Quantitative Longitudinal Data

Paul Lambert and Vernon Gayle
Stirling University

Prepared for *"Longitudinal Data Analysis for Social Science Researchers: Introductory Seminar"*, *Royal Statistical Society*, 28th April 2006

Five Approaches to Longitudinal Data Analysis

http://www.longitudinal.stir.ac.uk/

Introducing quantitative longitudinal data analysis	**1. Repeated cross-sections**
2. Panel datasets	**3. Cohort studies**
4. Event history datasets	**5. Time series analyses**

Quantitative longitudinal research in the social sciences

- **Survey resources**
  - Micro-data (individuals, households, ..)
  - Macro-data (aggregate summary for year, country..)

Data analysis is used to give a parsimonious summary of patterns of relations between variables in the survey dataset

- **Longitudinal**
  - Data concerned with more than one time point
  - Repeated measures over time

Motivations for QnLR

- **Focus on time / durations**
  - Trends in repeated information over time
  - Substantive role of durations (e.g., Unemployment)
- Focus on **change / stability**
- Focus on the **life course**
  - Distinguish age, period and cohort effects
  - Career trajectories / life course sequences
- **Getting the 'full picture'**
  - **Causality** and **residual heterogeneity**
  - *Examining multivariate relationships*
  - *Representative conclusions*

- **Specific features to QnLR**
  - Tends to use **'large and complex' secondary data**
    - Multiple points of measurement
    - Complex (hierarchical) survey structure / relations
    - Complex variable measures / survey samples
    - *Secondary data analysis positives: other users; cheap access; range of topics available*
  - **Particular techniques of data analysis**
    - *Algebra*
    - *Computer software manuals*
    - *Spectacles*

Some drawbacks

- **Dataset expense**
  - *mostly secondary; limited access to some data (cf. disclosure risk)*
- **Data analysis**
  - *software issues (complexity of some methods)*
- **Data management**
  - *complex file & variable management requires training and skills of good practice*

Five Approaches to Longitudinal Data Analysis

Introducing quantitative longitudinal research	**1. Repeated cross-sections**
2. Panel datasets	**3. Cohort studies**
4. Event history datasets	**5. Time series analyses**

Repeated Cross-sections

- **By far the most widely used** longitudinal analysis in contemporary social sciences

Whole surveys, with same variables, repeated at different time points

and

Same information extracted from different surveys from different time points

Illustration: Repeated x-sect data

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				

Some leading repeated cross-section surveys : UK

OPCS Census	British Crime Survey
Labour Force Survey	British Social Attitudes
New Earnings Survey	British Election Studies
Family Expenditure S.	Policy Studies (Ethnicity)
General Household Survey	Social Mobility enquires

Some leading repeated cross-section surveys : International

European Social Survey	PISA / TIMMS (schoolkid's aptitudes)
IPUMS census harmonisation	ISSP
LIS/LES (income and employment)	Eurobarometer

Repeated cross sections

- ✓ **Easy to communicate** & appealing: how things have changed between certain time points
- ✓ Partially distinguishes **age / period / cohort**
- ✓ Easier to analyse – **less data management**

However..

- ☹ Don't get other QnLR attractions (nature of changers; residual heterogeneity; causality; durations)
- ☹ Hidden complications: are sampling methods, variable operationalisations *really* comparable? (don't overdo: concepts are more often robust than not)

Repeated X-sectional analysis

1. **Present stats distinctively by time pts**
   - Analytically sound
   - Tends to be descriptive, limited # vars
2. **Time points as an explanatory variable**
   - More complex, requires more assumptions of data comparability
   - Can allow a more detailed analysis / models

Example 1.1: UK Census

- Directly access aggregate statistics from census reports, books or web, eg:

Wales: Proportion able to speak Welsh				
Year	1891	1981	1991	2001
%	54	19	19	21

- *Census not that widely used: larger scale surveys often more data and more reliable*

Eg1.2: UK Labour Force Survey

LFS: free download from **UK data archive**

http://www.data-archive.ac.uk/

Same questions asked yearly / quarterly

Example 1.2i: LFS yearly stats

Percent of UK workers with a higher degree, by employment category and gender (m / *f*)

Sample size ~35,000 m / 30,000 f each year

	1991	1996	2001
Profess.	**14.4**	**19.9**	**24.9**
Non-Prof.	**1.3**	**2.5**	**3.5**
Profess.	***11.0***	***24.4***	***28.3***
Non-Prof	***0.6***	***2.3***	***3.2***

Example 1.2ii: LFS and time

Log regression: odds of being a professional from LFS adult workers in 1991, 1996 and 2001

		B	Sig.	Exp(B)
a	Higher degree	2.383	.000	10.842
	Female	-.955	.000	.385
	Age in years (/10)	.777	.000	2.174
	Age in years squared (/1000)	-.857	.000	.424
	Time point 1991	.094	.000	1.098
	Time point 2001	-.195	.000	.823
	(Time in years)* (Higher Degree)	**-.030**	**.000**	**.971**
	Constant	-4.232	.000	.015

a. Nagelkere R2=0.11

Five Approaches to Longitudinal Data Analysis

Introducing quantitative longitudinal research	**1. Repeated cross-sections**
2. Panel datasets	**3. Cohort studies**
4. Event history datasets	**5. Time series analyses**

Panel Datasets

Information collected on the same cases at more than one point in time

- 'classic' longitudinal design
- incorporates 'follow-up', 'repeated measures', and 'cohort'

Panel data in the social sciences

- **Large scale studies**
  - ambitious and expensive; normally collected by major organisations; efforts made to promote use
- **Small scale panels**
  - are surprisingly common…
- **‘Balanced’ and ‘Unbalanced’** designs

Illustration: Unbalanced panel

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	******also 'sweep', 'contact',..*			

Panel data advantages

- **Study 'changers'** – how many of them, what are they like, what *caused* change
- Control for **individuals' unknown characteristics** ('residual heterogeneity')
- Develop a full and **reliable life history**
  - *eg family formation, employment patterns*
- Contrast **age / period / cohort effects**
  - *but only if panel covers long enough period*

Panel data drawbacks

- **Data analysis**
  - can be complex; methods advanced / developing
- **Data management**
  - tends to complexity, need training to get on top of
- **Dataset access**
  - Primary / Secondary data
- **Attrition**
- **Long Duration**
  - eg politics of funding; time until meaningful results

Some leading panel surveys : UK

British Household Panel Study (BHPS)
ONS Longitudinal Study (Census 1971->)
British Election Panel Studies
Labour Force Survey rotating panel
School attainment studies (various)
Health and medical progress studies (various)

Some leading panel studies : International

European Community Household Panel Study (1994-2001)
EU-SILC (2003 ->)
CHER, PACO, CNEF (individual projects harmonising panels)
Panel Study of Income Dynamics (US)

Analytical approaches

i) Study of Transitions / changers

- simple methods in any package, eg cross-tab if changed or not by background influence
- but complex data management

ii) Study of durations / life histories

- See section 5 ‘event histories’

Example 2.1: Panel transitions

Young people's household circumstance changes by subjective well-being between 1994 and 1995.

BHPS youth panel, 11-14yrs in 1994, row percents.

	Stays happy	Cheers up	Becomes miserable	Stays miserable	N
HH Stable	**54%**	**19%**	**10%**	**18%**	**499**
HH Changes	**42%**	**22%**	**14%**	**22%**	**81**

Analytical approaches

iii) Panel data models:

$$Y_{it} = BX_{it} + \ldots + \epsilon$$

Cases *i*	Year *t*	← 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

Panel data model types

- **Fixed and random effects**
  - Ways of estimating panel regressions
- **Growth curves**
  - Multilevel speak : time effect in panel regression
- **Dynamic Lag-effects models**
  - Theoretically appealing, methodologically not..

Analytically complex and often need advanced or specialist software

- ***Econometrics literature***
- ***STATA / GLLAMM; R; S-PLUS; SABRE / GLIM; LIMDEP; MLWIN; MPLUS; …***

Example 2.2: Panel model

BHPS 1994-8: Output from Variance Components Panel model for determinants of GHQ scale score (higher = more miserable), by individual factors for multiple time points per person[a]

Parameter	Estimate	Std. Error	Sig.	95% Confidence Interval	
				Lower Bound	Upper Bound
Intercept	12.69	.168	.000	12.4	13.0
Female	-1.36	.076	.000	-1.5	-1.2
In work	-1.23	.082	.000	-1.4	-1.1
Unemployed	.50	.131	.000	.2	.8
FT studying	-1.70	.141	.000	-2.0	-1.4
Age in years	.00	.002	.055	.0	.0
Holds degree or diploma	-.07	.076	.356	-.2	.1
Time point	.03	.014	.020	.0	.1

a. Variance components : Person level= 46%, individual level = 54%

Five Approaches to Longitudinal Data Analysis

Introducing quantitative longitudinal research	**1. Repeated cross-sections**
2. Panel datasets	**3. Cohort studies**
4. Event history datasets	**5. Time series analyses**

Cohort Datasets

Information on a group of cases which share a common circumstance, collected repeatedly as they progress through a life course

- Simple extension of panel dataset
- Intuitive type of repeated contact data
  - E.g. ‘7-up’ series

Cohort data in the social sciences

- Circumstances parallel other panel types:
  - Large scale studies ambitious & expensive
  - Small scale cohorts still quite common…

❖**Attrition problems** often more severe

❖Considerable **study duration problems –** have to wait for generations to age

Cohort data advantages

- **Study of 'changers'**
  - a main focus, looking at how groups of cases develop after a certain point in time
- **Full and reliable life history**
  - as often covers a very long span
- **Variety of issues**
  - Topics of relevance can evolve as cohort progresses through lifecourse
- **Age / period / cohort** effects
  - *Better chance of distinguishing (if >1 cohort studied)*

Cohort data drawbacks

- **{Data analysis / management demands}**
- **Attrition** problems more severe than panel
- **Longer Duration**
- **Very specific findings –** eg only for isolated people of a specific cohort

Some leading UK cohort surveys

Birth Cohort Studies
•1946 National Survey of Health and Development •1958 National Child Development Study •1970 Birth cohort study •2000 Millenium Cohort Study
Youth Cohort Studies (1985 onwards)
Health and medical progress studies (various)
Criminology studies of recidivism (various)

Cohort data analytical approaches

..parallel those of other panel data:

i. Study of **transitions** / changers
ii. Study of **durations** / life histories
iii. Panel data **models**

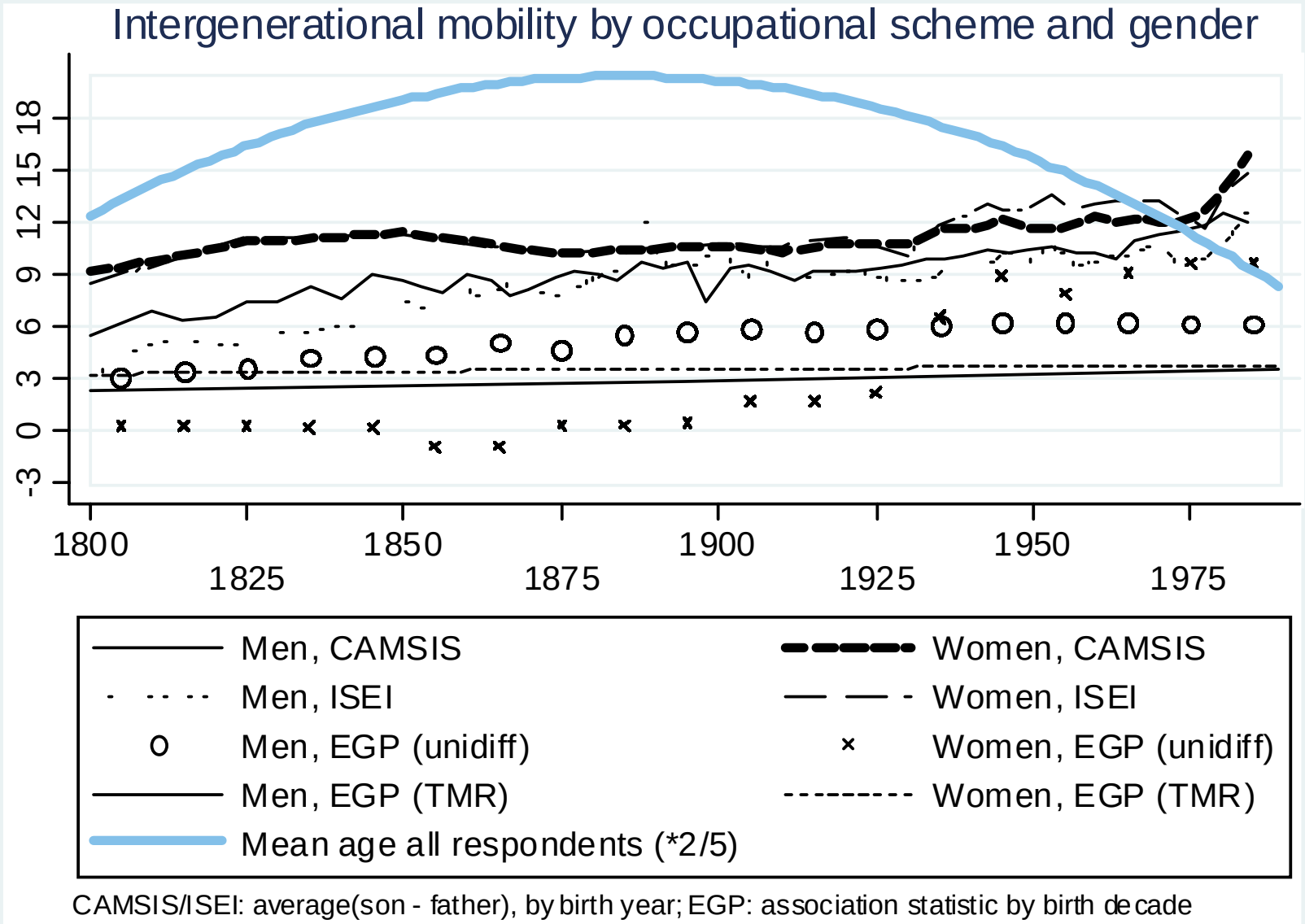
May focus more on life-course development than shorter term transitions

Cohort data analysis example

- Blanden, J. et al (2004) "Changes in Intergenerational Mobility in Britain", in Corak, M. (ed) *Generational Income Mobility in North America and Europe*. Cambridge University Press.
- Intergenerational mobility is declining in Britain:

Adj. Coefficient for father's income when aged 16		
	m	f
NCDS, age 33 in 1991	0.132	0.113
BCS, age 30 in 2000	0.253	0.239

..but with repeated cross-sections..



Five Approaches to Longitudinal Data Analysis

Introducing quantitative longitudinal research	**1. Repeated cross-sections**
2. Panel datasets	**3. Cohort studies**
4. Event history datasets	**5. Time series analyses**

Event history data analysis

Focus shifts to length of time in a 'state' - analyses determinants of time in state

- Alternative data sources:
  - Panel / cohort (more reliable)
  - Retrospective (cheaper, but recall errors)
- Aka: 'Survival data analysis'; 'Failure time analysis'; 'hazards'; 'risks'; ..

Social Science event histories:

- Time to labour market transitions
- Time to family formation
- Time to recidivism

Comment: Data analysis techniques relatively limited, and not suited to complex variates

⇨ *Many event history applications have used quite simplistic variable operationalisations*

Event histories differ:

- In **form of dataset** (cases are spells in time, not individuals)

➢ Some complex data management issues

- In types of **analytical method**

➢ Many techniques are new or rare, and specialist software may be needed

Key to event histories is ‘state space’

Episodes within state space : Lifetime work histories for 3 adults born 1935

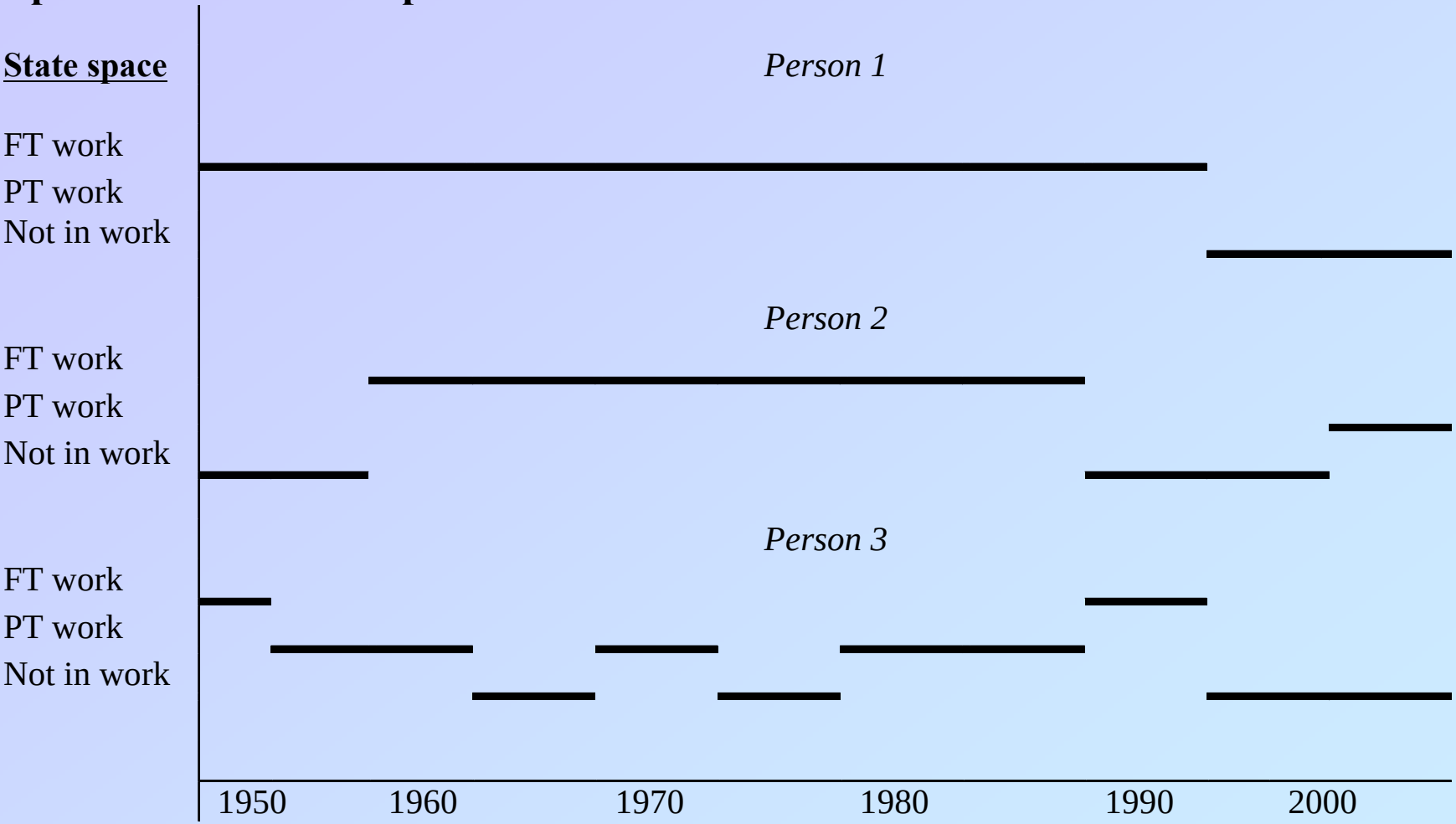


Illustration of a continuous time retrospective dataset

Case	Person	Start time	End time	Duration	Origin State	Destination state	{Other vars, person/state}
1	1	1	158	157	1 (FT)	3 (NW)	
2	1	158	170	12	3 (NW)	3(NW)	
3	2	1	22	21	3 (NW)	1 (FT)	
4	2	22	106	84	1 (FT)	3 (NW)	
5	2	106	149	43	3 (NW)	2 (PT)	
6	2	149	170	21	2 (PT)	2 (PT)	
7	3	1	10	9	1 (FT)	2 (PT)	
.	.	.	.	.	.		.

Illustration of a discrete time retrospective dataset

Case	Person	Discrete Time	Approx real time	State	End of state	{Other person, state, or time unit level variables}
1	1	1	5	1 FT	0	
2	1	2	20	1 FT	0	
3	1	3	35	1 FT	0	
4	1	4	50	1 FT	0	
5	1	5	65	1 FT	0	
6	1	6	80	1 FT	0	
7	1	7	95	1 FT	0	
8	1	8	110	1 FT	0	
9	1	9	125	1 FT	0	
10	1	10	140	1 FT	1	
11	1	11	155	3 NW	0	
12	1	12	170	3 NW	1	
13	2	1	5	3 NW	0	
14	2	2	20	3 NW	1	
15	2	3	35	1 FT	0	
16	2	4	50	1 FT	1	
.	.	.	.	.	.	

Event history data permutations

- **Single state single episode**
  - Eg Duration in first post-school job till end
- **Single episode competing risks**
  - Eg Duration in job until promotion / retire / unemp.
- **Multi-state multi-episode**
  - Eg adult working life histories
- **Time varying covariates**
  - Eg changes in family circumstances as influence on employment durations

Some UK event history datasets

British Household Panel Study (see separate ‘combined life history’ files)
National Birth Cohort Studies
Family and Working Lives Survey
Social Change and Economic Life Initiative
Youth Cohort Studies

Event history analysis software

SPSS – limited analysis options

STATA – wide range of pre-prepared methods

SAS – as STATA

S-Plus/R – vast capacity but non-introductory

GLIM / SABRE – some unique options

TDA – simple but powerful freeware

MLwiN; lEM; {others} – small packages targeted at specific analysis situations

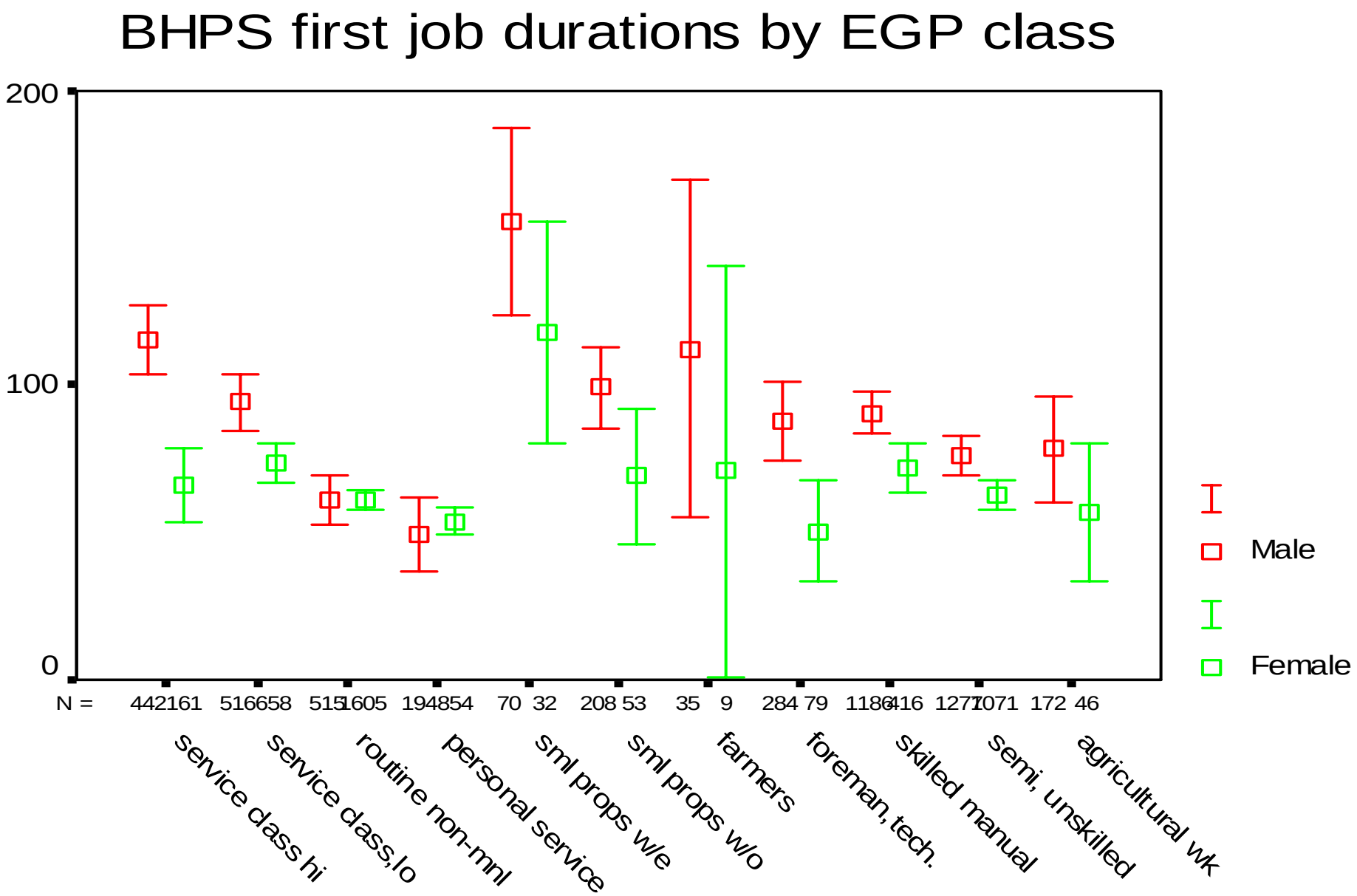
Types of Event History Analysis

i. **Descriptive:** compare times to event by different groups (eg survival plots)

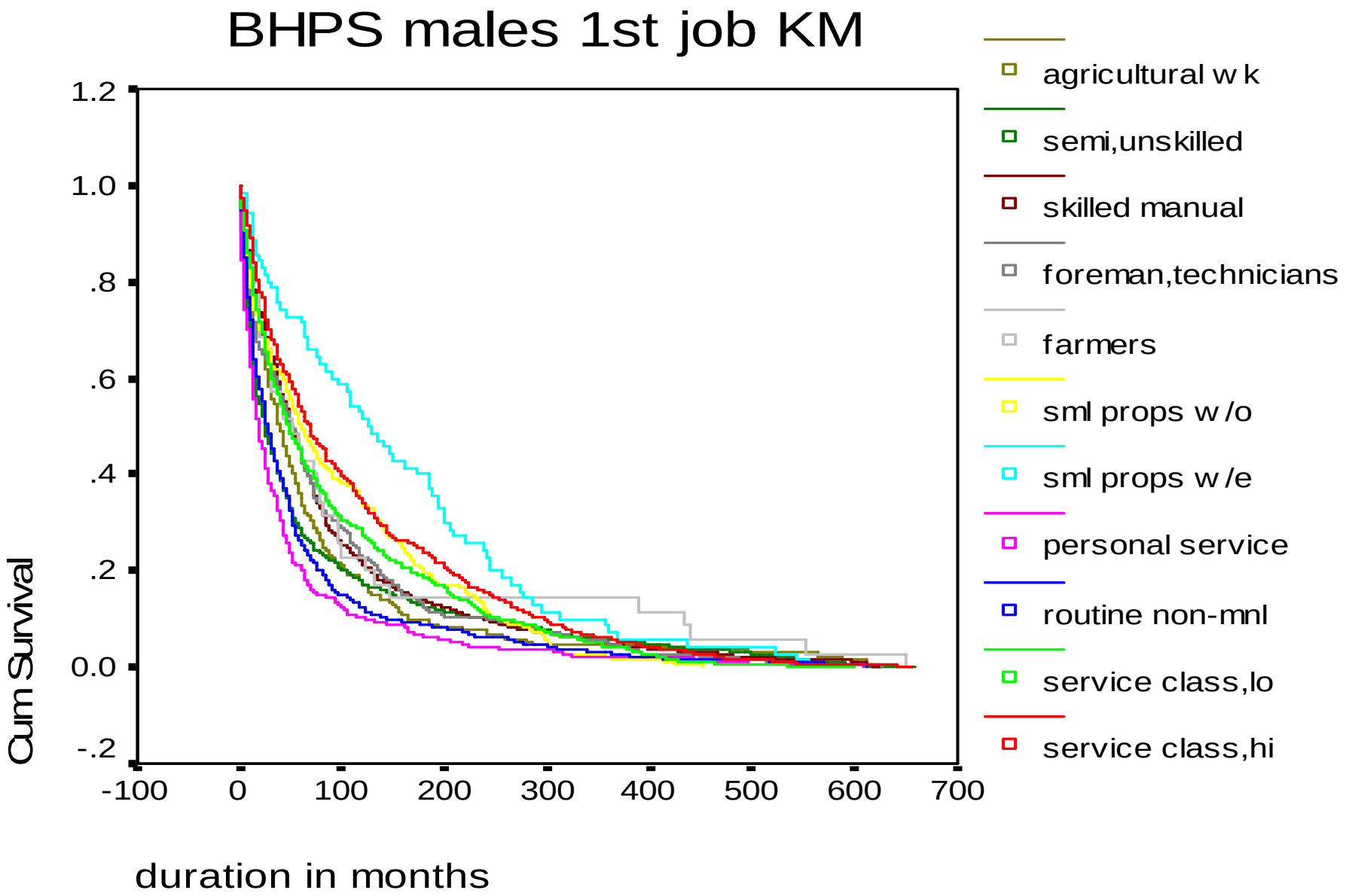
ii. **Modelling:** variations of Cox's Regression models, which allow for particular conditions of event history data structures

- Type of data permutations influences analysis – only simple data is easily used!

Eg 4.1 : Mean durations by states



Eg 4.1 : Kaplan-Meir survival



Eg 4.2: Cox's regression

Cox regression estimates: risks of quicker exit from first employment state of BHPS adults

	B	SE	Sig.
Female	.194	.081	.017
Self-employed	-.617	.179	.001
Age in 1990	-.062	.003	.000
Age in 1990 squared	.000	.000	.000
Hope-Goldthorpe scale	-.013	.001	.000
Female*self-employed	.214	.109	.049
Female* HG scale	-.003	.002	.061
Self-employed*HG scale	.000	.004	.897
Female*Age in 1990	.006	.001	.000

Five Approaches to Longitudinal Data Analysis

Introducing quantitative longitudinal research	**1. Repeated cross-sections**
2. Panel datasets	**3. Cohort studies**
4. Event history datasets	**5. Time series analyses**

Time series data

Statistical summary of one particular concept, collected at repeated time points from one or more subjects

Examples:

- Unemployment rates by year in UK
- University entrance rates by year by country

Comment:

- Panel = many variables few time points
  = *'cross-sectional time series' to economists*
- Time series = few variables, many time points

Time Series Analysis

i) Descriptive analyses

- charts / text commentaries on values by time periods and different groups
- **Widely used** in social science research
- But exactly **equivalent to repeated cross-sectional** descriptives.

Time Series Analysis

ii) Time Series statistical models

- **Advanced methods** of modelling data analysis are possible, require specialist stats packages
  - Autoregressive functions: $\mathbf{Y_t = Y_{t-1} + X_t + e}$
- Major strategy in business / economics, but **limited use in other social sciences**

Some UK Time Series sources

Time series databases (aggregate statistics)

- ONS Time series data
- ESDS International macrodata

Repeated cross-sectional surveys

- Census
- Labour Force Survey
- Many others..

Introducing quantitative longitudinal research	**1. Repeated cross-sections**
2. Panel datasets	**3. Cohort studies**
4. Event history datasets	**5. Time series analyses**

….Phew!

Summary: Quantitative approaches to longitudinal research

1) Pro's and cons to QnL research::

i. **Appealing analytical possibilities:** *eg analysis of change, controls for residual heterogeneity*
ii. **Pragmatic constraints:** *data access, management, & analytical methods; often applications over-simplify variables*
iii. **Uneven penetration** *of research applications between research fields at present*

Summary: Quantitative approaches to longitudinal research

2) Undertaking QnL research::

i. **Needs a bit of effort:** *learn software, data management practice – workshops and training facilities available; exploit UK networks*

ii. **Remain substantively driven:** *'methodolatry' widespread in QnL: applications 'forced' into desired techniques; often simpler techniques make for the more popular & influential reports*

iii. **Learn by doing** *(..try the syntax examples..)*

Some research resources

See website for text and links to further internet resources:

- **Many training courses in UK** – e.g. see ESRC Research Methods Programme

➢ **Practical exemplar data analysis and data management in SPSS and STATA:**

http://www.longitudinal.stir.ac.uk/