Geometric data analysis (e.g. correspondence analysis)
  - Propensity score matching (more tomorrow)
  - Social network analysis
  - (Quasi) experimental approaches

Many of these techniques require “organising” or “preparing” data in a specific way before analysis<sub>23</sub>

## Conclusion

At a practical level we must engage with data management when working with social surveys

- Number of variables
- Data files structure
- Survey designs
- Emerging techniques

## Conclusion

Maintaining high standards and good practices in data management is integral to working at the frontiers of social survey analysis

25