INTRODUCTION TO MULTILEVEL MODELLING FOR REPEATED MEASURES DATA

Belfast
9th June to 10th June, 2011

Dr James J Brown
Southampton Statistical Sciences Research Institute (UoS)
ADMIN Research Centre (IoE and NCRM)

jjb1@soton.ac.uk

Introduction

Aim: Introduce participants to analysing repeated measures data within the multilevel framework.

Learning Outcomes: By the end of this unit, you should understand the importance of correlation structures when modelling repeated measures and how complex structures can be incorporated within the multilevel framework.

SESSION ONE
Repeated Measures and the Random Intercepts Model

Content of Session

- Types of Longitudinal Data
- Reminder of the 2-Level Random Intercepts Model
- Basic Model for Repeated Measures
- *Issues with the Model*

Types of Longitudinal Data

- Simple transitions
  - Outcome at time t depends on the outcome at time t-1

- Extend to ‘time to event’
  - Classic *survival analysis* leading to the whole area of *event history analysis*

- Repeated measures over time
  - Child growth over time
  - Wages over time

We will concentrate on the last type with a continuous outcome.

- Can extend to categorical outcomes (Griffiths *et al*, 2004)

Basic 2-Level Random Intercepts Model

The classic example would be pupils within schools…

- *Pupil i's performance at 16, y_{ij}, will be a function of their performance at 11 and other background characteristics, $\underline{x}_{ij}$, as well as an unmeasured residual school effect, u_j*

$$y_{ij} = \alpha + \beta x_{ij} + u_j + e_{ij}$$

As u_j is a residual it has certain properties…

- Independence between residuals
- *Normal distribution with constant variance*
- Exogenous to the x's

Implied Correlation

Now each observation as has two independent random errors, u_j and e_{ij}

- This implies that two pupils i and k from the same school j will be correlated because they share the common random school effect

$$Cov(y_{ij}, y_{kj}) = Cov(u_j + e_{ij}, u_j + e_{kj}) = \sigma^2_{u0}$$

$$Corr(y_{ij}, y_{kj}) = \frac{Cov(y_{ij}, y_{kj})}{\sqrt{Var(y_{ij}) \times Var(y_{kj})}} = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_{e0}}$$

- This is the intra-cluster correlation (*proportion of the residual variance due to the school*)
  - Constant correlation across all pupils within the same school…

Basic Longitudinal Model

Once we see that a random effects model allows correlation between observations this leads us to a simple model for repeated measures…

- An individual i's wages at time t, y_{ti}, will be a function of time, time varying covariates, time-constant characteristics, and an unobserved individual effect…

$$y_{ti} = \alpha + \beta x_{ti} + \gamma z_i + u_i + e_{ti}$$

As u_i is a residual it has certain properties…

- Independence between residuals (*in this case independence between individuals*)
- *Normal distribution with constant variance*
- Exogenous to the x's and the z's

Implied Correlation

Now each observation as has two independent random errors, u_i and e_{ti}

- This implies that wages at two time points t and s from the same individual i will be correlated because they share the common random effect

$$Cov(y_{ti}, y_{si}) = Cov(u_i + e_{ti}, u_i + e_{si}) = \sigma^2_{u0}$$

$$Corr(y_{ti}, y_{si}) = \frac{Cov(y_{ti}, y_{si})}{\sqrt{Var(y_{ti}) \times Var(y_{si})}} = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_{e0}}$$

- Constant correlation between wages for individuals across time…

Example from US

Extract from the National Longitudinal Survey (NLS) containing young women aged 14 to 26 years of age in 1968.

- Wages collected once a year covering 1968 to 1973 (when respondent working at the time of the survey)
  - Missing time points are fine as long as we assume MAR…
- Information on location (living in the South – time-varying)
- Information on education (college graduate)
- Information on total work experience (adjusts for spells of unemployment)

This is actually simulated data available with STATA but it is based on the real data…

Simple Model

$\ln\_wage_{ij} \sim N(XB, \Omega)$

$\ln\_wage_{ij} = \beta_{0ij}\text{const} + 0.035(0.002)\text{time}_{ij} + -0.136(0.011)\text{south}_{ij} + 0.331(0.020)\text{collgrad}_{j} + 0.154(0.010)\text{mean\_exp\^1}_{j} + -0.013(0.001)\text{mean\_exp\^2}_{j}$

$\beta_{0ij} = 1.170(0.016) + u_{0j} + e_{0ij}$

$\left[u_{0j}\right] \sim N(0, \Omega_u) : \Omega_u = \left[0.085(0.003)\right]$

$\left[e_{0ij}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.059(0.001)\right]$

*-2*loglikelihood(IGLS Deviance)* = 5099.467(9818 of 9818 cases in use)

$$Corr(y_{ti}, y_{si}) = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_{e0}^2} = \frac{0.085}{0.085 + 0.059} = 0.590$$

- A strong residual correlation in wages across time for an individual after controlling to average experience (*increasing with diminishing returns*), education, and region
  - *Time effect or '**growth curve**' suggests level of wages is rising…*

Criticism of the Model

While a constant correlation structure might make sense for pupils within schools

- All pupils share the common school environment

It makes less sense when looking at data over time

- It seems plausible that my wages this year will have a stronger correlation with my wages last year than with my wages five years ago
  - *Unless I can really control out all the time-varying effects…*

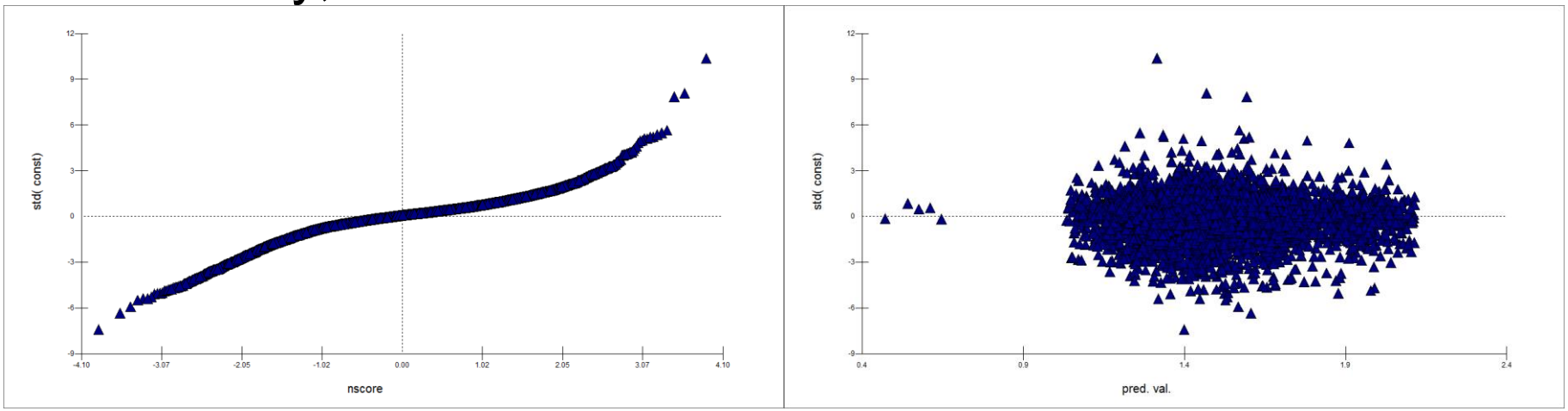
Endogeneity

- The u_j is correlated with time varying x's
  - *Important issue if we are assessing change over time…*

Residual Analysis – *how does it help?*

Of course we should check residuals for the usual things…

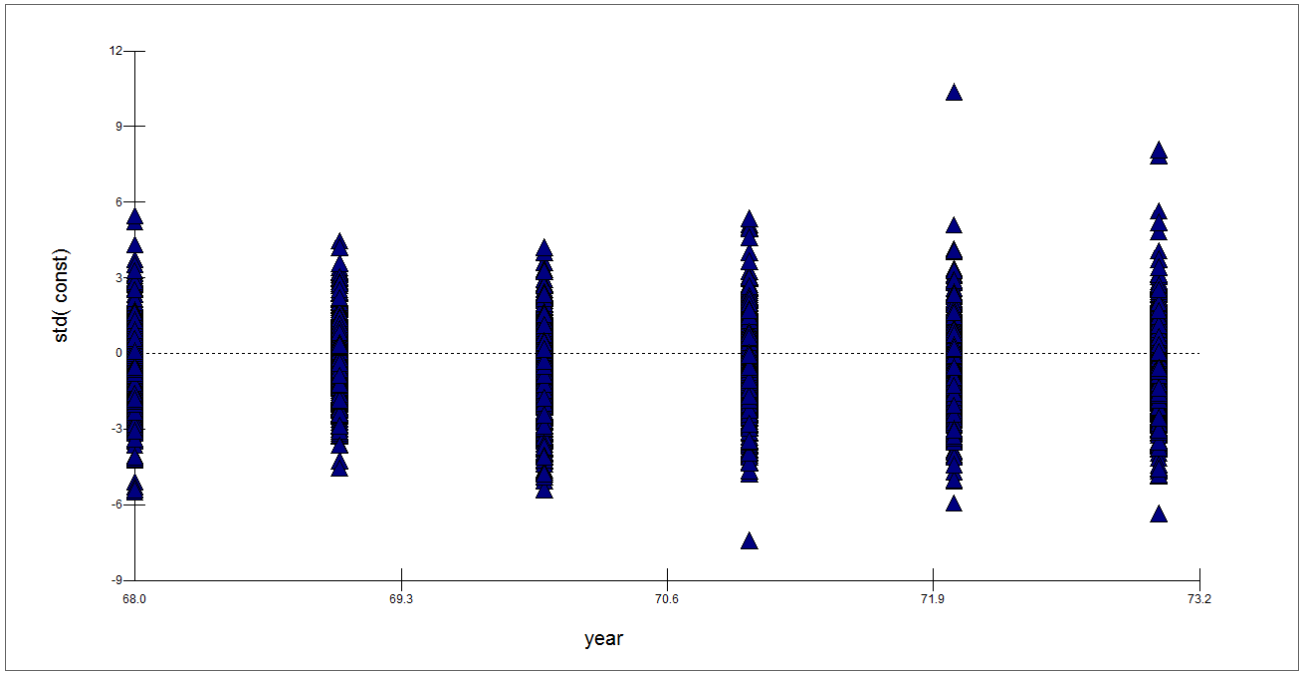
- Normality, constant variance at both levels…



- Mmmmm…

Can it help with spotting issues with the correlation and/or endogeneity?

Residual Analysis – *how does it help?*



- We can try plotting level one residuals against time to see if any patterns emerge…
  - *Plotting level two residuals against x's can help suggest endogeneity issues…*

Alternatives (to be introduced in the following sessions)

- Extending the correlation structure…
  - Random slopes
  - Multivariate models
  - *Additional levels…*

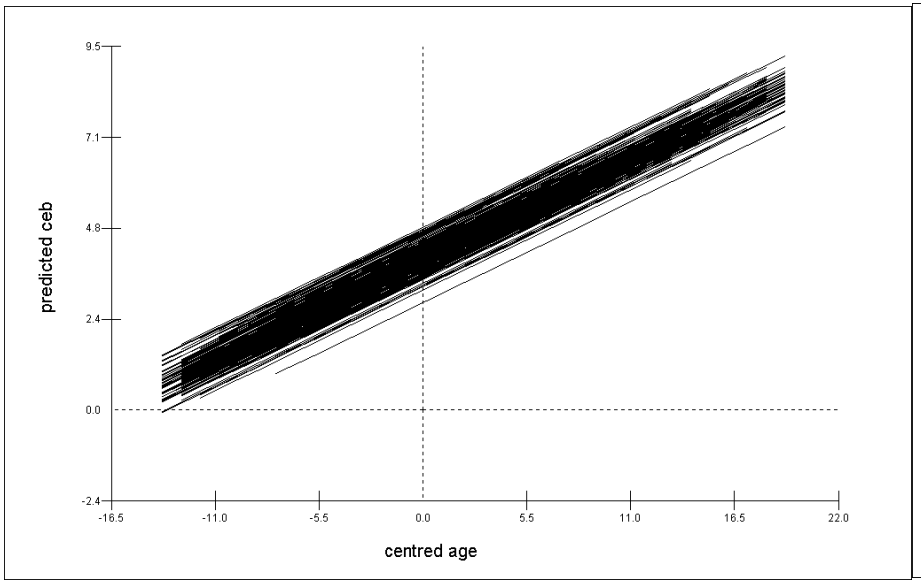
- Handling endogeneity
  - Role of contextual effects
  - Combined within and between models

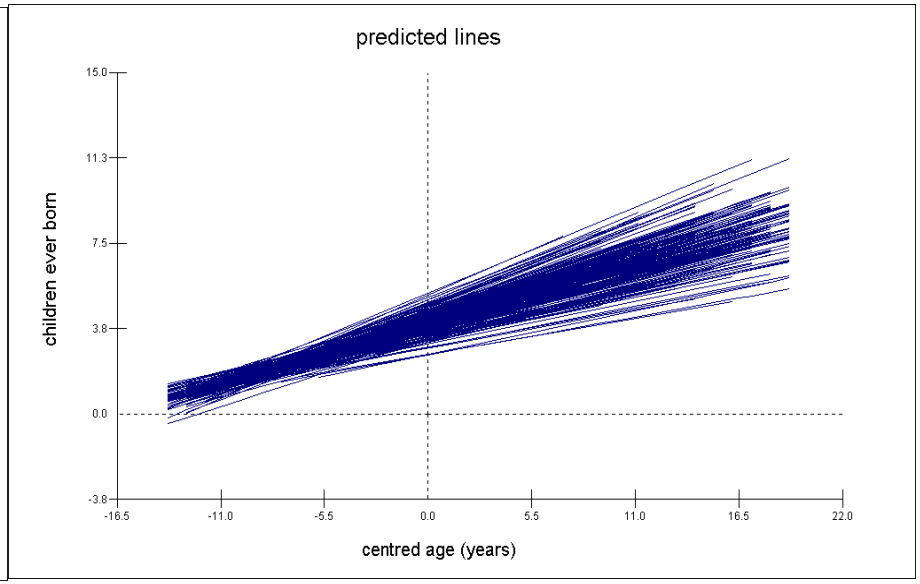
SESSION TWO
Complex Correlation Structures

Content of Session

- ‘Traditional’ motivation for random slopes
- Using random slopes to capture complex correlation structures
- Multivariate model approach (with time dummies)
- Examples of different structures…

Why random slopes?





It often is implausible that the residual grouping effect should be constant for all individuals within the group…

- The school effect may well be differential for pupils with differing prior attainments – the slopes vary from school to school
  - Interaction between x and group…

Random Slopes Model

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{0j} + u_{1j} x_{1ij} + e_{ij}$$

$$y_{ij} = \beta_{0ij} + \beta_{1j} x_{1ij}$$
$$\beta_{0ij} = \beta_0 + u_{0j} + e_{ij},\ \beta_{1j} = \beta_1 + u_{1j}$$
$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$$

We now have two residuals at level two, one to vary the intercept and one to vary the slope

- Independence across groups, independence between levels, independence between residuals and the x's
  - ***Covariance between the slope and intercept for the same group…***

Implied Correlation

Now each observation as has three random errors; u_{0j} and u_{1j}, which are correlated, and e_{ij}

- Consider the covariance between two pupils i and k from the same school j with prior attainment x_{1ij} and x_{1kj}

$$Cov(y_{ij}, y_{kj}) = Cov(u_{0j} + u_{1j}x_{1ij} + e_{ij}, u_{0j} + u_{1j}x_{1kj} + e_{kj})$$
$$= \sigma^2_{u0} + x_{1ij}\sigma_{u01} + x_{1kj}\sigma_{u01} + x_{1ij}x_{1kj}\sigma^2_{u1}$$
$$Var(y_{ij}) = \sigma^2_{u0} + 2x_{1ij}\sigma_{u01} + x^2_{1ij}\sigma^2_{u1} + \sigma^2_{e0}$$

$$Corr(y_{ij}, y_{kj}) = \frac{Cov(y_{ij}, y_{kj})}{\sqrt{Var(y_{ij}) \times Var(y_{kj})}}$$

- *No longer constant across pupils within the same school…*

Repeated Measures Context

$$y_{ti} = \beta_0 + \beta_1 x_{1ti} + u_{0i} + u_{1i} x_{1ti} + e_{ti}$$

We now have a model where x_{1ti} is capturing time for individual i

- *In this case linear 'growth' with time…*

$$Cov(y_{ti}, y_{si}) = \sigma^2_{u0} + x_{1ti}\sigma_{u01} + x_{1si}\sigma_{u01} + x_{1ti}x_{1si}\sigma^2_{u1}$$

- We can now get differing correlations between two time points s and t for individual i
  - **It is structured through the random slope (but hard to see and/or interpret that structure)**
  - **But has the intuitive appeal that each individual has their own underlying growth curve…**

Repeated Measures Context

We can now extend our simple random effects model (p.11) to allow for a more complex correlation structure for wages across time…

$$\text{ln\_wage}_{ij} \sim N(XB,\ \Omega)$$

$$\text{ln\_wage}_{ij} = \beta_{0ij}\text{const} + \beta_{1j}\text{time}_{ij} + -0.132(0.011)\text{south}_{ij} + 0.337(0.020)\text{collgrad}_{j} + 0.152(0.009)\text{mean\_exp\^{}1}_{j} + -0.013(0.001)\text{mean\_exp\^{}2}_{j}$$

$$\beta_{0ij} = 1.177(0.016) + u_{0j} + e_{0ij}$$

$$\beta_{1j} = 0.032(0.002) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.098(0.004) & \\ -0.009(0.001) & 0.004(0.000) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 0.047(0.001) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 4781.754(9818 of 9818 cases in use)

- -2LL has dropped by over 200 from adding two parameters
  - **HIGHLY SIGNIFICANT** (*should adjust p-values for a one-sided test on a variance*)

Implied Correlation

We can now plug-in the values of the variance and covariance parameters to get the implied variance-covariance matrix and then the correlation structure

- Remember – time is coded 0, 1,…,5 representing measurements in the years 1968 to 1973
  - When working with random slopes you always want zero to be a meaningful value (why we often centre)

Covariance

TIME	0	1	2	3	4	5
0	0.145					
1	0.089	0.131				
2	0.08	0.079	0.125			
3	0.071	0.074	0.077	0.127		
4	0.062	0.069	0.076	0.083	0.137	
5	0.053	0.064	0.075	0.086	0.097	0.155

Implied Correlation

Correlation

TIME	1968	1969	1970	1971	1972	1973
1968	1.000					
1969	0.646	1.000				
1970	0.594	0.617	1.000			
1971	0.523	0.574	0.611	1.000		
1972	0.440	0.515	0.581	0.629	1.000	
1973	0.354	0.449	0.539	0.613	0.666	1.000

- As we might expect the residual correlation between wages decreases as the time-points become more separated…
  - *Pattern down the diagonals is forced by the random slope structure – is it the right thing?*

Multivariate Model

Ideally, we want an approach that offers full flexibility for the residual correlation structure

- with the ability to impose specific structures
  - We will see later that MLwiN does not quite allow this as we cannot fit AR type structures (can in STATA and SAS)

Let's assume we have t=1,..,T observations over time for each individual i

- Can have intermittent non-response as long as its MAR…

We then treat the measurements for each individual as coming from a multivariate normal distribution.

- The specification of the mean function can vary with time (full multivariate) if needed…

Multivariate Model

$$\begin{bmatrix} y_{1i} \\ \dots \\ y_{Ti} \end{bmatrix} \sim N\left(\beta \boldsymbol{X}, \begin{bmatrix} \boldsymbol{\sigma_1^2} & & \\ \dots & \dots & \\ \boldsymbol{\sigma_{1T}} & \dots & \boldsymbol{\sigma_T^2} \end{bmatrix} \right)$$

- Here we assume the β's do not change across time and we have the same x's at each time-point…
  - Makes sense with repeated measures data (wages across time)
    - ***Model if we use a 'trick' in MLwiN to fit it***
  - If modelling different y's (say heights and weights) for individuals it then makes sense to vary the mean function for each outcome
    - ***Default if we tell MLwiN its multivariate***

Repeated Measures Context

We can now extend our model to have this fully flexible covariance (correlation) structure…

$$\ln\_wage_{ij} \sim N(XB,\ \Omega)$$

$$\ln\_wage_{ij} = 1.185(0.016)\text{const} + 0.029(0.002)\text{time}_{ij} + -0.129(0.011)\text{south}_{ij} + 0.338(0.020)\text{collgrad}_{j} + 0.151(0.009)\text{mean\_exp\^1}_{j} + -0.013(0.001)\text{mean\_exp\^2}_{j} + u_{6j}\text{year68}_{ij} + u_{7j}\text{year69}_{ij} + u_{8j}\text{year70}_{ij} + u_{9j}\text{year71}_{ij} + u_{10j}\text{year72}_{ij} + u_{11j}\text{year73}_{ij}$$

$$\begin{bmatrix} u_{6j} \\ u_{7j} \\ u_{8j} \\ u_{9j} \\ u_{10j} \\ u_{11j} \end{bmatrix} \sim N(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.127(0.005) & & & & & \\ 0.077(0.004) & 0.128(0.005) & & & & \\ 0.074(0.004) & 0.100(0.004) & 0.140(0.005) & & & \\ 0.059(0.004) & 0.085(0.004) & 0.093(0.004) & 0.142(0.004) & & \\ 0.057(0.004) & 0.074(0.004) & 0.077(0.004) & 0.095(0.004) & 0.159(0.005) & \\ 0.050(0.004) & 0.063(0.004) & 0.068(0.004) & 0.080(0.004) & 0.121(0.004) & 0.156(0.005) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 4482.365(9818 of 9818 cases in use)

- -2LL has dropped by nearly 300 from adding 17 parameters
  - **HIGHLY SIGNIFICANT**

Implied Correlation

This model now has a fully flexible structure for the residual covariance matrix

- Unstructured correlation

Correlation

TIME	1968	1969	1970	1971	1972	1973
1968	1.000					
1969	0.604	1.000				
1970	0.555	0.747	1.000			
1971	0.439	0.630	0.660	1.000		
1972	0.401	0.519	0.516	0.632	1.000	
1973	0.355	0.446	0.460	0.538	0.768	1.000

NOTE: *The 'trick' in MLwiN is to have a dummy for each time point BUT only have variance at level two (the individual)*

Other Structures – compound symmetry & heterogeneous compound symmetry

Compound Symmetry

- This is the standard random effects structure
  - Constant residual variance across time $(\sigma_u^2 + \sigma_e^2)$ and a constant covariance between time points (σ_u^2)

From p.11 we see this model fits with -2LL = 5099.467

Heterogeneous Compound Symmetry

- Allows the residual variance to change with time BUT assumes the underlying correlation between time points is constant
  - Can't quite do this in MLwiN – we can make the covariance constant…

Heterogeneous Compound Symmetry (well almost)

$$\ln\_wage_{ij} \sim N(XB,\ \Omega)$$

$$\ln\_wage_{ij} = 1.223(0.013)\text{const} + \beta_{1j}\text{time}_{ij} + -0.155(0.009)\text{south}_{ij} + 0.362(0.016)\text{collgrad}_{j} + 0.157(0.007)\text{mean\_exp\^{}1}_{j} + -0.013(0.001)\text{mean\_exp\^{}2}_{j} + e_{6ij}\text{year68}_{ij} + e_{7ij}\text{year69}_{ij} + e_{8ij}\text{year70}_{ij} + e_{9ij}\text{year71}_{ij} + e_{10ij}\text{year72}_{ij} + e_{11ij}\text{year73}_{ij}$$

$$\beta_{1j} = 0.017(0.002) + u_{1j}$$

$$\begin{bmatrix} u_{1j} \end{bmatrix} \sim N(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.006(0.000) \end{bmatrix}$$

$$\begin{bmatrix} e_{6ij} \\ e_{7ij} \\ e_{8ij} \\ e_{9ij} \\ e_{10ij} \\ e_{11ij} \end{bmatrix} \sim N(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 0.119(0.005) & & & & & \\ 0.000(0.000) & 0.092(0.004) & & & & \\ 0.000(0.000) & 0.000(0.000) & 0.092(0.003) & & & \\ 0.000(0.000) & 0.000(0.000) & 0.000(0.000) & 0.075(0.003) & & \\ 0.000(0.000) & 0.000(0.000) & 0.000(0.000) & 0.000(0.000) & 0.045(0.002) & \\ 0.000(0.000) & 0.000(0.000) & 0.000(0.000) & 0.000(0.000) & 0.000(0.000) & 0.030(0.003) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 6056.254(9818 of 9818 cases in use)

Heterogeneous Compound Symmetry (well almost)

- Time dummies varying at level one give a flexible residual variance
- Random effect at level two gives a constant covariance
  - *So the underlying correlation is actually changing…*

From p.30 we see this model fits with -2LL = 6056.254

- Poor compared to the simple random effects model…

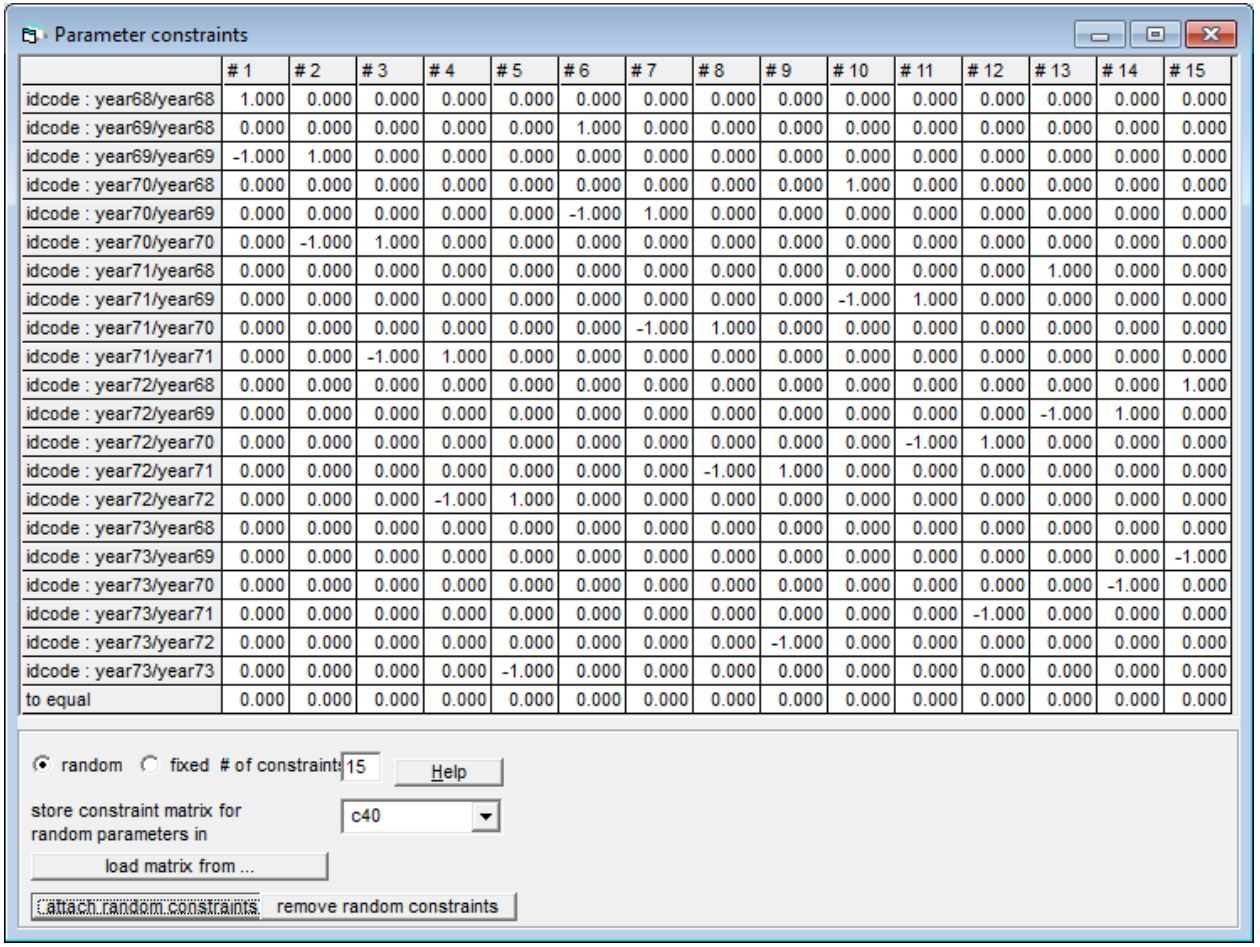
Other Structures – Toeplitz

This makes elements of the covariance matrix constant down the diagonals

$$\begin{bmatrix} \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_2 & \sigma_1 & \sigma^2 \end{bmatrix}$$

- This is fairly easy to fit in MLwiN as it can be specified using linear constraints on our fully unstructured model (p.27)
  - *Will fit if you have an AR(1) structure BUT does not assume the inter-relationships that would exist down the columns…*

Other Structures – Toeplitz



	# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8	# 9	# 10	# 11	# 12	# 13	# 14	# 15
idcode : year68/year68	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
idcode : year69/year68	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
idcode : year69/year69	-1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
idcode : year70/year68	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
idcode : year70/year69	0.000	0.000	0.000	0.000	0.000	-1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
idcode : year70/year70	0.000	-1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
idcode : year71/year68	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
idcode : year71/year69	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-1.000	1.000	0.000	0.000	0.000	0.000
idcode : year71/year70	0.000	0.000	0.000	0.000	0.000	0.000	-1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
idcode : year71/year71	0.000	0.000	-1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
idcode : year72/year68	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
idcode : year72/year69	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-1.000	1.000	0.000
idcode : year72/year70	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-1.000	1.000	0.000	0.000	0.000
idcode : year72/year71	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
idcode : year72/year72	0.000	0.000	0.000	-1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
idcode : year73/year68	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
idcode : year73/year69	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-1.000
idcode : year73/year70	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-1.000	0.000
idcode : year73/year71	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-1.000	0.000	0.000	0.000
idcode : year73/year72	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-1.000	0.000	0.000	0.000	0.000	0.000	0.000
idcode : year73/year73	0.000	0.000	0.000	0.000	-1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
to equal	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Other Structures – Toeplitz

$$\ln\_wage_{ij} \sim N(XB,\ \Omega)$$

$$\ln\_wage_{ij} = 1.172(0.016)\text{const} + 0.031(0.002)\text{time}_{ij} + -0.128(0.011)\text{south}_{ij} + 0.332(0.020)\text{collgrad}_{j} + 0.155(0.009)\text{mean\_exp\^{}1}_{j} + -0.014(0.001)\text{mean\_exp\^{}2}_{j} + u_{6j}\text{year68}_{ij} + u_{7j}\text{year69}_{ij} + u_{8j}\text{year70}_{ij} + u_{9j}\text{year71}_{ij} + u_{10j}\text{year72}_{ij} + u_{11j}\text{year73}_{ij}$$

$$\begin{bmatrix} u_{6j} \\ u_{7j} \\ u_{8j} \\ u_{9j} \\ u_{10j} \\ u_{11j} \end{bmatrix} \sim N(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.144(0.003) & & & & & \\ 0.100(0.003) & 0.144(0.003) & & & & \\ 0.081(0.003) & 0.100(0.003) & 0.144(0.003) & & & \\ 0.068(0.003) & 0.081(0.003) & 0.100(0.003) & 0.144(0.003) & & \\ 0.060(0.003) & 0.068(0.003) & 0.081(0.003) & 0.100(0.003) & 0.144(0.003) & \\ 0.053(0.004) & 0.060(0.003) & 0.068(0.003) & 0.081(0.003) & 0.100(0.003) & 0.144(0.003) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 4635.171(9818 of 9818 cases in use)

We see this model fits with -2LL = 4635.171

- Significant improvement on RE (5099.467) and RS (4781.754)
- Significantly worse than unstructured (4482.365)

Just to Finish

- Using random slopes with time (as a continuous variable) allows for more complex residual correlations
  - **Can be difficult to interpret in terms of a correlation structure**
  - **Intuitive interpretation with a unique 'growth curve' for each individual**

- Using random structure on time dummies allows us to be completely unstructured
  - Can impose linear constraints on the parameters of the covariance matrix
  - Can use time dummies in the **X** as well to get a totally flexible shape for the ***growth curve*** over time

SESSION THREE
Improving the fixed part of the model

Content of Session

- The ‘within’ and ‘between’ models
- A strong assumption behind random effects
- Using contextual effects to get at the within effect
- Other extensions
  - Extra levels
  - Cross-level interactions

Within and Between Effects

Let's go back to a simple random effects model with a single continuous x

$$y_{ij} = \alpha + \beta x_{ij} + u_j + e_{ij}$$

We can also define a model for the **between (school) effects** as

$$\bar{y}_j = \alpha + \beta \bar{x}_j + v_j$$

where under expectation u_j and v_j are the same random error as the e_{ij}'s have expected mean zero in the group

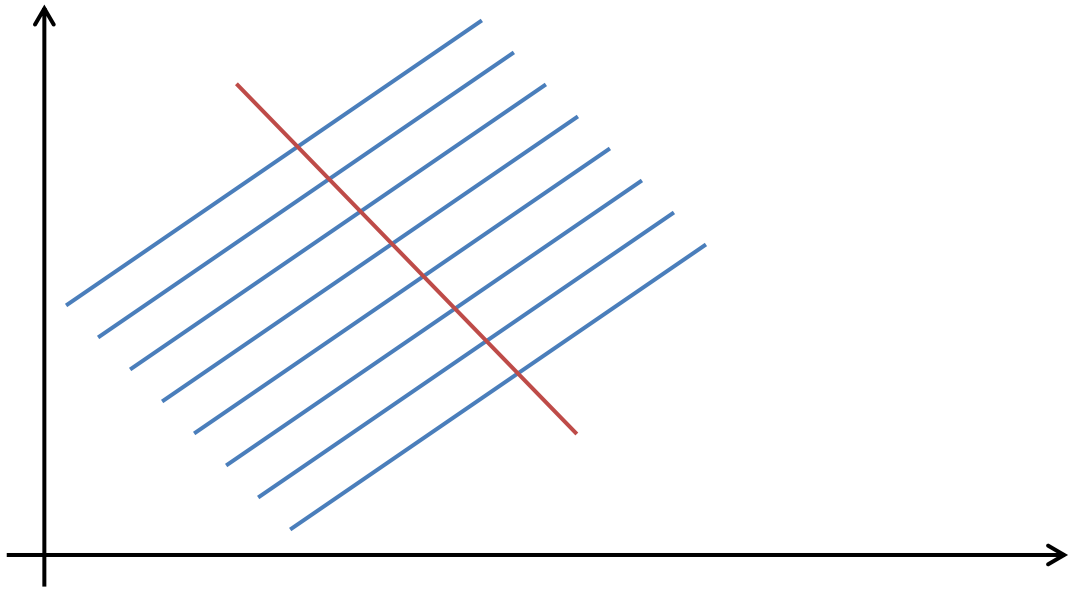
The difference gives the model for the **within (school) effects**

$$(y_{ij} - \bar{y}_j) = \beta (x_{ij} - \bar{x}_j) + \tilde{e}_{ij}$$

Strong Assumption of Random Effects

The β for the overall effect is the same as the within effect and the between effect!

- ***But the within school relationship can be positive while between schools its negative!***



Strong Assumption of Random Effects

This is what economists often worry about!

- If I am trying to evaluate impact (a change within each individual) then I want a good estimate of the within effect…
  - *I do not want it to be biased by mixing-up between and within effects with the random effect*

Hausmann specification test is used by STATA to test whether the β's can be assumed the same

- We can adjust our model to bring in the group mean as a contextual effect (and centre the original x)
  - Rabe-Hesketh, S. and Skrondal, A. (2005)

$$y_{ij} = \alpha + \beta_w\left(x_{ij} - \bar{x}_j\right) + \beta_b\bar{x}_j + u_j + e_{ij}$$

Simple Example (with the NLS data)

We want to evaluate the impact of moving to the South of the US on wages

- South in the data is a time-varying covariate…

$$\ln\_wage_{ij} \sim N(XB,\ \Omega)$$

$$\ln\_wage_{ij} = \beta_{0ij}\text{const} + 0.040(0.002)\text{time}_{ij} + -0.155(0.012)\text{south}_{ij}$$

$$\beta_{0ij} = 1.438(0.009) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0,\ \Omega_u) : \Omega_u = \begin{bmatrix} 0.112(0.003) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0,\ \Omega_e) : \Omega_e = \begin{bmatrix} 0.058(0.001) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 5747.158(9818 of 9818 cases in use)

- A move to the south reduces loge(wages) by 0.155; a 15% drop in wages

Simple Example (with the NLS data)

Is it that simple?

- Surely we would expect time in the South to impact
  - Different 'growth curve' on wages in the South…

$$\ln\_wage_{ij} \sim N(XB,\ \Omega)$$

$$\ln\_wage_{ij} = \beta_{0ij}\text{const} + 0.040(0.002)\text{time}_{ij} + -0.153(0.015)\text{south}_{ij} + -0.001(0.004)\text{time.south}_{ij}$$

$$\beta_{0ij} = 1.438(0.010) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0,\ \Omega_u) : \Omega_u = \begin{bmatrix} 0.112(0.003) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0,\ \Omega_e) : \Omega_e = \begin{bmatrix} 0.058(0.001) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 5747.132(9818 of 9818 cases in use)

- No evidence to support a significant interaction…
  - Just a level shift **IMPACT**…

Simple Example (with the NLS data)

What about our strong assumption behind the random effects model?

- Are we really getting the true within effect here of moving to the South?

$$\ln\_wage_{ij} \sim N(XB,\ \Omega)$$

$$\ln\_wage_{ij} = \beta_{0ij}\text{const} + 0.040(0.002)\text{time}_{ij} + -0.058(0.021)(\text{south-m(idcode)})_{ij} + -0.196(0.014)\text{prop\_south}_{j}$$

$$\beta_{0ij} = 1.454(0.010) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.112(0.003) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 0.058(0.001) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 5716.835(9818 of 9818 cases in use)

Simple Example (with the NLS data)

- This parameterisation is a significantly better fit and both β's strongly significant
  - Clear evidence that the between and within effects are NOT the same…
- Suggests the direct impact of a move is about a 6% drop in wages
- Also suggests that the proportion of time spent in the South impacts significantly on the overall level of an individual's wages

Simple Example (with the NLS data)

- The effects are still there when we bring in our additional controls…

$$\ln\_wage_{ij} \sim N(XB,\ \Omega)$$

$$\ln\_wage_{ij} = \beta_{0ij}\text{const} + 0.035(0.002)\text{time}_{ij} + -0.057(0.021)(\text{south-m(idcode)})_{ij} + -0.163(0.012)\text{prop\_south}_{j} + 0.329(0.020)\text{collgrad}_{j} + 0.152(0.010)\text{mean\_exp\^{}1}_{j} + -0.013(0.001)\text{mean\_exp\^{}2}_{j}$$

$$\beta_{0ij} = 1.183(0.016) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.085(0.003) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 0.059(0.001) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 5080.906(9818 of 9818 cases in use)

- ***Still fits significantly better than the standard random effects model (p.11)***

Simple Example (with the NLS data)

- What about if we extend our correlation structure?

$$\ln\_wage_{ij} \sim N(XB,\ \Omega)$$

$$\ln\_wage_{ij} = 1.202(0.016)\text{const} + 0.029(0.002)\text{time}_{ij} + -0.021(0.021)(\text{south-m(idcode)})_{ij} + -0.166(0.012)\text{prop\_south}_{j} + 0.149(0.009)\text{mean\_exp\^{}1}_{j} + -0.013(0.001)\text{mean\_exp\^{}2}_{j} + 0.336(0.020)\text{collgrad}_{j} + u_{6j}\text{year68}_{ij} + u_{7j}\text{year69}_{ij} + u_{8j}\text{year70}_{ij} + u_{9j}\text{year71}_{ij} + u_{10j}\text{year72}_{ij} + u_{11j}\text{year73}_{ij}$$

$$\begin{bmatrix} u_{6j} \\ u_{7j} \\ u_{8j} \\ u_{9j} \\ u_{10j} \\ u_{11j} \end{bmatrix} \sim N(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 0.126(0.005) & & & & & \\ 0.076(0.004) & 0.127(0.005) & & & & \\ 0.074(0.004) & 0.100(0.004) & 0.140(0.005) & & & \\ 0.058(0.004) & 0.084(0.004) & 0.093(0.004) & 0.142(0.004) & & \\ 0.057(0.004) & 0.073(0.004) & 0.077(0.004) & 0.095(0.004) & 0.159(0.005) & \\ 0.049(0.004) & 0.063(0.004) & 0.067(0.004) & 0.080(0.004) & 0.122(0.004) & 0.156(0.005) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 4446.532(9818 of 9818 cases in use)

- Proportion of time in the South impacts on wages but not the movement (***needs a bit more thought***)…

Further Extensions – extra levels

By fitting our repeated measures models within the multilevel framework we can account for extra clustering

- We can relax the assumption that individuals are independent…

$$y_{tij} = \alpha + \beta x_{tij} + \gamma z_{ij} + v_j + u_{ij} + e_{tij}$$

- Work by a recent PhD student at Southampton has been exploring these models in the context of the Brazilian Labour Force Survey
  - ***Rotation panel survey with area clustering…***

Further Extensions – cross level interactions

By fitting our repeated measures models within the multilevel framework we can allow the impact of time to change for different individual level (time-constant) characteristics

- Different *growth curve* for college grads? **NO**

$$\text{ln\_wage}_{ij} \sim \text{N}(XB,\ \Omega)$$

$$\text{ln\_wage}_{ij} = \beta_{0ij}\text{const} + 0.035(0.002)\text{time}_{ij} + \text{-}0.136(0.011)\text{south}_{ij} + 0.348(0.029)\text{collgrad}_{j} + 0.154(0.010)\text{mean\_exp\^{}1}_{j} + \text{-}0.013(0.001)\text{mean\_exp\^{}2}_{j} + \text{-}0.005(0.006)\text{time.collgrad}_{ij}$$

$$\beta_{0ij} = 1.169(0.016) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.085(0.003) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 0.059(0.001) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 5098.840(9818 of 9818 cases in use)

Further Extensions – cross level interactions

- Different *South Effect* for college grads? **NO**

$$\ln\_wage_{ij} \sim N(XB,\ \Omega)$$

$$\ln\_wage_{ij} = \beta_{0ij}\text{const} + 0.035(0.002)\text{time}_{ij} + -0.138(0.011)\text{south}_{ij} + 0.323(0.023)\text{collgrad}_{j} + 0.154(0.010)\text{mean\_exp\^1}_{j} + -0.013(0.001)\text{mean\_exp\^2}_{j} + 0.027(0.038)\text{south.collgrad}_{ij}$$

$$\beta_{0ij} = 1.171(0.016) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.085(0.003) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 0.059(0.001) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 5098.958(9818 of 9818 cases in use)

- *You might want to explore South with Race…*

Just to Finish

- If you are interested in the direct impact of a change then we need to be careful…
  - Between and within effects can get messed-up with simple random effects

- Might want to set the data so that there is a variable ***Move_South*** that just takes the value 1 the year the move south takes place and is zero otherwise
  - ***Move_North*** would pick-up a move from the South
  - ***South*** picks-up living in the South
- Now the β for ***South*** will tell us if wages are generally lower in the South
  - The β's on the ***Move_South*** and ***Move_North*** will try and measure the immediate impact of the change on wages…

SESSION FOUR
Some alternative approaches

Content of Session

- Fixed effects
- GEE
- Other frameworks…

Basic Repeated Measures Model

This is the basic model structure we have been using for the past two days…

$$y_{ti} = \alpha + \beta x_{ti} + \gamma z_i + u_i + e_{ti}$$

- u_i represents some residual unobserved effect that individual i has on the outcome y
  - It also generates a correlation structure for the observations
- In the random effects situation we assume the u_i's are drawn from some distribution for the population
  - So a sample can estimate the population parameter σ^2_{u0} capturing the residual correlation that exists within individuals ***in the population***

Basic Repeated Measures Model – fixed effects approach

Suppose I'm only interested in the '***within effects***'

- **Time-constant effects** (including u_i) are just a nuisance to me…
  - Treat them as fixed and then difference them out

$$(y_{ij} - \bar{y}_j) = \beta(x_{ij} - \bar{x}_j) + \tilde{e}_{ij}$$

  - Just like adding a dummy for each individual BUT you don't need all the parameters
    - Computer drops one observation per individual to account for estimating the means

Classic approach in econometrics (Wooldridge, 2010) to handle panel data

- *Very easy to fit in STATA…*

Basic Repeated Measures Model – general estimating equations (GEE) approach

We still have the same model

$$y_{ti} = \alpha + \beta x_{ti} + \gamma z_i + u_i + e_{ti}$$

BUT

- We are worried that the correlation structure is more complex than say our simple random effects…
  - The correlation structure is a nuisance getting in the way of us estimating…
- Need an approach to get good estimates of the fixed parameters in the model
  - While recognising the existence of complex correlations…

General Estimating Equations (xtgee in STATA)

For full details see Liang and Zeger (1986) although there is a basic description in Diggle, Liang, and Zeger (1994)

Basic Idea

The correlation structure is a *nuisance parameter*. However, if we specify a *working correlation structure* that is sensible we can get efficient estimates of the β's (with robust standard errors) and therefore estimate the 'marginal' model for the population.

- We iterate between a moments estimator for the correlation structure and an efficient estimator for the β's
- We can then use a marginal model to contrast sub-populations in the data BUT not necessarily to talk about individuals (*unless the model follows a Gaussian distribution*).

General Estimating Equations

This leads to a very flexible estimation strategy that can be applied to linear and non-linear data with a whole class of within group/individual correlation structures.

- In fact we do not even need to get the correlation structure correct as long as we use robust methods to estimate the standard errors of the β's.
  - However, getting a 'good' approximation will lead to more efficient estimation of the β's.

Alternative Framework – Structural Equation Modelling

Not my area at all BUT it is an alternative way of handling the u_i term

- This is an unobserved effect, which we must put some structure on…

GLaMM (Generalised Latent Mixed Models)

This is work by Rabe-Hesketh, Skrondal (and others) that gives a framework in which you can fit either a multilevel approach or a structural equation approach

- Just depends on the assumptions you impose on the latent structure
  - Steele (2008) gives a good review of this...

Extra References

Berrington, A., Smith, P. W. F. and Sturgis, P. (2006) An overview of methods for the analysis of panel data. *ESRC National Centre for Research Methods Briefing Paper*.

Griffiths P., Brown J. and Smith P. W. F. (2004) A comparison of univariate and multivariate multilevel models for repeated measures using uptake of antenatal care in Uttar Pradesh. *JRSS(A)*, **167**, pp. 597-611.

Steele, F. (2008) Multilevel Models for Longitudinal Data. *JRSS(A)*, **171**, pp. 5-19.

Singer, J. and Willett J. (2003) *Applied Longitudinal Data Analysis: Modelling change and event occurrence*. New York: Oxford University Press.