Data sources and Data structure: Panel data

Paul Lambert
Stirling University

Prepared for "Longitudinal Data Analysis for Social Science Researchers" , Stirling University, 2-6th September 2006

Data sources and data structure: Panel data

Large and complex data

Longitudinal data analysis in the social sciences is fundamentally complicated by a number of consistent features to longitudinal data resources:

- They are **LARGE**

and

- They are **COMPLEX**

Moreover,

- **PANEL DATA** is the largest and most complex..

The nature of 'large and complex' longitudinal data: complicating the variable by case matrix

Cases ↓	**← Variables →**							
1	1	17	1.73	A	.	.	.	.
2	1	18	1.85	B	.	.	.	.
3	2	17	1.60	C	.	.	.	.
4	2	18	1.69	A	.	.	.	.
.	.	.	.	.	.	.	.	.
N								

Large and complex =

⇒ **Complexity in:**

i. Multiple points of measurement
ii. Lots and lots of cases and variables
iii. Multiple hierarchies of measurement
iv. Sample collection and weighting
v. Multiple data sources

i) Multiple measurement points

*Longitudinal = information collected at or referring to multiple time points; linking data from different time points in a **comparable** manner*

- **Repeated cross-section:** same information at different time points, different cases
- **Panel or cohort:** several records via repeated contact for each individual
  - Social science: **'unbalanced' panel**
- **Event history**: nature of 'episodes' or 'events' and their duration
  - **Censoring**: if event fully observed

Repeated cross-sectional dataset

Survey	**Person**	← **Person-level Vars** →			
1	**1**	1	38	1	1
1	**2**	2	34	2	2
1	**3**	2	6	-	-
2	**4**	1	45	1	3
2	**5**	2	41	1	1
3	**6**	1	20	2	2
3	**7**	1	25	2	2
3	**8**	1	20	1	1
N_s=3	N_c=8				

7

Unbalanced panel dataset, long format

Wave	Person				
1	**1**	38	5	1	1,500
1	**2**	34	4	1	1,500
1	**3**	6	-	9	1,500
2	**1**	39	5	1	1,610
2	**2**	35	2	1	1,610
3	**1**	40	5	1	1,640
3	**2**	36	3	1	1,640
3	**3**	8	-	9	1,640
N_w=3	N_p=3				

8

Unbalanced panel, wide format

Person	**Sex**	Age_97	V_97	V_01	V_05
1	**2**	18	2	-1	-9
2	**2**	67	1	1	1
3	**1**	49	2	3	3
4	**1**	36	2	2	-1
5	**2**	34	3	-9	2
N_p=5		*N_w=3*			

Event history dataset

Person	Event	Duration	Start		
1	**1**	3	1962	1	4
1	**2**	26	1965	2	1
2	**1**	30	1958	1	5
3	**1**	7	1986	1	7
4	**1**	5	1948	1	2
4	**2**	10	1950	6	-
4	**3**	30	1960	1	2
N_p=4	N_e=3				

More complexities: (ii) cases & variables

- **Lots of cases**
  - Impossibility of visual review of data
  - False confidences – large n v's survey data quality
- **Lots of variables**
  - Recoding / dummy indicators / indexes
  - Alterations in variables over time (quality and quantity)
- ***Longitudinal variable management***
  - *Variables spread across multiple data files*
  - Trying to construct comparable longitudinal variables:
    - 'Universalist' v's 'Specific' approaches
    - Contextualisation through multivariate analysis

iii) Multiple hierarchies of measurement

Independence between cases preferred, but we often have data from:

- Both individuals and households; schools and pupils; people and local districts and regions
- Individuals with repeated measures at different times

- **Strategies:**
  - (Separate VxC matrix for each level – 'structural breaks')
  - Merged VxC matrix at lowest level, with hierarchical analysis:
    - **Similarity relations**
      - Hierarchical structural or fixed effects
      - Hierarchical clustering or random effects *(eg random effects estimators; fixed effects estimators; multilevel models)*
    - **Dependence relations** (relations between cases)
      - link data between related cases
      - *Dynamic / lag models in panel data*

Illustration: Hierarchical dataset

Cluster	Person	← Person-level Vars			→
1	**1**	1	38	1	1
1	**2**	2	34	2	2
1	**3**	2	6	-	-
2	**1**	1	45	1	3
2	**2**	2	41	1	1
3	**1**	1	20	2	2
3	**2**	1	25	2	2
3	**3**	1	20	1	1
n1=3	n2=8				

13

iv) Sample collection / weighting

- Sample may have been clustered, stratified
  - 'Multistage cluster'
- Longitudinal
  - Attrition
  - Censoring
  - Study design priorities changing over time
- Limitations of longitudinal sample weights:
  - Complex in application / not suited to all analysis techniques (eg Stata functionality)
  - Limited input – eg simple demographic differences
  - Attrition weights usually only apply to re-constructed balanced panels (throw out 'unbalanced' data)

v) Multiple data sources

- Many linkages between files
  - *Eg linking data on same cases as collected in different years; linking individual level and household level files*
- Many complexities: what want to link to what
- **Essential to keep records**

➢ *Need to use log command (syntax) files!!*

Summary: Complex panel data

- Physical linkages between different datasets
- Long and wide formats (see below)
- Likely to incorporate multiple other complexities:
  - Missing data
  - Sampling weights
  - Household clusterings

Data sources and data structure: Panel data

1. Structural features of panel data
2. General purpose panel studies
3. Cohort and follow-up studies
4. Issues in working with panel data

2) General purpose panel studies

- **Household panels**
  - PSID, BHPS, others
  - Longitudinal households?
- **Focussed panels**
  - BES; LFS; Consumer panels; ELSA; Census LS; SLS
- **Improving data quality**
  - Reliability
  - Households: Household sharers' data
- **Longitudinal questions**
  - Time-varying behaviours
  - Getting the 'full picture'
  - Retrospective recall errors
  - Household life-courses
  - Household transitions
  - Intra-relations: similarity and dependence

The British Household Panel Survey 1991-2006

- **Panel study of 5k households re-contacted annually since 1991**
- **Major UK research investment**

For lots more introductions, see:
http://www.longitudinal.stir.ac.uk/

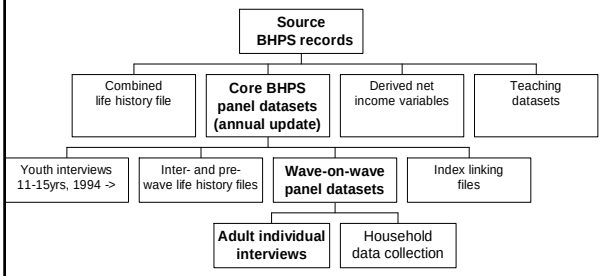
The 'Essex' BHPS

- **ISER** - Institute for Social and Economic Research
- **ULSC** - UK Longitudinal Studies Centre
- ***Design, coordinate, release, analyse and promote the BHPS***
- Data supplied by **UK Data Archive** at University of Essex
- Online documentation and support: *http://iserwww.essex.ac.uk/ulsc/bhps/doc/*

Annual survey since 1991

- Sample re-interviewed once a year
- Each new panel is a 'wave'
- Interviews start each September
- Datasets updated and re-released annually
- Government funding to at least 2009

BHPS data file structures



You'll most likely use..

- Adult individual interviews
  - All adults within household contribute and individual record
- Youth records
  - All 11-15's within household
- Combined life-history files
  - Oriented around event history analyses (durations)

Sampling design

- **W1 (1991): Stratified random sample of 5,500 households**

➢ 14,000 'OSM' household members

➢ Later waves: trace all OSM's; their descendants; and their household sharers (TSM's)

*NB: longitudinal trace of **individuals** and their surrounding household, but **not** of 'longitudinal households'*

Extension samples

- W7-11 -> ECHP supplement (low incomes)
- W9- -> Scottish and Welsh boosts
- W11- -> Northern Irish boosts

Future: possible extension / supplement samples
possible minority group boosts?

- **These are important!!**
  - affect representativeness
  - use of weights is complicated
  - catches every user out at least once…

	BHPS sampling structure							
	OSM	**TSM**	**ECHP boost**	**Scot. boost**	**Wales boost**	**N. Irel boost**	**Total sample**	**Tot adults interviewed**
Wave:								
A: 1991	13,840						13,840	10,264
B: 1992	12,567	584					13,151	9,845
C: 1993	12,219	885					13,104	9,600
D: 1994	11,821	1030					12,851	9,481
E: 1995	11,425	1124					12,549	9,249
F: 1996	11,412	1308					12,720	9,438
G: 1997	11,251	1301	2490				15,042	11,193
H: 1998	11,161	1300	2374				14,835	10,906
I: 1999	10,996	1337	2258	3397	3577		21,565	15,624
J: 2000	10,773	1481	2193	3584	3573		21,604	15,605
K: 2001	10,624	1610	2125	3518	3523	5188	26,588	18,869
L: 2002	10,470	1664		3329	3385	4589	23,437	16,599

BHPS Unbalanced panel & Data Management:

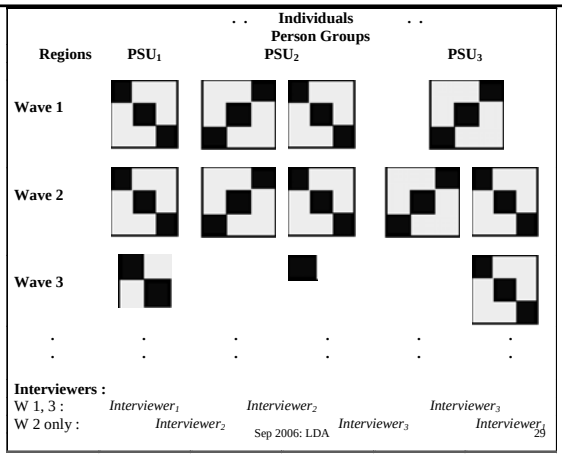
Below data may have come from 6 different BHPS source files

Wave	**Person**	**← Person-level Vars →**			
1	**1**	1	38	1	36
1	**2**	2	34	2	0
1	**3**	2	6	9	-
2	**1**	1	39	1	38
2	**2**	2	35	1	16
3	**1**	1	40	1	36
3	**2**	2	36	1	18
3	**3**	2	8	9	-
N_w=3	N_p=3				

The household structure of the BHPS

- All adults within a household are interviewed
  - Clustering analysis issues
  - Person groups?
- All persons within a household are ennumerated
- Children records
  - Rising 16's
  - Siblings and migration

➢ *BHPS Household analysis possibilities are exciting but complex..*



International household panels

- The BHPS is part of:
  - ECHP (1997-2001)
  - CHER (1991-2000)
  - PACO (1991-1998)
  - CNEF (1991->)
  - EU-SILC (2003 onwards: under-discussion)
  - Numerous stand-alone comparative projects
- Source project : US PSID
  - Use of comparable questionnaire design
  - Comparative methodolgocial literature [esp. Rose 2000]
- Market leaders(?)
  - GSOEP: http://www.diw.de/english/sop/
  - BHPS: http://iserwww.essex.ac.uk/ulsc/bhps/
  - PSID: http://psidonline.isr.umich.edu/
  - HILDA: http://melbourneinstitute.com/hilda/

BHPS in summary

Assets:

- Large scale panel
- Household info
- wYOUTH records
- Occupational data
- Income data
- Sub-populations

Drawbacks:

- Complexity
- Esp life history ?'s
- Short term coverage
- Dropout / non-resp.
- Panel conditioning(?)
- Regional
- Clustering

Summary on general purpose panels

- Gap between methodological textbooks and general purpose panel data [cf Rose 2000]
- Research evidence and value:
  - Major resources
  - ??

Data sources and data structure: Panel data

1. Structural features of panel data
2. General purpose panel studies
3. Cohort and follow-up studies
4. Issues in working with panel data

- **Major cohort studies**
  - Birth cohorts [UK: 1946; 1958; 1970; 2000]
  - Age cohorts [YCS; retirement surveys]
- **Focussed cohorts**
  - Offender's studies; Smoking cessation; Voting patterns; psychometric testing;
- **Follow-up studies**
  - Mobility studies; SCELI;

➢ *Characterised by fewer contacts spread over a longer period of time*

Features: Data management

- Ordinarily **wide format** structures
- **Longer time-span**
  - Variable construction difficulties over time
- Management of multi-cohort designs / comparisons
  - Difficult harmonisations between cohorts
- **High attrition**
- Typically complex sampling strategies

1990s Cohorts of the YCS

YCS Cohort	YEAR										
	91	**92**	**93**	**94**	**95**	**96**	**97**	**98**	**99**	**00**	**01**
5	**1**	**2**	**3**								
6		**1**	**2**	**3,4a**							
7				**1**		**2**					
8						**1**		**2**		**3a**	
9									**1**	**2**	**3,4a**

Sweeps of data collection usually take place in the spring (e.g. Easter). An autumn sweep is denoted "a". In Cohort 6 and Cohort 9 there were two sweeps of data collections for the same cohort both in spring and autumn of the same calendar year.

Features: Data analysis

- Truly temporal analyses – eg childhood maths tests impacting adult employment
  - Wide format: $y_{it3}=x_i + x_{it3} + x_{it1} + \ldots$
  - Long format: $y_{it}=x_{it} + u_i + v_t$ ('Growth curve')
- Attrition rates make missing data models desirable
- Administrative v's substantive definitions of the cohort => interpretation of cohort differences

Data sources and data structure: Panel data

1. Structural features of panel data
2. General purpose panel studies
3. Cohort and follow-up studies
4. Issues in working with panel data

See Stata & SPSS data management example command files:

LDA Web site www.longitudinal.stir.ac.uk

Key points:

- Importance of logging your work ('sytnax' / 'do' files)
- Array of related **'variable management'** functions
- **Merge separate files** via 'identifier variables'; check on merges via '_merge' indicator variables
- **Stata documentation** – extensive range of internet; manual; textbook guides (v9 'Data management' manual)

Stata bias?

Stata offers both a wide range of general purpose functions, and extended suites of purpose built longitudinal data management and data analysis functions

Some other benefits

- Succinct command syntax
- Diverse analytical and graphics capabilities
- Extension packages – eg GLLAMM
- User communities – contributed programmes; user support

Comparing SPSS & Stata for data management

Claim: For data management, Stata is ultimately much more powerful, but it is not always well designed [cf PEAS]

- Batch files / interactive syntax / programs:
  - *Stata has more flexibility, but SPSS interactive syntax is easier to use (Stata doesn't allow 'delimiter' character; other system requirements)*
- Direct data entry / browsing
  - *Stata is clumsy – easier to use SPSS or another package*
- Variable and value labels and presenting outputs
  - *SPSS quicker and better presentation; Stata needs more effort*
- Computing / recoding
  - *Stata more extensive (eg 'by' and 'if'); Stata prevents overwriting existing var*
- Missing values
  - *Stata's default settings cause more confusion than SPSS*
- Weighting data
  - *Stata has some restrictions on its weights / SPSS easier*
- Survey estimators (svy)
  - *Unique and advantageous feature of Stata*

Handling Panel datasets (Unbalanced)

Cases	**Year**	**← Variables →**			
1	**1**	1	17	1	1
1	**2**	1	18	2	1
1	**3**	1	19	2	-
2	**1**	1	17	1	3
2	**2**	1	18	1	1
3	**2**	2	20	2	2
3	**3**	2	21	2	2
3	**4**	2	22	1	1
n1=3	n2=8				

Handling panel (repeated-contacts) datasets

Panel datasets are essentially rectangular so any data management package is basically ok

- SPSS
  - no specific extension facilities for summarising, managing etc repeated contacts data
  - must be programmed in from first principles – possible but laborious
- Stata
  - ✓ possible to programme in extensive range of panel data operations
  - ✓ multitude of useful specifically designed extension data management facilities for summarising and managing repeated-contacts data ('xt' commands, eg, 'xtdes')
  - unhelpfully refers to repeated-contacts data as 'cross-sectional time series' (ie – XT)

Analysing panel datasets

- Many techniques work on any rectangular dataset – fine in Stata & other packages
  - *eg analysis of transitions; simple variance components models; computations of lag effects*
- SPSS has no specific panel data analysis extension functions
- ✓ Stata has an array of specific panel data analysis extension functions ('xt' commands, plus user-contributed extensions)

'Wide' versus 'Long' format panels

- 'Wide' = 1 case per record (person), additional vars for time points :

  Person 1 Sex YoB Var1_92 Var1_93 Var1_94 …
  Person 2 …
- 'Long' = 1 case per time point within person

  (as panel data example)

✓ Stata: 'reshape' command allows easy transfer between the two formats *(SPSS equivalent is more convoluted)*

The problems of variables

'Long' and 'Wide' panel datasets requires merging data from different years; we usually rename variables as equal across time

You know that:

- Questions and codings differ between interviews
- Time gaps differ
- Respondents drop out
- Interviewer modes differ

The computer sees:

- A nice neat file with lots of equivalent variables

Matching files

- Complex panel data inevitably involves **more than one related data file**
  - ***Linking them is a vital data analysis skill!!***
- Link data between files by connecting them according to **key linking variable(s)**
  - Eg, 'person identifier' variable 'pid'
  - Eg : iserwww.essex.ac.uk/ulsc/bhps/doc/

See SPSS and Stata example command files within 'lab 0' and 'lab 2', from www.longitudinal.stir.ac.uk

Types of file matching

- **Addition of files**
  - eg two files with same variables for different people
    - **Stata: append *using file2.dta***
    - **SPSS: add *files file="file1.sav" /file="file2.sav"*** .
- **Case-to-case matching**
  - *One-to-one link*, eg two files with different sets of variables for same people
    - **Stata: *merge pid using file2.dta***
    - **SPSS: *match files file="file1.sav" /file="file2.sav" /by=pid.***
- **Table distribution**
  - *One-to-many link*, eg one file has individuals, another has households -> match household info to individuals
    - **Stata: *merge pid using file2.dta***
    - **SPSS: *match files file="file1.sav" /table="file2.sav" /by=pid* .**

Types of file matching ctd

- **Aggregating**
  - Summarise over multiple cases
    - **Stata: -** ***collapse (mean) inc , by(pid)***
      or - egen avinc=mean(inc), by(pid)
    - **SPSS:** ***aggregate outfile="file2.sav" /break=pid /avinc=mean(inc)***
  - Output files from aggregate / collapse often linked back into the micro-data from which they are derived
- **Related cases matching**
  - Link info from one related case to another case, eg info on spouse put on own case
    - **Stata: -** ***merge pid using file2.dta***
      or - ***joinby …***
    - **SPSS:** ***match files file="file1.sav" /file="file2.sav" /by=pid.***

File matching crib:

Stata:

_merge = indicator of cases present for:

1 = Master file but not input file

2 = Input file but not Master file

3 = Master and input file

- Remember to drop auto-generated _merge before performing next merge command..

- See the Stata documentation (v8: User Guide, sections 15,25; v9: manual)