Event history data structures and data management

Paul Lambert
Stirling University

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

Event history data structures and data management

1. Structures of event history data
2. Measuring and analysing life-courses

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'; ..
- Specific analytical techniques required (SPSS has some; Stata has more)

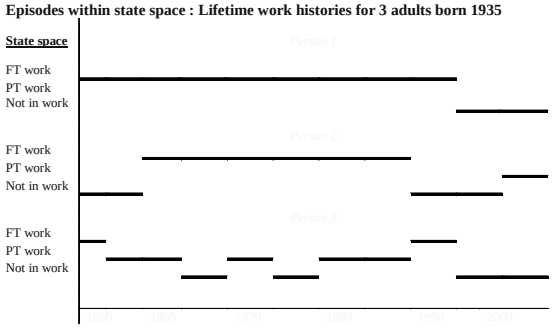
Event history dataset

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				

4

Key to event histories is 'state space'

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



Sep 2005:LDA - ONS 5

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

Sep 2005:LDA - ONS 6

Using Stata with event history data

- See the Stata manual *'Survival Analysis and Epidemiological Tables'*
- stset: declares survival time data:

  stset dur, failure(status)

  (each case is an episode; variable *dur* is the length of the episode; variable *status* indicates whether record was right censored - value 0 means it was censored, ie, the end of event didn't occur within observation period)
- Many specialist event history data analysis functions built into Stata
- Common pitfall: panel ~= duration data

'Wide' versus 'Long' in event history data

Relevant to multi-state or multi-episode data

- 'Wide' = state space typologies / sequences
- 'Long' = multiple states stacked above each other
  - Model controls for residual heterogeneity

Continuous v's Discrete time

- Continuous time ('spell files', 'event oriented')
  - One episode per case, time in case is a variable
- Discrete time
  - One episode per time unit, type of event and event occurrence as variables
  - More flexible: time-varying covariates
- Stata
  - Oriented to continuous time data
  - Discrete time data formats are usually analysed by first transforming to continuous time

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	
.	.	.	.	.	.	

Transforming between continuous and discrete time

- SPSS
  - Discrete to continuous: simple aggregation
  - Continuous to discrete: bespoke programme
- Stata
  - Discrete to continuous: simplete aggregation (collapse)
  - Continuous to discrete: stsplit

Continuous v's Discrete time

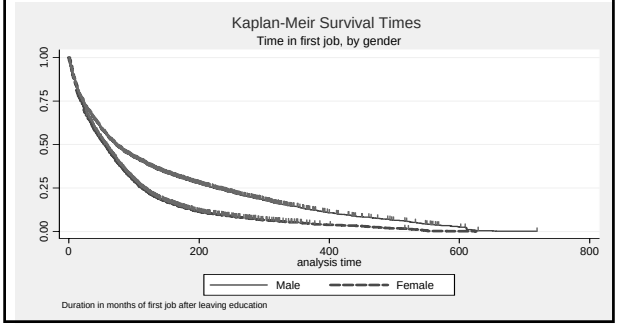
- Continuous
  - Gives full data on length of time period
  - More manageable data
- Discrete time
  - More flexible: time-varying covariates
  - More easily estimated: basic logistic regression with residual heterogeneity
  - Usually some time period detail is lost
    - [Germany: annual panels as discrete time EH data]

Multiple comparisons

– Event history: multivariate=chopping up data

– SPSS and Stata usually do this automatically

- (may use 'select if' / 'by')



Continuous time and time varying covariates?

Continuous time				'Episode-splitting'				summary	
id	start	dur	cohab	Id	sp	dur	coh	id	coh
1	26	12	0	1	1	26	0	1	0
2	40	3	1	2	1	40	1	2	1
3	35	20	*?From mth 6*	3	1	5	0	3	0.25
				3	2	15	1		

•Discrete time captures Time-Varying variables easily.

•Continuous time needs some work:

- •Summary totals obtained through aggregation
- •Episode splitting / discretising - automated in Stata, not SPSS

Ties

- In continuous time data, no two events should end at precisely the same time
- Estimation of survival functions assumes no tied ending points
- In practice, measurement units mean ties are common
- Analytial solution
  - Breslow 1974 (Stata default)
  - Alternative calculations

Event history data structures and data management

1. Structures of event history data
2. Measuring and analysing life-courses

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)	
.	.	.	.	.	.		.

Obtaining event history data

We need

- **State** (*simple* characterisation of circumstance)
  - Start state (single v's multi-state)
  - End state (single v's competing risks)
- **Duration** (start and end time)
- **Censoring** indicator (right-censoring)
  - Did the event end during observation, or not?
- Other covariates
  - Must be able to map them to the event record

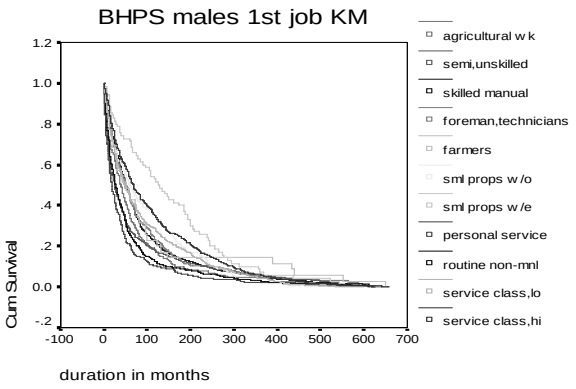
Obtaining event history data

- Data collection
  - 'Diary' records / sequences of events
  - See questionnaire schedules
- Data construction : calculating duration
  - Dur = (end – start + 1)
  - Time[end | start] = data in an absolute unit
  - E.g. {calendar months; years; days} since 1900
  - Stata and SPSS date conversion functions

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

Descriptive : Kaplan-Meir survival



Modelling: Cox's regression

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

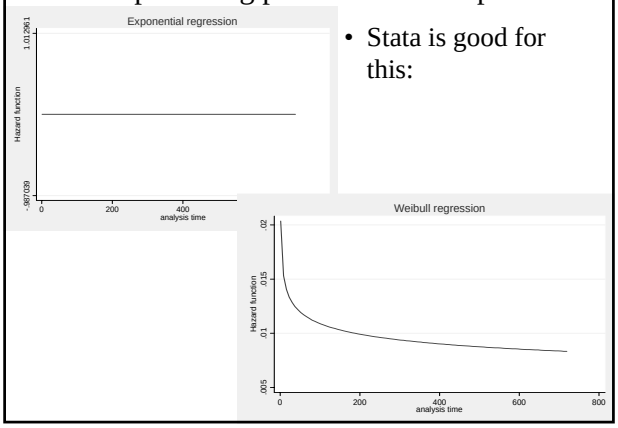
	B	SE	Sig.
Female	.194	.081	.017
Self-employed	-.617	.179	.001
Age in 1990	-.062	.003	.000
Age in 1990 squared	.000	.000	.000
Hope-Goldthorpe scale	-.013	.001	.000
Female*self-employed	.214	.109	.049
Female* HG scale	-.003	.002	.061
Self-employed*HG scale	.000	.004	.897
Female*Age in 1990	.006	.001	.000

Models – Loglinear models

- [e.g. Gilbert 1993; Vermunt 1997]
- Event histories are categorical measures
- Discrete time approach allows log-linear model

	Men		*Women*	
	End of spell (last record)			
Year	0	1	0	1
1	90	10	40	10
2	60	30	20	20
3	40	20	5	15
4	25	15	1	4

Conceptualising parametric assumptions



- Stata is good for this:

Sequence analyses

- Descriptive techniques to characterise state sequences
  - 'career centred mode of analysis' (Taris 2000)
- Categorical sequences
  - (c.f. growth curves / regression trajectories)
- Various possible implementations:
  - Latent class growth curves
  - Optimal matching analysis
  - Discriminant analysis
  - Sequence analysis
- *Relies upon a theory of the nature of the career and state spaces*

Overview: Event history data structures

- Focus upon speed until event
- Importance of censoring
- Categorisation of social science information
- Causality
  - Causal effects upon speed to transitions
    - 'event history models provide a time-relatd empirical representation of the structure of causal arguments' [Blossfeld and Rohwer 2003, p24]
    - But: restrictive state spaces and analytical options make for a limited description..