A note on methodology for analysing longitudinal event histories using repeated partnership data from the national Child Development Study (NCDS).

Harvey Goldstein[1], Huiqi Pan[2] and J. Bynner [3]

Abstract

We show how repeated durations of particular events within individuals can be modelled using a flexible discrete time event history model that incorporates individual level random effects. The model is applied to the analysis of partnership episodes for adult members of the National Child Development Study followed up between the ages of 16 and 33.

Keywords

Event history model, longitudinal data, multilevel model, repeated measures.

Acknowledgements

We are very grateful to Fiona Steele for comments.

1, 3. Bedford Group, Institute of Education, University of London

2. Institute of Child health, University of London

email: h.goldstein@ioe.ac.uk

Introduction

A major aim of longitudinal surveys is to identify the factors shaping the form and timing of transitions between states and the time spent in different states. In domains such as employment, family formation and health there are many kinds of states with varying transitions out of them, such as getting a job, forming a partnership or having a child. These can be viewed for each individual as a set of episodes comprising an 'event history' and within which the durations of the episodes vary between and within individuals. This creates a 2-level structure with episodes nested within individuals and thus requires a multilevel modelling approach to analysis. Several options are available (see Goldstein, 1995) including a fully parametric approach modelling duration lengths, semi-parametric procedures such as the Cox model and piecewise discrete time models. While all these models can, in principle be extended to multilevel data structures (Goldstein, 1995, Chapter 9) the latter models generalise very straightforwardly to a multilevel structure and can be fitted by any software that can handle multilevel binary response data. This paper is intended to provide an exposition of the use of these models with longitudinal survey data and we explore this with data from the National Child Development Study on partnership durations (Bynner et al., 2002). We are interested in the duration of two states – not being in a partnership and belonging to a partnership. We have chosen as explanatory variables, age at the start of each episode, and social class based upon own occupation since both are known to be associated with the age at which partnerships begin and the number that occur (Bynner et al., 2002).

Methodology

We assume that, given individual characteristics and any time-varying effects, and including any random effects, the durations for each state are independently distributed. The total time interval is divided into short time intervals, in our case 3 months, within which at most one transition is assumed to have taken place. These actual time intervals are re-coded into *modelled time intervals* (z) grouped within an episode, determined by the event state. For example data for individual 1 may look as follows:

Individual	Actual time interval	Modelled time interval	Response	Event state
1	1	1	0	no partnership
1	2	2	0	no partnership
1	3	3	1	no partnership
1	4	1	0	partnership
1	5	2	0	partnership
1	6	3	0	partnership
1	7	4	1	partnership
1	8	1	0	no partnership
1	9	2	0	no partnership
1	10	3	0	no partnership

Thus, starting in state 'no partnership', individual 1 moves in time interval 3 to state 'partnership' and in time interval 7 to state 'no partnership'. The response yariable, y, takes

the value zero if no move takes place and one if a change in partnership status occurs during the interval. Thus the hazard at modelled time *t* is

$$\pi_{ijk(t)} = P(y_{ijk(t)} = 1 \mid y_{ijk(t-1)} = 0)$$

where *k* indexes individual, *j* indexes episode and *i* indexes the state. The states (partnership, non-partnership) are modelled by dummy variables and this leads to a 3-level binary response model where, as we have assumed, the (conditional) responses within episodes within individuals are independent. Level 3 is the individual, level 2 represents variation between repeated episodes within individuals and level 1 refers to the time interval within repeated episodes. Using the conventional logit link function this model can be written in the form

$$\begin{aligned} &\text{logit}(\pi_{ijk(t)}) = \beta_0 + \sum_{h=1}^{p} \alpha_h^* (z_{it}^*)^h + \sum_{l=1}^{m} \beta_l x_{lijk(t)} + u_{jk}^{(i)} + v_k^{(i)} \\ &y_{ijk(t)} \sim Bin(1, \pi_{ijk(t)}) \end{aligned} \qquad (1)$$

where z_{it}^* is an index for the modelled interval at time *t* and x_{kij} the covariates, including the dummy variable for partnership state. The term $v_k^{(i)}$ is the random effect for individual *k* for state *i* and $u_{jk}^{(i)}$ is the random effect associated with the *j*-th episode for the *k*-th individual. In fact, for the present data we can detect no variation at level 2 so that, for simplicity we shall assume just a 2-level model in the following exposition. We also have no time varying covariates. Thus the model becomes

$$\text{log}it(p_{ijk(t)}) = b_0 + \sum_{h=1}^{p} a_h^* (z_{it}^*)^h + \sum_{l=1}^{m} b_l x_{lijk} + v_k^{(i)} \qquad (2)$$

We retain the subscript j in the covariate expression to allow for episode varying covariates. Following Goldstein (1995, Chapter 9) we make use of a high order polynomial rather than a full set of blocking factors to define the underlying hazard at each time interval and the order of this polynomial, p, is typically 4 or 5. Where there are many discrete time intervals this avoids the estimation of a large number of nuisance terms; one for each ordered time interval. An alternative is to group the blocking factors into a small number of relatively homogeneous longer intervals but we have not pursued this. At the individual level we may have several random effects, in particular for each state, and these will covary.

The population probability of survival to the end of modelled time interval t for state i is

$$\prod_{h=1}^{t}(1-\pi_{ijk(h)})$$

and these can be averaged over individual random effects to provide population estimates. MLwiN version 1.10 is used for the analysis (Rasbash et al, 2000).

Data

The National Child Development Study (NCDS) is a longitudinal study which takes as its subjects all those living in Great Britain who were born between 3 and 9 March, 1958. The fifth follow-up of the National Child Development Study (NCDS5) took place in 1991 when the cohort members were age 33. 'Your Life Since 1974' was a self-completion questionnaire posted to the cohort members during the course of NCDS5 which asked for retrospective information on relationships, children, jobs and housing from the age of 16 until the time of the 1991 survey. Altogether, 11178 persons filled in

either all or some of this section of the survey. In addition the 'Cohort Member Interview' of NCDS5, carried out by trained interviewers, also contained a retrospective partnership history. The final cleaned partnership histories (Bynner et al., 2002) are used to derive the three-monthly duration data for the study.

All but 39 of the cohort members had no more than four partnerships by the age of 33. Partnership involves cohabitation or marriage. Very few cohort members (61) had gone back to partners with whom they had lived before. The present analysis uses only a subset of the variables available. These are the 'start age', the age of the cohort member at the start of the current episode; their social class (manual or non-manual – 2% had missing data or other codes), and whether the episode is a partnership or non-partnership.

The data are for cohort males only and we carry out two analyses. The first ignores the initial time to establish a partnership, so that the first state is always a partnership. The second analysis includes the time to first partnership. The analysis has essentially two response variables for each individual; the time to first partnership from age 16, that is the first episode, and the duration for subsequent episodes. At level 2 we therefore have three random effects, one for the first episode, one for partnership and one for non-partnership durations. We can alternatively, if we wish, regard this as a three variate response model. In this analysis the start age is omitted as an explanatory variable since, for the first partnership, it is effectively confounded with the time to first partnership response. The total number of 3-month periods is 140420 with 3737 male cohort members who had partnership histories.

Results

A fifth-order polynomial was used to smooth the blocking factors. Main effects for the above factors are fitted, together with interactions between the partnership status dummy variable and all the others, including the polynomial terms. Since the polynomial coefficients are essentially nuisance parameters used only to define the time dependent hazard, we do not interpret them, although they are used in the construction of the survivor functions. We have retained several 'non signficant' polynomial coefficients in order to ensure that the underlying hazard is sufficiently accurately estimated; the order was chosen by noting when the predicted probabilities changed only negligibly when a further polynomial term was added.

Table 1 shows the estimates for a variance components model fitted using both quasilikelihood estimation (PQL1, see Goldstein, 1995 Chapter 7) and Markov Chain Monte Carlo (MCMC, see Browne and Draper, 2000). The MCMC model uses a gamma prior (see Rasbash et al, 2000) and was run for 10,000 iterations with a burn in of 1000. It gives somewhat different estimates from PQL1, which can happen with binary response data, and in subsequent tables we quote only the MCMC estimates. There is some evidence that those with manual occupations have longer durations when in a partnership (the negative coefficient is associated with a lower probability of terminating an episode at any given time), but a small and non significant difference (-0.113+0.142=-0.029) for non-partnerships. The later the starting age the longer the duration for partnerships but there is only a small (-0.041+0.026=-0.015) and non significant relationship for non-partnership durations. Table 2 introduces a random coefficient for the partnership effect at level 2. The estimates are close to those in table 1 and there is more variation between individuals for partnership durations than for non-partnership durations. We also note that

there is a negative correlation of -0.52 between partnership and non-partnership episodes indicating that long partnerships tend to be associated with short non-partnerships and vice-versa. Thus individuals can, tentatively, be classified on this basis as either long partnership/short non-partnership or long non-partnership/short partnership individuals. Figures 1, and 2 plot the 'survivor function', i.e. the probability of remaining in a state, for different combinations, derived from the estimates in Table 2. This is the median population estimate derived from the model using the fixed part predictor, i.e. at the mean of the random effect distributions.

Table 3 introduces the time to first partnership as a further response. This response is at the individual level, and may covary with the partnership and non-partnership durations. We find, however, that the variance for the first episode duration is small and poorly estimated with a large standard error. We have therefore omitted it from the analysis. Being in a non-manual occupation increases the probability of leaving the first episode, that is the duration of time to first partnership after age 16 is shorter, and this is shown in Figure 3.

Discussion

We have shown how longitudinal event history data involving repeated episodes for individuals can be modelled so that the between-individual variation in duration can be estimated. The model can deal with multiple states, in the present case partnership and non-partnership and allows correlations to be estimated between the individual level random effects for the different states. While we have only fitted age and social class as

explanatory variables in the present case, we can readily extend this to more variables. Education level achieved and parent's partnership status would be some of the most obvious candidates.

An early age of first partnership is associated with having a manual social class. There is a suggestion that for non-partnership episodes, other than the very first, those in the manual social class also have shorter durations, but the coefficient is non-significant. The earlier the age the partnership starts the shorter the partnership duration, but for non-partnership episodes there is little relationship between age of starting and duration.

The models used in this paper can be extended in a number of directions. For example we can model different kinds of transitions from an episode state, so that partnerships may end in separation, divorce, death etc. We can also fit multivariate models that study the simultaneous durations of different kinds of states such as partnership and employment. Steele et al., (2002) describe how such models can be specified and fitted.

Table 1 Repeated measures models for partnership and non-partnership episodes: starting from first partnership.

Parameter	Estimate (s.e.). PQL1	Estimate (s.e.) MCMC
Fixed		
intercept	-3.936	-4.065
z	$-0.289(0.063)*10^{-1}$	$-0.283\ (0.064) * 10^{-1}$
z^2	$0.132(0.070)*10^{-2}$	$0.160\ (0.078) * 10^{-2}$
z^3	$-0.510(2.283)*10^{-5}$	$-0.584\ (2.331) * 10^{-5}$
z^4	$-0.167(0.160)*10^{-5}$	$0.230\ (0.181) * 10^{-5}$
z^5	$0.347(0.357)*10^{-7}$	$0.451\ (0.410) * 10^{-7}$
start age	-0.039(0.010)	-0.041 (0.011)
manual	-0.113(0.065)	-0.113 (0.072)
np(non-partnership)	1.612(0.396)	1.710 (0.377)
np*z	$0.578(0.164)*10^{-1}$	$0.616\ (0.182) *10^{-1}$
np*z^2	$-0.224(0.135)*10^{-2}$	$-0.249\ (0.147) *10^{-2}$
np* z^3	$-0.145(0.096)*10^{-3}$	$-0.151\ (0.113) *10^{-3}$
np* z^4	$0.695(0.332)*10^{-6}$	$0.754\ (0.383) *10^{-6}$
np* z^5	$0.112(0.132)*10^{-6}$	$0.090\ (0.165) *10^{-6}$
np*start age	0.027(0.016)	0.026 (0.016)
np*manual	0.131(0.105)	0.142 (0.255)
Random		
σ^2_{v0}	0.289 (0.045)	0.377 (0.064)

Note z is centred at 20. In all tables the dummy variable for non-partnership combines with the hazard polynomial, age and social class variables to form interaction terms. Thus, for example the manual – non manual difference coefficient for non-partnerships is given from column B as –0.113+0.142=0.029.

Table 2. Random coefficient model for partnership and outside partnership. MCMC estimates: starting from first partnership.

Parameter	Estimate (s.e.)
Fixed	
intercept	-3.709
z	$-0.244(0.067)*10^{-1}$
z^2	$0.076(0.053)*10^{-2}$
z^3	$-0.140(0.240)*10^{-4}$
z^4	$-0.089(0.114)*10^{-5}$
z^5	$0.093(0.264)*10^{-7}$
start age	-0.063(0.010)
manual	-0.116(0.075)
np(non-partnership)	1.753(0.469)
np*z	$0.499(0.208)*10^{-1}$
np*z^2	$-0.181(0.134)*10^{-2}$
np* z^3	$-0.112(0.114)*10^{-3}$
np* z^4	$0.024(0.332)*10^{-5}$
np* z^5	$0.055(0.158)*10^{-6}$
np*start age	0.049(0.017)
np*manual	0.152(0.118)
Random	
σ_{v0}^2 (non partnership)	0.400(0.211)
σ_{v01}	-0.119(0.118)
σ_{v1}^2 (partnership)	1.145(0.171

Table 3. Random coefficient model for first partnership, partnership and outside partnership. MCMC estimates.

Parameter	Estimate (s.e.)
Fixed	
intercept	-5.227
z	-0.316 (0.036)*10^{-1}
z^2	0.119 (0.021)*10^{-2}
z^3	0.067 (0.083)*10^{-4}
z^4	-0.155 (0.025)*10^{-5}
z^5	0.270 (0.070)*10^{-7}
manual	-0.044 (0.064)
np(non-partnership)	2.746 (0.150)
np*z	0.682 (0.141)*10^{-1}
np*z^2	-0.232 (0.144)*10^{-2}
np* z^3	-0.134 (0.078)*10^{-3}
np* z^4	0.041 (0.344)*10^{-5}
np* z^5	0.850 (1.180)*10^{-7}
np*manual	0.103 (0.120)
fp (time to first partnership)	1.453 (0.067)
fp*z	0.144 (0.005)
fp*z^2	-0.625 (0.034)*10^{-2}
fp* z^3	0.233 (0.129)*10^{-4}
fp* z^4	0.479 (0.074)*10^{-5}
fp* z^5	-0.830 (0.130)*10^{-7}
fp*manual	0.326 (0.071)
Random	
σ^2_{v0} (non partnership)	0.462 (0.061)
σ_{v01}	-0.313 (0.095)
σ^2_{v1} (partnership)	0.773 (0.141)

References

Bynner, J., Elias, P., McKnight, A., Pan, H., et al. (2002). *Young people in transition: changing pathways to employment and independence.* York, Joseph Rowntree Foundation.

Goldstein, H. (1995). Multilevel Statistical Models. London, Edward Arnold: New York. Wiley.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., et al. (2000). *A user's guide to MlwiN (Second Edition).* London, Institute of Education:

Browne, W. and Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational statistics* **15**: 391-420.

Steele, F., Goldstein, H. and Browne, W. (2002). A general multilevel multistate competing risks model for event history data. (submitted for publication).

Figure 1.

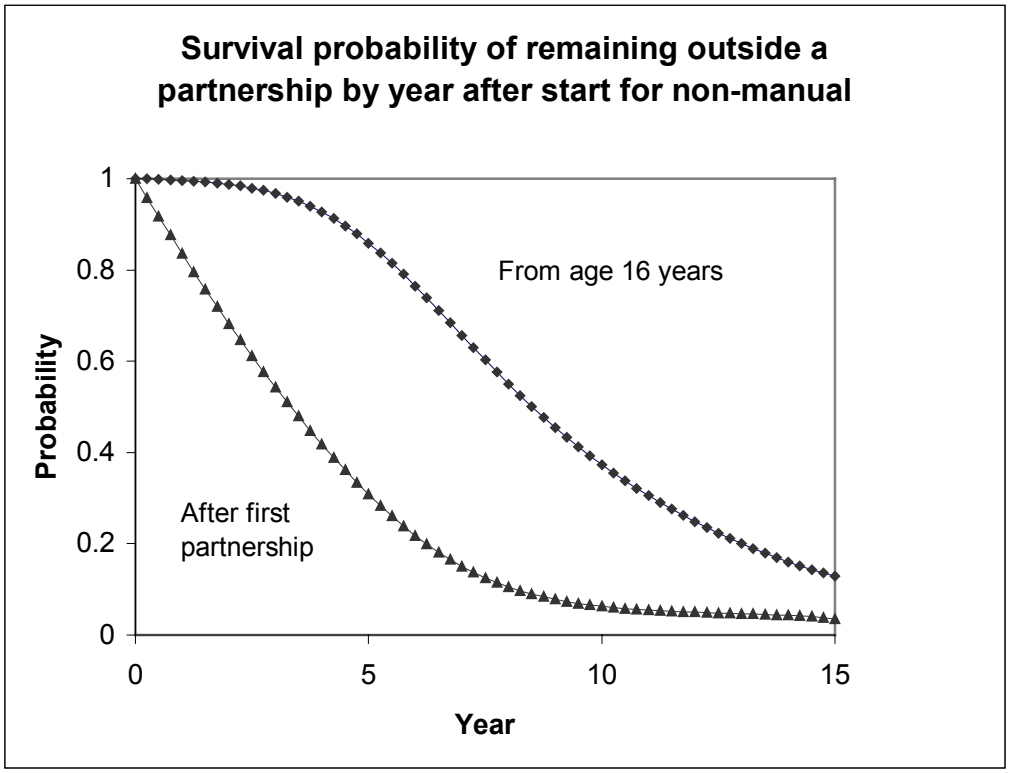

Survival probability of remaining outside a partnership by year after start for non-manual
Probability
1
0.8
0.6
0.4
0.2
0
From age 16 years
After first partnership
0
5
10
15
Year

Figure 2.

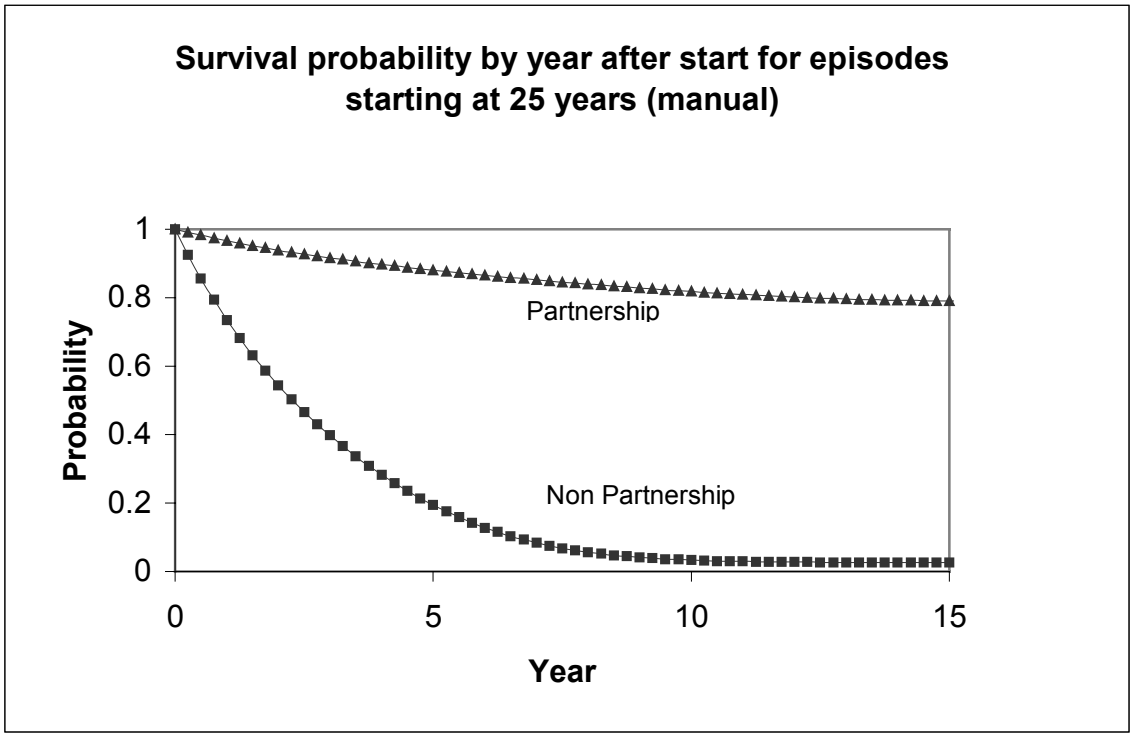
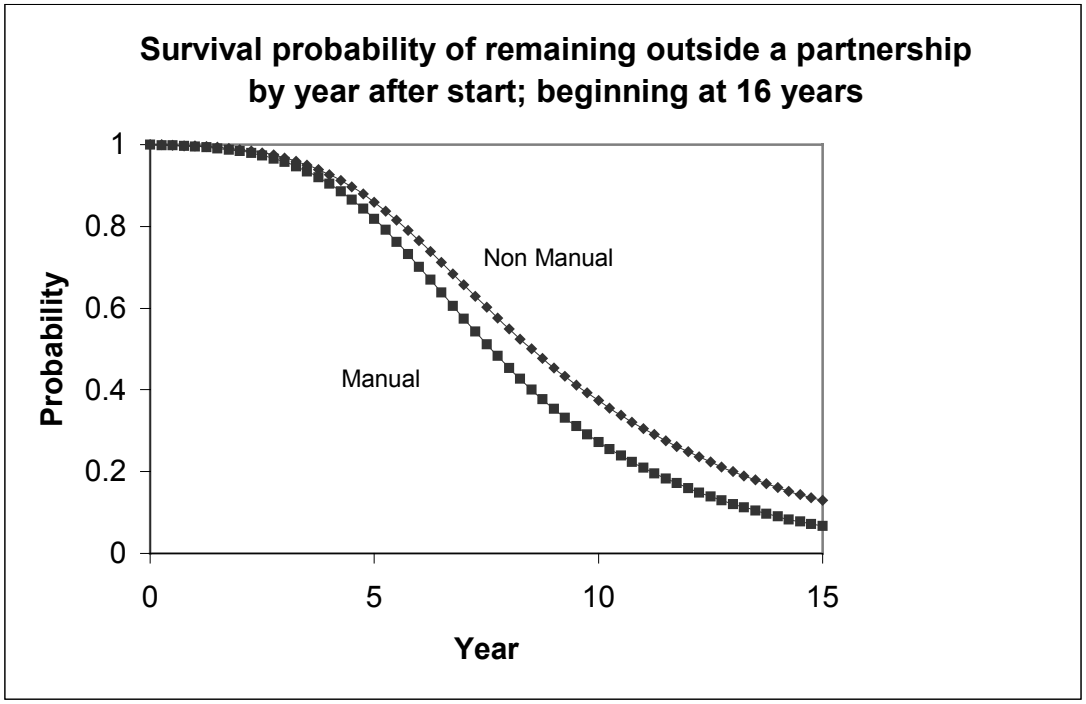

Survival probability by year after start for episodes starting at 25 years (manual)
Probability
1
0.8
0.6
0.4
0.2
0
Partnership
Non Partnership
0
5
10
15
Year

Figure 3.


Survival probability of remaining outside a partnership by year after start; beginning at 16 years
Probability
1
0.8
0.6
0.4
0.2
0
Non Manual
Manual
0
5
10
15
Year