Multilevel Event History Modelling of Birth Intervals

Fiona Steele

Centre for Multilevel Modelling

Graduate School of Education

University of Bristol

Outline

- Why model fertility using event history analysis?
- Overview of discrete-time methods
- Multilevel models for recurrent events and unobserved heterogeneity
- Example: Hutterite birth intervals
- Choice and interpretation of time-varying covariates: endogeneity
- Software

Why Model Fertility Using Event History Analysis?

- Could simply fit a linear/ordered regression to number of children at time of survey (controlling for age)
  - But tells us nothing about *timing* of births and how chance of having a child changes over the life course

- An alternative which allows us to study *timing* of births is to analyse length of birth intervals. Need to use EHA:
  - Many women will not have completed childbearing by time of survey → right-censoring
  - Covariates may have time-varying values, e.g. family composition at start of birth interval, household income, family planning or childcare provision.

Event History Analysis

Methods for analysis of length of time until the occurrence of some event (e.g. a birth or conception). The dependent variable is the duration until event occurrence.

EHA also known as:

- Survival analysis (particularly in biostatistics and when event is not repeatable)
- Duration analysis
- Hazard modelling

Event Times and Censoring Times

Denote the event time (also known as **duration, failure, or survival time**) by the random variable T.

t_j event time for individual j

δ_j censoring indicator
=1 if uncensored (i.e. observed to have event)
=0 if censored

But for right-censored case, we do not observe t_j. We only observe the time at which they were censored, c_j.

Our dependent variable is $y_j = \min(t_j, c_j)$.

Our observed data are (y_j, δ_j).

The Discrete-time Approach

Event times measured in discrete time periods t = 1, 2, 3, . . . (e.g. months, years).

Can think of event history as a series of independent success/failure trials. In each time period t we observe a binary response indicating whether an event has occurred.

Main Advantages of the Discrete-time Approach

- Events times often measured in discrete-time units, particularly when collected retrospectively (e.g. birth dates to nearest month)

- Straightforward to test and allow for non-proportional hazards

- Analysis straightforward as we can use models for discrete response data – important for more complex data structures and processes

Disadvantages of the Discrete-time Approach

- Data must first be restructured so that for each individual we have a sequence of observations, one for each time period until event occurrence or censoring.

- If observation period is long relative to the width of the time periods in which durations are measured, the dataset may become very large.

Restructuring Data for a Discrete-time Analysis: Individual-based File

E.g. records for 2 individuals

INDIVIDUAL (j)	DURATION (t_j)	CENSOR (δ_j)	AGE (x_j)
1	5	0	20
2	3	1	35

CENSOR=1 if uncensored and 0 if censored.

Restructuring Data for a Discrete-time Analysis: Discrete-time (Person-period) File

j	t	$y_j(t)$	x_j
1	1	0	20
1	2	0	20
1	3	0	20
1	4	0	20
1	5	0	20
2	1	0	35
2	2	0	35
2	3	1	35

$y_j(t)$= 1 if event occurs to individual j in period t
= 0 if event has not occurred

The Discrete-time Logit Model (1)

The response variable for a discrete-time model is the binary indicator of event occurrence $y_j(t)$.

The discrete-time **hazard function** is the probability of event in time period t, given event has not occurred before start of t

$$h_j(t) = \Pr(y_j(t) = 1 \mid y_j(t-1) = 0)$$

The Discrete-time Logit Model (2)

We can fit a logit regression model of the form:

$$\text{logit}\,[h_j(t)] = \log\left[\frac{h_j(t)}{1-h_j(t)}\right] = \alpha(t) + \beta\, x_j(t) \qquad (*)$$

The covariates $x_j(t)$ can be constant over time or time-varying.

$\alpha(t)$ is some function of time, called the logit of the **baseline hazard function**. This needs to be specified.

Modelling the Time-dependency of the Hazard (1)

Changes in *h*(*t*) over time are captured by α(*t*). This might be a **linear** or **quadratic** function.

Linear: $\alpha(t) = \alpha_0 + \alpha_1 t$

Quadratic: $\alpha(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2$

Modelling the Time-dependency of the Hazard (2)

In the most flexible model, time is treated as a categorical variable with a category for each time period:

$$\alpha(t) = \alpha_1 D_1 + \alpha_2 D_2 + \ldots + \alpha_q D_q$$

where D_1, D_2, . . ., D_q are dummies for time periods t=1, 2, . . ., q, and q is the maximum observed event time. (Alternatively choose one period as the reference and fit overall intercept.)

If q is very large, categories can be grouped together – piecewise constant hazard model.

The Proportional Hazards Assumption

Model (*) assumes that the effects of covariates $x(t)$ are constant over time. This is known as the **proportional hazards** assumption. (Strictly it is the ***odds*** that are assumed proportional as we are fitting a logit model.)

We can relax this assumption by introducing interactions between $x(t)$ and $\alpha(t)$.

Unobserved Heterogeneity

- Some individuals more at risk of an event than others for reasons that are not fully captured by covariates
  - E.g. some women will conceive quicker than others for biological reasons.

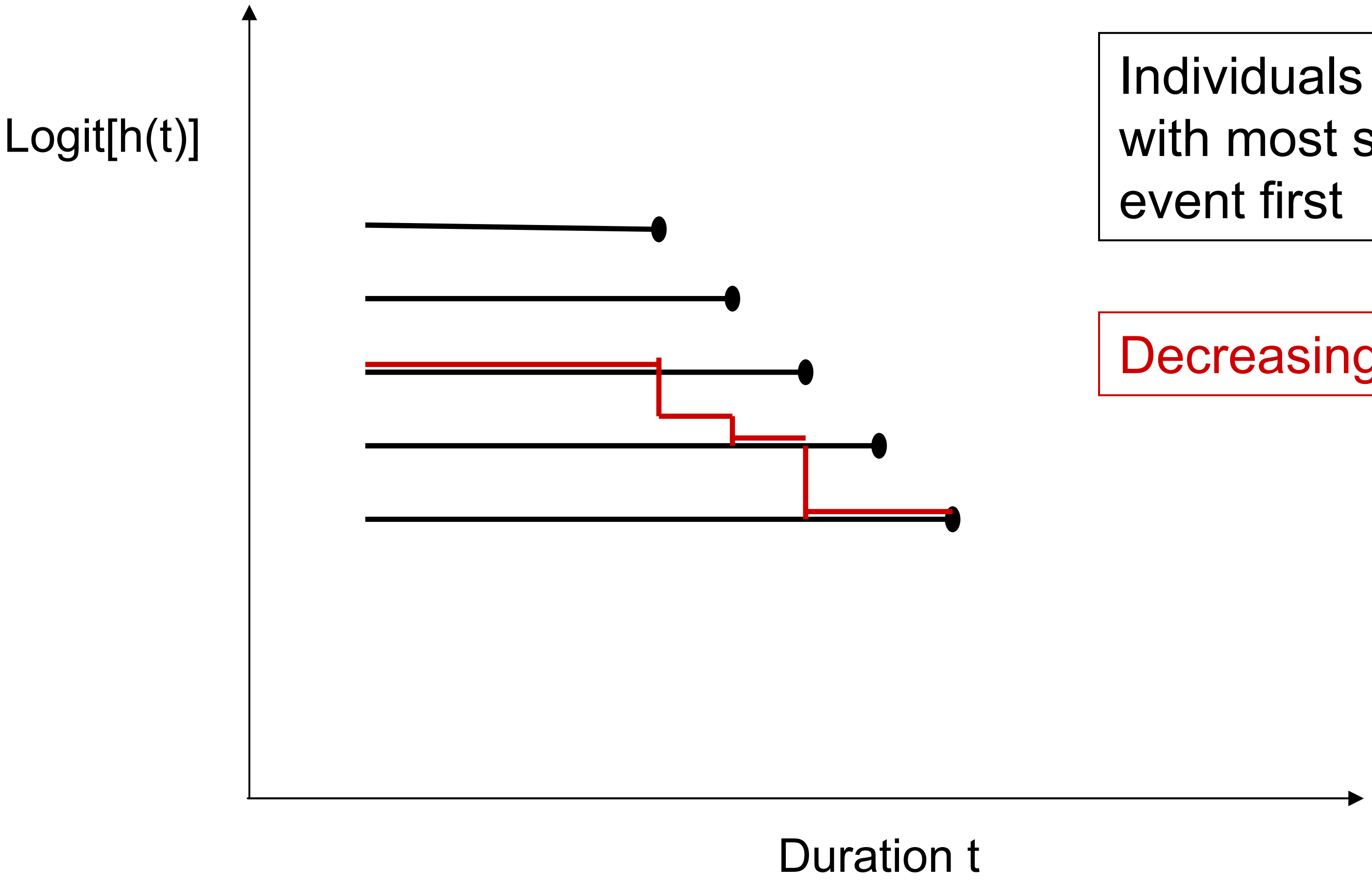
- Presence of unobserved individual-specific risk factors leads to **unobserved heterogeneity** in the hazard
  - Also referred to as **frailty**, particularly in biostatistics (e.g. more ‘frail’ individuals have a higher mortality risk)

Consequences of Unobserved Heterogeneity

- If there are individual-specific unobserved factors that affect the hazard, the observed form of the hazard function at the aggregate population level will tend to be different from individual hazards

- Even if all individuals have a constant hazard, the aggregate population hazard may be time-dependent, typically decreasing. This may be explained by a **selection effect** operating on individuals

Illustration of Selection



Individuals with constant hazard with most susceptible having event first

Decreasing population hazard

Allowing for Unobserved Heterogeneity in a Discrete-Time Model

We can introduce a random effect which represents individual-specific unobservables:

$$\operatorname{logit}[h_j(t)] = \alpha(t) + \beta x_j(t) + u_j$$

Usually assume $u_j \sim N(0, \sigma_u^2)$

But difficult to estimate var(u) if only one event per individual. Really need recurrent events

Examples of Recurrent Events

Many types of event can occur more than once to an individual. Define an episode as the time between the start of the ‘risk’ period and the occurrence of an event or censoring.

Birth intervals: duration between birth of one child and conception of the next

Employment episode: duration from starting a new job to leaving that job.

Marriage episode: duration of a marriage.

Problem with Analysing Recurrent Events

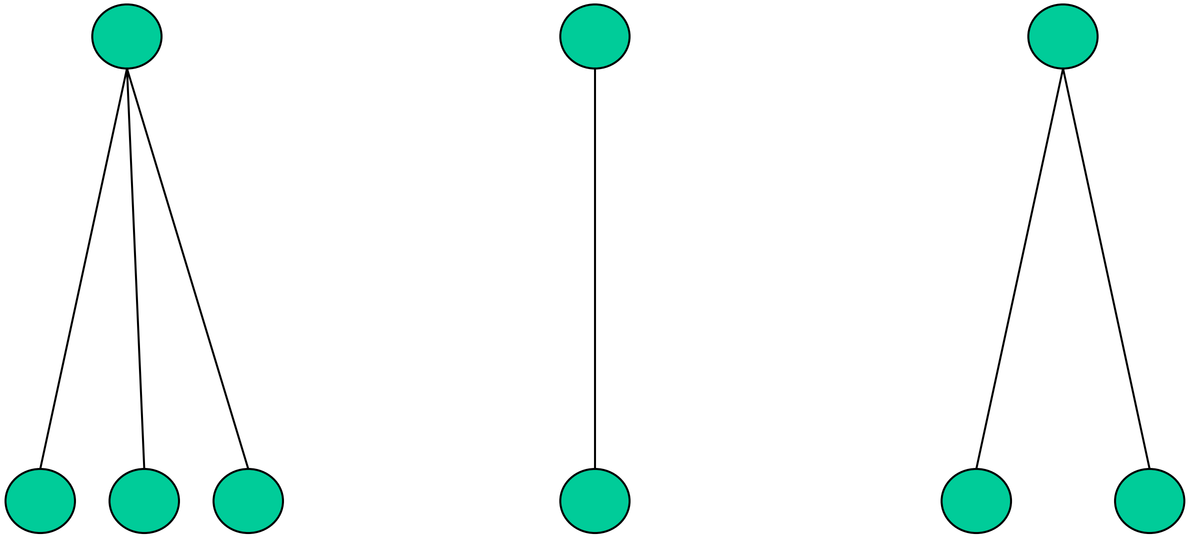
We cannot assume that the durations of episodes from the same individual are independent.

There may be unobserved individual-specific factors (i.e. constant across episodes) which affect the hazard of an event for all episodes.

Presence of such unobservables, and failure to account for them in the model, will lead to correlation between the durations of episodes from the same individual.

Hierarchical Data Structure

Recurrent events lead to a two-level hierarchical structure.



Level 2: Individuals

Level 1: Episodes

Simple Two-level Model for Recurrent Events (1)

$$\text{logit}[h_{ij}(t)] = \alpha(t) + \beta x_{ij}(t) + u_j$$

$h_{ij}(t)$ is hazard of event in time period *t* during episode *i* of individual *j*

$x_{ij}(t)$ are covariates which might be time-varying or defined at the episode or individual level

u_j random effect representing unobserved characteristics of individual *j* – **shared 'frailty'** (common to all episodes)

Assume $u_j \sim N(0, \sigma_u^2)$

Simple Two-level Model for Recurrent Events (2)

- The model for recurrent events is essentially the same as the model for unobserved heterogeneity, and is therefore estimated in exactly the same way.
- Recurrent events allow better identification of the random effect variance.
- The expansion of data to discrete-time format is carried out for each episode within an individual.

Why Not Fit A Separate Model for Each Episode?

- Inefficient as many covariates will have the same effect for first, second, third etc. event
  - If we analyse episodes jointly, we can test if this is the case and allow for episode-specific effects if necessary
- Modelling each event separately may lead to misleading conclusions because of unobserved heterogeneity (see Kravdal, 2001)

Modelling Birth Rates in Norway (Kravdal, 2001)

- In separate analyses of birth intervals, find counter-intuitive positive effects of education on rates of 2nd and 3rd births
- Effect of education becomes negative when intervals modelled jointly with control for unobserved heterogeneity
- **Explanation**: Compare 2 women each having 2nd birth at age 30. This might be the norm for an educated women but late for a less-educated woman, so we are not comparing like with like. They differ on unobserved factors associated with fertility timing

Episode-specific Effects

We can allow the duration and covariate effects to vary between episodes. E.g. we might expect factors affecting the timing of the first birth to differ from those affecting timing of subsequent births (or the same factors to have different effects).

Include dummy variable(s) for order of the event and interact with *t* and covariates.

Example

Repeated birth intervals for Hutterite women, a natural fertility (no contraceptive use) population in North America.

Data on 944 birth intervals from 159 women. Interval is duration between a birth and conception of next child.

Only closed intervals (i.e. ending in a conception before the survey). Long open intervals may be due to primary or secondary sterility. Therefore there is no censoring.

Duration and Covariates

Duration of a birth interval is represented by a categorical variable (in a piecewise constant hazards model):

MONTH Month of exposure to risk of conception
[<6, 6-11, 12-23, 24-35, 36+ (ref.)]

Consider one covariate:

AGE Age at start of birth interval (years)

Results

	Coeff.	(SE)
Const	0.38	(0.39)
MONTH		
<6	-0.96	(0.30)
6-11	-0.21	(0.30)
12-23	0.12	(0.30)
24-35	-0.24	(0.36)
AGE	-0.07	(0.01)
σ_u^2	0.31	(0.06)

Random Coefficient Models

So far assumed that unobserved heterogeneity is constant. We have considered a random intercept model, where hazard is shifted up or down by an amount u_j on the logit scale.

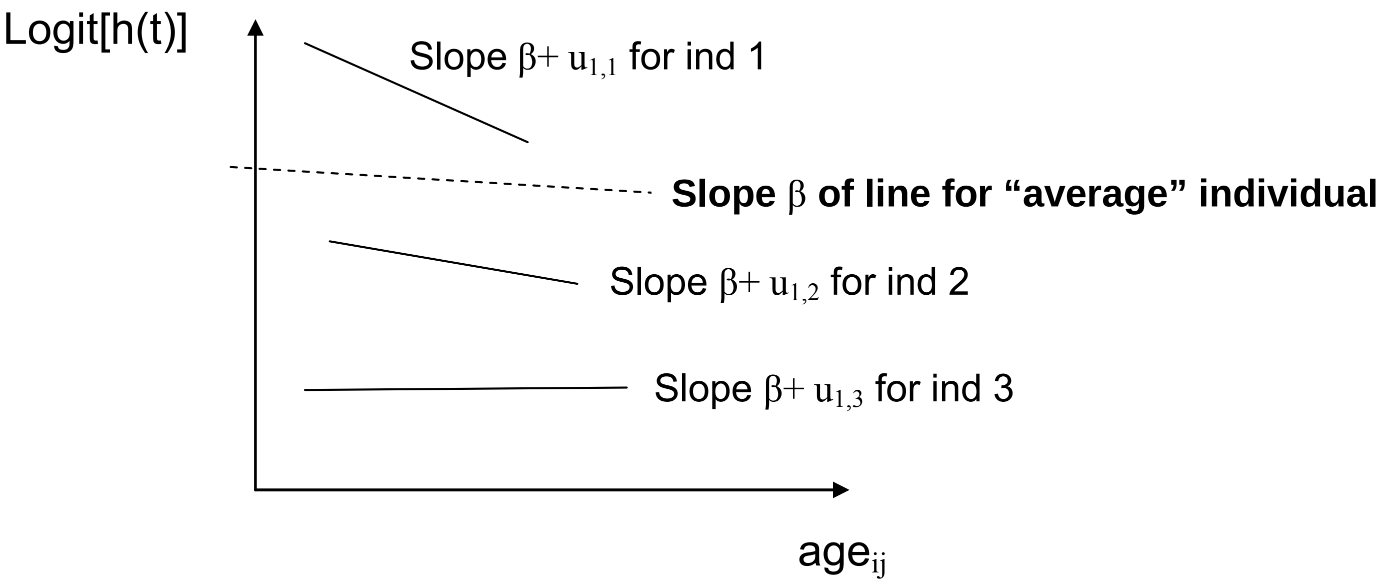
But the duration and covariate effects are assumed to be the same for each individual.

To test these assumptions, we can consider random coefficient/slope models.

But note there may be insufficient information to estimate random coefficients if the number of individuals with recurrent events is small.

Example: Random Coefficient for Age

For any time interval *t*:


Logit[h(t)]
Slope $\beta + u_{1,1}$ for ind 1
Slope β of line for "average" individual
Slope $\beta + u_{1,2}$ for ind 2
Slope $\beta + u_{1,3}$ for ind 3
age_{ij}

Fitting a Random Coefficient for AGE

$$\operatorname{logit}[h_{ij}(t)] = \alpha(t) + \beta_j \, AGE_{ij} + u_{0j}$$

where $\beta_j = \beta + u_{1j}$

Also written as $\operatorname{logit}[h_{ij}(t)] = \alpha(t) + \beta \, AGE_{ij} + u_{0j} + AGE_{ij} u_{1j}$

Assume $u_{0j} \sim N(0, \sigma^2_{u0}), \qquad u_{1j} \sim N(0, \sigma^2_{u1})$

$\text{Cov}(u_{0j}, u_{1j}) = \sigma_{u01}$

Unobserved Heterogeneity as a Function of AGE

A random coefficient for AGE implies that the unobserved heterogeneity between women is:

$$\mathrm{Var}(u_{0j} + AGE_{ij}u_{1j})$$

$$= Var(u_{0j}) + 2Cov(u_{0j}, u_{1j})AGE_{ij} + Var(u_{1j})AGE_{ij}^2$$

$$= \sigma_{u0}^2 + 2\sigma_{u01}AGE_{ij} + \sigma_{u1}^2 AGE_{ij}^2$$

i.e. a quadratic function in AGE.

Results from Random Coefficient Model (AGE centred around mean)

	Coeff.	(SE)
Const	-1.55	(0.30)
MONTH		
<6	-0.99	(0.31)
6-11	-0.21	(0.30)
12-23	0.14	(0.30)
24-35	-0.21	(0.37)
AGE	-0.07	(0.01)
σ_{u0}^2	0.29	(0.85)
σ_{u01}	-0.007	(0.005)
σ_{u1}^2	0.001	(0.001)

Testing Significance of Random Coefficient

2 additional parameters introduced to random intercept model:

σ_{u1}^2 and σ_{u01}

Test the null hypothesis that $H_0 : \sigma_{u1}^2 = \sigma_{u01} = 0$

The (approximate) Wald test statistic is 2.84 on 2 d.f.

So fail to reject null and conclude that the effect of AGE is constant across women.

Interpretation of Random Coefficient for AGE

In practice, we would revert to the random intercept model after finding little evidence of a random coefficient. But, for illustration, we consider the interpretation of the random coefficient model.

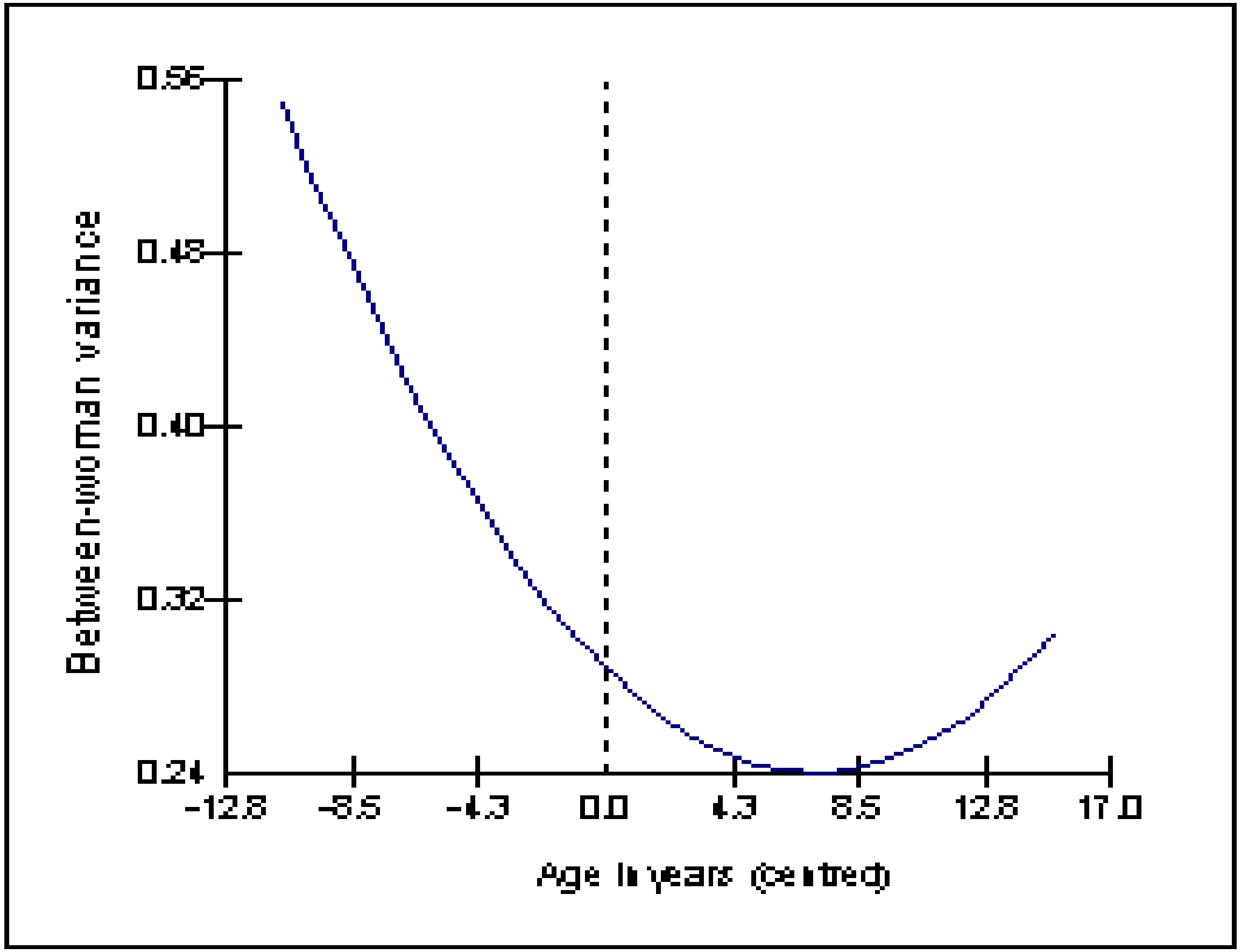
The between-woman variance in the logit-hazard of conception (after accounting for duration effects) is:

$$\mathrm{Var}(u_{0j} + AGE_{ij}u_{1j})$$

$$= 0.29 - 0.014\,AGE_{ij} + 0.001\,AGE_{ij}^{2}$$

We can plot the between-woman variance as a function of AGE.

Between-woman Unobserved Heterogeneity as a Function of AGE



Choice and Interpretation of Time-Varying Covariates

- Think carefully about choice of covariates X, especially those that are time-varying. Could values of X be co-determined with timing of event of interest?
  - E.g. timing of births and women's employment. Does having children delay career progression and/or does having a high-level job lead a woman to delay childbearing?

- One approach to the problem is to model birth intervals and employment jointly in a multilevel multiprocess model

Software

- Essentially multilevel models for binary responses
- Mainstream software: Stata (xtlogit, xtmixed), SAS (PROC NLMIXED), R
- Specialist multilevel modelling software: aML, MLwiN, Sabre (standalone or plug-ins for Stata and R)

Training Resources

- CMM online course. Module 7 on ‘Multilevel Models for Binary Responses’ with MLwiN practical (Stata practical to be added soon)
- Materials from 2-day course (including practicals), MLwiN macros and aML syntax from http://www.cmm.bris.ac.uk/research/Multiprocess/index.shtml
- Multilevel modelling software reviews (including syntax for a range of models) from http://www.cmm.bris.ac.uk/learning-training/multilevel-m-software/index.shtml

Bibliography:

Modelling the Time to a Single Event

Allison, P.D. (1982) "Discrete-time methods for the analysis of event history data." In Sociological Methodology (Ed. S. Leinhardt). San Francisco: Jossey-Bass.

Singer, J.D. and Willet, J.B. (1993) "It's about time: Using discrete-time survival analysis to study duration and the timing of events." *Journal of Educational Statistics*, **18**: 155-195.

Bibliography:

Models for Recurrent Events and Unobserved Heterogeneity

Davis, R.B., Elias, P. and Penn, R. (1992) “The relationship between a husband’s unemployment and a wife’s participation in the labour-force.” *Oxford Bulletin of Economics and Statistics*, **54**:145-71.

Kravdal, Ø. (2001) “The high fertility of college educated women in Norway: An artefact of the separate modelling of each parity transition.” *Demographic Research*, **5**, Article 6.

Bibliography:

Multiprocess Models for Correlated Histories

Brien, M.J., Lillard, L.A. and Waite, L.J. (1999) “Interrelated family-building behaviors: Cohabitation, marriage, and nonmarital conception.” *Demography*, **36**: 535-551.

Steele, F., Kallis, C., Goldstein, H. and Joshi, H. (2005) “The relationship between childbearing and transitions from marriage and cohabitation in Britain.” *Demography*, **42**: 647-673.

Upchurch, D.M., Lillard, L.A. and Panis, C.W.A. (2002) “Nonmarital childbearing: Influences of education, marriage, and fertility.” *Demography*, **39**: 311-329.

Waite, L.J. and Lillard, L.A. 1991. “Children and marital disruption.” *American Journal of Sociology*, **96**: 930-53.