Multilevel Models for Longitudinal Data

Fiona Steele

Aims of Talk

- Overview of the application of multilevel (random effects) models in longitudinal research, with examples from social research
- Particular focus on joint modelling of correlated processes using multilevel multivariate models, e.g. to adjust for selection bias in estimating effect of parental divorce on children's education

Longitudinal Research Questions and Models

Consider multilevel models for:

- Change over time
  - Growth curve (latent trajectory) models
    E.g. Do child developmental processes (academic ability, behaviour etc.) differ for boys and girls, or by parental characteristics?
  - Dynamic (autoregressive) models
    E.g. Is there a *causal* effect of test score at time t on the score at $t+1$?
- Time to event occurrence
  - Event history analysis
    E.g. What are risk factors of divorce? What is the impact of divorce on children's educational careers?

Modelling Change

Repeated Measures Data

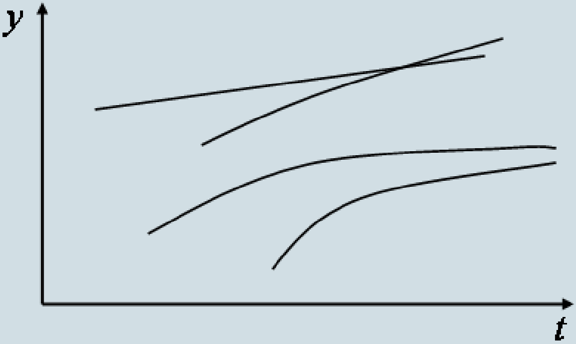
Denote by y_{ti} the response at occasion t ($t = 1, \ldots, T_i$) for individual i ($i = 1, \ldots, n$).

- Occasions need not be equally spaced
- In many applications time $\equiv$ age (e.g. developmental processes) and, at a given t, individuals vary in age
- Individuals may have missing data

Can view data as having a 2-level hierarchical structure: responses (level 1) within individuals (level 2).

Examples of Growth Curves

Growth curve models posit the existence of individual underlying trajectories. The pattern of y over time provides information on these trajectories.



Individuals may vary in their initial level of y *and* their growth rate.

Linear Growth Model

Denote by z_{ti} the timing of occasion t for individual i. Suppose y_{ti} is a linear function of z_{ti} and covariates $\mathbf{x}_{ti}$.

$$y_{ti} = \alpha_{0i} + \alpha_{1i} z_{ti} + \beta x_{ti} + e_{ti}$$

$$\alpha_{0i} = \alpha_0 + u_{0i} \qquad \text{(individual variation in level of } y\text{)}$$

$$\alpha_{1i} = \alpha_1 + u_{1i} \qquad \text{(individual variation in growth rate)}$$

where u_{0i} and u_{1i} are individual-level residuals $\sim$ bivariate normal and e_{ti} are i.i.d. normally distributed occasion-level residuals. Residuals at both levels assumed uncorrelated with x_{ti}.

Linear Growth Model

Denote by z_{ti} the timing of occasion t for individual i. Suppose y_{ti} is a linear function of z_{ti} and covariates $\mathbf{x}_{ti}$.

$$y_{ti} = \alpha_{0i} + \alpha_{1i} z_{ti} + \beta x_{ti} + e_{ti}$$

$$\alpha_{0i} = \alpha_0 + u_{0i} \qquad \text{(individual variation in level of } y\text{)}$$

$$\alpha_{1i} = \alpha_1 + u_{1i} \qquad \text{(individual variation in growth rate)}$$

where u_{0i} and u_{1i} are individual-level residuals $\sim$ bivariate normal and e_{ti} are i.i.d. normally distributed occasion-level residuals. Residuals at both levels assumed uncorrelated with x_{ti}.

Frame as a multilevel 'random slopes' model *or* a SEM (Curran 2003[†]).

[†] *Multivariate Behavioral Research*, **38**: 529-568.

Advantages of Multilevel Approach

- No need to have balanced design or equally spaced measurements
  - Individuals may vary in their number of measurements by design or due to attrition
  - Individuals with missing y included under a **missing at random** assumption

Advantages of Multilevel Approach

- No need to have balanced design or equally spaced measurements
  - Individuals may vary in their number of measurements by design or due to attrition
  - Individuals with missing y included under a **missing at random** assumption
- Straightforward to allow for between-individual variation in the timing of measurement t

Advantages of Multilevel Approach

- No need to have balanced design or equally spaced measurements
  - Individuals may vary in their number of measurements by design or due to attrition
  - Individuals with missing y included under a **missing at random** assumption
- Straightforward to allow for between-individual variation in the timing of measurement t
- Flexibility in specification of dependency of y on z, e.g. polynomial, spline, step function

Advantages of Multilevel Approach

- No need to have balanced design or equally spaced measurements
  - Individuals may vary in their number of measurements by design or due to attrition
  - Individuals with missing y included under a **missing at random** assumption
- Straightforward to allow for between-individual variation in the timing of measurement t
- Flexibility in specification of dependency of y on z, e.g. polynomial, spline, step function
- Can allow for clustering at higher levels, e.g. geography

Example: Development in Reading Ability

- Reading scores for 221 children on four occasions (only complete cases considered)*
- Occasions spaced two years apart (1986, 1988, 1990 and 1992); children aged 6-8 in 1986

*Dataset from `http://www.duke.edu/~curran/`

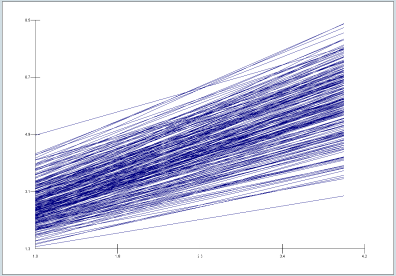
Example: Development in Reading Ability

- Reading scores for 221 children on four occasions (only complete cases considered)*
- Occasions spaced two years apart (1986, 1988, 1990 and 1992); children aged 6-8 in 1986
- Model children's reading trajectories over the four occasions as a linear function of time ($z_{ti} = z_t$), with origin at 1986
- Allow initial reading score (intercept) and progress (slope) to vary across individuals

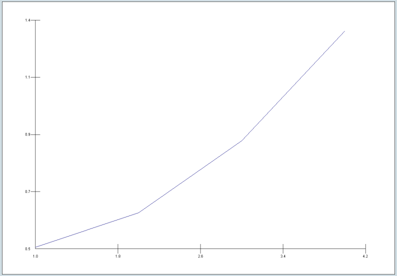
---

*Dataset from `http://www.duke.edu/~curran/`

Predicted Trajectories



Between-Child Variance



Multivariate Growth Curve Models

Suppose x_{ti} and y_{ti} are outcomes of correlated processes, e.g. reading and maths ability.

Unmeasured influences on y_{ti} (represented by u_{0i} and u_{1i}) might also affect x_{ti}. We can model changes in y_{ti} and x_{ti} jointly:

$$
\begin{aligned}
y_{ti} &= \alpha_{0i}^{(y)} + \alpha_{1i}^{(y)} z_{ti}^{(y)} + e_{ti}^{(y)} \\
x_{ti} &= \alpha_{0i}^{(x)} + \alpha_{1i}^{(x)} z_{ti}^{(x)} + e_{ti}^{(x)}
\end{aligned}
$$

where $\alpha_{ki}^{(y)} = \alpha_k^{(y)} + u_{ki}^{(y)}$ and $\alpha_{ki}^{(x)} = \alpha_k^{(x)} + u_{ki}^{(x)}$, $k = 0, 1$.

Multivariate Growth Curve Models

Suppose x_{ti} and y_{ti} are outcomes of correlated processes, e.g. reading and maths ability.

Unmeasured influences on y_{ti} (represented by u_{0i} and u_{1i}) might also affect x_{ti}. We can model changes in y_{ti} and x_{ti} jointly:

$$
\begin{aligned}
y_{ti} &= \alpha_{0i}^{(y)} + \alpha_{1i}^{(y)} z_{ti}^{(y)} + e_{ti}^{(y)} \\
x_{ti} &= \alpha_{0i}^{(x)} + \alpha_{1i}^{(x)} z_{ti}^{(x)} + e_{ti}^{(x)}
\end{aligned}
$$

where $\alpha_{ki}^{(y)} = \alpha_k^{(y)} + u_{ki}^{(y)}$ and $\alpha_{ki}^{(x)} = \alpha_k^{(x)} + u_{ki}^{(x)}$, $k = 0, 1$.

Equations linked via cross-process correlations among residuals defined at the same level.

Joint Model of Reading (R) & Antisocial Behaviour (B)

Correlations between child-specific random effects:

	R intercept	B intercept	R slope	B slope
R intercept	1			
B intercept	−0.08	1		
R slope	0.12	−0.44	1	
B slope	−0.06	0.42	0.30	1

Only the correlation between the behaviour intercept and the reading slope is significant at 5%. Worse-than-average behaviour at year 1 ($u_{0i}^{(B)} > 0$) associated with below-average reading progress ($u_{1i}^{(R)} < 0$).

Note that we cannot infer that behaviour at $t = 1$ predicts future reading in any causal sense.

Dynamic (Autoregressive) Models

1st order autoregressive, AR(1), model:

$$y_{ti} = \alpha_0 + \alpha_1 y_{t-1,i} + \beta x_{ti} + u_i + e_{ti}, \qquad t = 2, 3, \ldots, T$$

where $u_i \sim N(0, \sigma_u^2)$, $\text{cov}(x_{ti}, u_i) = 0$ and $e_{ti} \sim N(0, \sigma_e^2)$

Dynamic (Autoregressive) Models

1st order autoregressive, AR(1), model:

$$y_{ti} = \alpha_0 + \alpha_1 y_{t-1,i} + \beta x_{ti} + u_i + e_{ti}, \qquad t = 2, 3, \ldots, T$$

where $u_i \sim N(0, \sigma_u^2)$, $\text{cov}(x_{ti}, u_i) = 0$ and $e_{ti} \sim N(0, \sigma_e^2)$

- α_1 is assumed the same for all individuals (and often for all t)
- Effect of y_1 on a subsequent y_t is α_1^{t-1}, so diminishes with t for $|\alpha_1| < 1$
- Residual correlation between y_{ti} and $y_{t-1,i}$ is $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$

State Dependence vs. Unobserved Heterogeneity

Is correlation between y_t and y_{t-1} due to:

- **Causal** effect of y_{t-1} on y_t?
  $\Rightarrow |\alpha_1|$ close to 1 and ρ close to 0 (state dependence)

- Mutual dependence on time-invariant omitted variables?
  $\Rightarrow |\alpha_1|$ close to 0 and ρ close to 1 (unobserved heterogeneity)

State Dependence vs. Unobserved Heterogeneity

Is correlation between y_t and y_{t-1} due to:

- **Causal** effect of y_{t-1} on y_t?
  $\Rightarrow |\alpha_1|$ close to 1 and ρ close to 0 (state dependence)

- Mutual dependence on time-invariant omitted variables?
  $\Rightarrow |\alpha_1|$ close to 0 and ρ close to 1 (unobserved heterogeneity)

E.g. Explanations for persistently high/low crime rates in areas:

- Current crime rate determined by past crime rate
- Dependence of crime rate at all t on unmeasured area-specific characteristics (unemployment, social cohesion etc)

Example: Dynamic Analysis of Reading Ability

Effects on standardised reading score at year t

Variable	Estimate	SE
Reading at $t-1$	0.34	0.07
Year 1990 (vs 1988)	0.28	0.07
Year 1992 (vs 1988)	0.56	0.11
Cognitive support at home	0.10	0.03
Male	-0.05	0.06

$\hat{\alpha}_1$=0.34 (se=0.07) and $\hat{\rho} = 0.60$

No clear pattern: evidence of state dependence and substantial unobserved heterogeneity

The Initial Conditions Problem

y_1 may not be measured at the start of the process

Can view as a missing data problem:

$$\begin{array}{ll} \text{Observed} & (y_1, \ldots, y_T) \\ \text{Actual} & (y_{-k}, \ldots, y_0, y_1, \ldots, y_T) \end{array}$$

where first $k+1$ measures are missing.

We need to specify a model for y_1 (not just condition on y_1).

Modelling the Initial Condition

Common assumptions:

- **Short-run** Treat $t = 1$ as the start of the process, but need to allow for time-invariant unobservables affecting y_{1i} and $(y_{2i}, \ldots, y_{Ti})$
- **Long-run** Allow for possibility that process is already underway by $t = 1$, and regard y_{1i} as informative (about past and future y)

In a random effects framework, specify a model for y_{1i} and estimate jointly with the model for $(y_{2i}, \ldots, y_{Ti})$.

A widely used alternative approach (without parametric assumptions) is Generalised Method of Moments.†

† e.g. Arellano and Honoré (2001) *Handbook of Econometrics*, vol. 5.

Dynamic Model with Endogenous x_{ti}

x_{ti} may be jointly determined with y_{ti} (subject to shared or correlated time-invariant unobserved influences), i.e. $\text{cov}(x_{ti}, u_i) \neq 0$

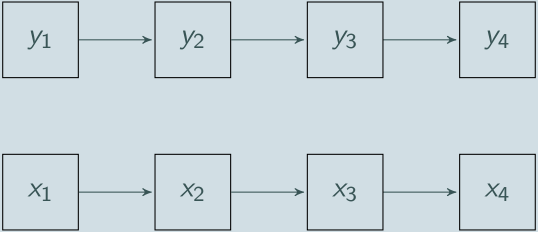
In addition, the relationship between x and y may be bi-directional.

$\Rightarrow$ Fully simultaneous bivariate model (cross-lagged SEM):

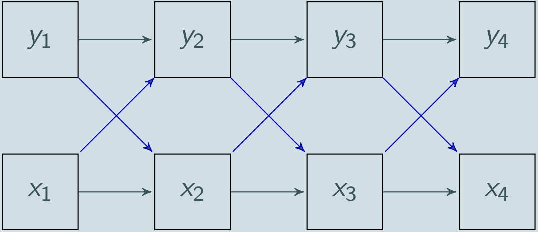
$$
\begin{aligned}
y_{ti} &= \alpha_0^{(y)} + \alpha_1^{(y)} y_{t-1,i} + \beta^{(y)} x_{t-1,i} + u_i^{(y)} + e_{ti}^{(y)} \\
x_{ti} &= \alpha_0^{(x)} + \alpha_1^{(x)} x_{t-1,i} + \beta^{(x)} y_{t-1,i} + u_i^{(x)} + e_{ti}^{(x)}
\end{aligned}
$$

where $\text{cov}\left(u_i^{(y)}, u_i^{(x)}\right)$ is freely estimated.

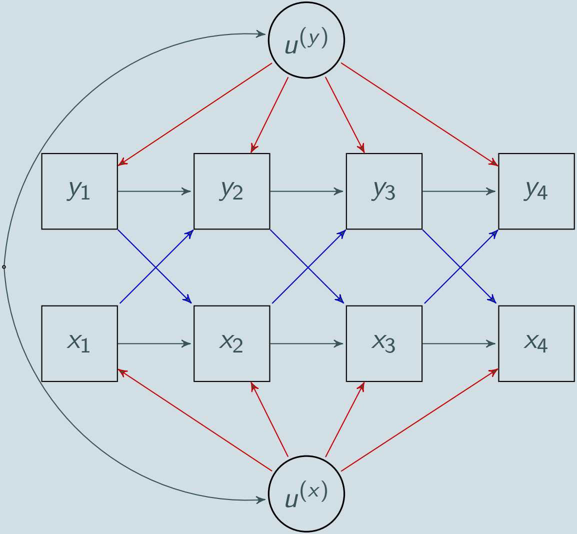
Cross-lagged Structural Equation Model for $T=4$



Cross-lagged Structural Equation Model for T=4



Cross-lagged Structural Equation Model for $T=4$



Modelling Event Occurrence

Multilevel Event History Analysis

Multilevel event history data arise when events are repeatable (e.g. births, partnership dissolution) or where individuals are organised in groups.

Suppose events are repeatable, and define an **episode** as a continuous period for which an individual is at risk of experiencing an event.

Denote by y_{ij} the duration of episode j of individual i, which is fully observed if an event occurs ($\delta_{ij} = 1$) and right-censored if not ($\delta_{ij} = 0$).

Discrete-Time Data

In social research, event history data are usually collected in one of two ways:

- retrospectively in a cross-sectional survey, where dates are recorded to the nearest month or year
- prospectively in irregularly-spaced waves of a panel study (e.g. annually)

Both give rise to discretely-measured durations.

We can convert the observed data (y_{ij}, δ_{ij}) to a sequence of binary responses $\{y_{tij}\}$ where y_{tij} indicates whether an event has occurred in time interval $[t, t+1)$.

Discrete-Time Data Structure

individual i	episode j	y_{ij}	δ_{ij}
1	1	2	1
1	2	3	0

Discrete-Time Data Structure

individual i	episode j	y_{ij}	δ_{ij}
1	1	2	1
1	2	3	0

$\downarrow$

individual i	episode j	t	y_{tij}
1	1	1	0
1	1	2	1
1	2	1	0
1	2	2	0
1	2	3	0

Multilevel Discrete-time Event History Model

Hazard function

$$h_{tij}(t) = Pr[y_{tij} = 1 | y_{t-1,ij} = 0]$$

Logit model

$$\text{logit}(h_{tij}) = \alpha_t + \beta x_{tij} + u_i$$

α_t is a function of cumulative duration t

$u_i \sim N(0, \sigma_u^2)$ allows for unobserved heterogeneity ('shared frailty') between individuals due to time-invariant omitted variables

Multilevel Event History Analysis: Extensions

- Competing risks More than one type of transition or event may lead to the end of an episode, e.g. different reasons for leaving a job $\rightarrow$ multinomial event occurrence indicator y_{tij}

Multilevel Event History Analysis: Extensions

- Competing risks More than one type of transition or event may lead to the end of an episode, e.g. different reasons for leaving a job $\longrightarrow$ multinomial event occurrence indicator y_{tij}

- Multiple states Individuals may pass through different 'states' (e.g. employed, unemployed). Allow for residual correlation among transitions in a joint model (negative correlation between transitions in and out of unemployment?)

Multilevel Event History Analysis: Extensions

- Competing risks More than one type of transition or event may lead to the end of an episode, e.g. different reasons for leaving a job $\longrightarrow$ multinomial event occurrence indicator y_{tij}

- Multiple states Individuals may pass through different 'states' (e.g. employed, unemployed). Allow for residual correlation among transitions in a joint model (negative correlation between transitions in and out of unemployment?)

- Multiple processes Time-varying covariates are often outcomes of a correlated process, and can be modelled jointly with process of interest. E.g. employment, childbearing and partnership transitions all co-determined†

---

†Aassve et al. (2006) *J. Roy. Stat. Sci. A*, **169**: 781-804.

Correlated Event Processes

Example: Marital and birth histories[†]

y_{ij} is duration of marriage j of person i

z_{tij} is number of children from marriage j at start of interval $[t, t+1)$, an outcome of a birth history

---

[†] Lillard and Waite (1993) *Demography*, **30**: 653-681.

Correlated Event Processes

Example: Marital and birth histories[†]

y_{ij} is duration of marriage j of person i

z_{tij} is number of children from marriage j at start of interval $[t, t+1)$, an outcome of a birth history

Unobserved individual characteristics affecting risk of marital dissolution may be correlated with those affecting probability of a birth (or conception) during marriage, i.e. z_{tij} may be endogenous w.r.t. y_{ij}

[†] Lillard and Waite (1993) *Demography*, **30**: 653-681.

Simultaneous Equations Model for Multiple Event Processes

h^D_{tij} Hazard of marital dissolution in interval $[t, t+1)$

h^C_{tij} Hazard of a conception (leading to live birth) in $[t, t+1)$

Bivariate hazards model:

$$\begin{aligned} \text{logit}(h^D_{tij}) &= \alpha^D_t + \beta^D x^D_{tij} + \gamma z_{tij} + u^D_i \\ \text{logit}(h^C_{tij}) &= \alpha^C_t + \beta^C x^C_{tij} + u^C_i \end{aligned}$$

where $\text{cov}(u^D_i, u^C_i)$ is freely estimated

Simultaneous Equations Model: Identification

Two approaches:

- **Covariate exclusion restrictions** Find at least one variable that affects hazard of conception but not hazard of dissolution (an instrument). Often difficult to find in practice.
- **Replication** Use fact that some individuals have more than one marriage and more than one birth, allowing estimation of within-person effect of number of children. Assume residual correlation is between time-invariant characteristics.

Effect of Children on Log-hazard of Marital Dissolution

No. kids	Model A		Model B	
(ref=0)	Est	(SE)	Est	(SE)
1	-0.56	(0.10)	-0.33	(0.11)
2+	-0.01	(0.05)	0.27	(0.07)
corr(u_i^D, u_i^F)	0	–	-0.75	(0.20)

Effect of Children on Log-hazard of Marital Dissolution

No. kids	Model A		Model B	
(ref=0)	Est	(SE)	Est	(SE)
1	−0.56	(0.10)	−0.33	(0.11)
2+	−0.01	(0.05)	0.27	(0.07)
corr(u_i^D, u_i^F)	0	–	−0.75	(0.20)

Negative residual correlation implies women with low risk of dissolution tend to have high hazard of a conception

$\Rightarrow$ selection of low dissolution risk women into categories 1 and 2+
$\Rightarrow$ downward bias in estimated dissolution risk among women with children

Example with Multiple Processes, Multiple States and Competing Risks

Extend Lillard & Waite model to include cohabiting unions. In modelling partnership transitions we have to consider:

- Multiple states (unpartnered, married, cohabiting)
- Competing risks (cohabitation can be converted to marriage or be dissolved)

Partnership transition response is therefore binary for marriage, and multinomial for cohabitation.

Joint Modelling of Partnership Transitions and Fertility†

5 equations:

- Partnership process
  - 3 transitions: marriage $\longrightarrow$ single (dissolution), cohabitation $\longrightarrow$ single, cohabitation $\longrightarrow$ marriage
- Birth process
  - 2 equations distinguishing births in marriage and cohabitation

Equations include woman-specific random effects $\sim$ multivariate normal to allow correlation across transitions.

A discrete-time model can be fitted as a multilevel bivariate model for mixtures of binary and multinomial responses.

---

†Steele, Kallis, Goldstein and Joshi (2005) *Demography*, **42**: 647-673.

Effect of Family Disruption on Children's Educational Outcomes in Norway[†]

Previous research suggests that children whose parents divorce fare poorly on a range of adolescent and adult outcomes.

To what extent is association between parental divorce and children's education due to selection on unobserved characteristics of the mother?

[†] Steele, Sigle-Rushton and Kravdal (2008) *Submitted*.

Effect of Family Disruption on Children's Educational Outcomes in Norway†

Previous research suggests that children whose parents divorce fare poorly on a range of adolescent and adult outcomes.

To what extent is association between parental divorce and children's education due to selection on unobserved characteristics of the mother?

Outcome is educational attainment (5 categories from 'compulsory only' to 'postgraduate').

Explanatory variables: indicator of divorce and child's age at divorce.

† Steele, Sigle-Rushton and Kravdal (2008) *Submitted*.

Simultaneous Equations Model for Parental Divorce and Children's Education

Event history model for divorce

- Outcome is duration of marriage j for woman i
- Include woman-specific random effect u_i^D

Simultaneous Equations Model for Parental Divorce and Children's Education

Event history model for divorce

- Outcome is duration of marriage j for woman i
- Include woman-specific random effect u_i^D

Sequential probit model for educational transitions

- Convert 5-category educational outcome y_{ij} for child j of woman i into binary indicators of 4 sequential transitions (compulsory $\rightarrow$ lower secondary, ..., undergrad $\rightarrow$ postgrad)
- Include woman-specific random effect u_i^E

Simultaneous Equations Model for Parental Divorce and Children's Education

Event history model for divorce

- Outcome is duration of marriage j for woman i
- Include woman-specific random effect u_i^D

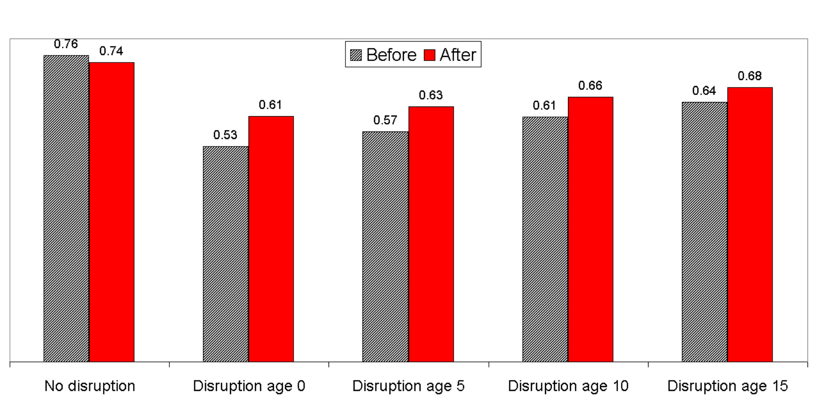
Sequential probit model for educational transitions

- Convert 5-category educational outcome y_{ij} for child j of woman i into binary indicators of 4 sequential transitions (compulsory $\rightarrow$ lower secondary, ..., undergrad $\rightarrow$ postgrad)
- Include woman-specific random effect u_i^E

$\left(u_i^D, u_i^E\right) \sim$ bivariate normal with correlation ρ

Probability of Continuing Beyond Lower Secondary (Before and After Allowing for Selection)

$\hat{\rho} = -\mathbf{0.431}$ **(SE=0.023)**



Final Remarks

- Multilevel modelling a flexible approach for analysing longitudinal data, and can now be implemented in several software packages

- BUT need to be especially careful in treatment of time-varying covariates – are values of x_t and y_t jointly determined?

- Multilevel multiprocess models can be useful for modelling selection effects (endogenous x_t)
  - Increasingly used in social sciences (e.g. demography)
  - Can be framed as multilevel multivariate response models

Software for Multilevel Longitudinal Data Analysis

- Growth curve models
  - Basic model in any mainstream statistics package (e.g. SAS, Stata, SPlus) and specialist multilevel modelling software (e.g. HLM, MLwiN)
  - Autocorrelated residuals in SAS and MLwiN
- Dynamic models
  - Allowing for initial conditions requires flexible environment (e.g. SAS, MLwiN)
- Event history analysis
  - Discrete-time models for one type of event, competing risks, or multiple states in any of the above
  - Discrete-time models for multiple processes in software that can handle bivariate discrete responses (e.g. SAS, MLwiN, aML)
  - aML most flexible for multiple processes (focus on continuous-time models)

Resources on Multilevel Longitudinal Data Analysis

Hedeker, D. and Gibbons, R.D. (2006) *Longitudinal Data Analysis*. John Wiley & Sons, New Jersey. [See also online resources at `http://tigger.cc.uic.edu/~hedeker/long.html`]

Singer J.D. and Willett J.B. (2003) *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, New York.

Steele F. (2008) Multilevel Models for Longitudinal Data. *Journal of the Royal Statistical Society, Series A*, **171**, 5-19.

Centre for Multilevel Modelling website (`http://www.cmm.bris.ac.uk`) includes materials from workshop on Multilevel Event History Analysis.