Introduction to Multilevel Modelling
and the software MLwiN

Lecture at China National Institute for Educational Research, China
7 March 2008

Wen-Jung Peng
Graduate School of Education
University of Bristol

Review some concepts

The data set used to address the related issues in this lecture – MLwiN tutorial sample

Basic information of the data set

- Number of schools: 65
- Number of pupils: 4059

Normexam: pupil's exam score at age 16
Standlrt: pupil's score at age 11 on the London Reading Test

We are interested in

Q1: The relationship between ‘pupil’s exam score at age 16’ and ‘pupil’s score at age 11 on the London Reading Test’ - the effect of ‘standlrt’ (prior attainment) on ‘normexam’ (outcome)
Simple scatter plot

All 4059 pupils

Ordinary linear regression method

Unit of analysis – pupil
One regression line
Simple scatter plot
Aggregated data for 65 schools

Data source: MLwiN tutorial sample

Ordinary linear regression method
Unit of analysis – school
One regression line

Data source: MLwiN tutorial sample
What went missing in the analysis?

The hierarchical structure of the data set

- The fact that those 4059 pupils were from 65 schools was ignored

Ordinary linear regression method

Unit of analysis – pupil
One regression line
Ordinary linear regression method

Unit of analysis – school
One regression line

- The information about individual pupils was discarded

How this structure affects the measurement of interest?
How this structure affects the measurement of interest?

Data source: MLwiN tutorial sample

Simple scatter plots

For School 3 and School 64
Data source: MLwiN tutorial sample
Underlying Meaning

In this case, pupils within a school will be more alike, on average, than pupils from different schools.
We are also interested in

Q1: The relationship between ‘pupil’s exam score at age 16’ and ‘pupil’s score at age 11 on the London Reading Test’ - the effect of ‘standlrt’ (prior attainment) on ‘normexam’ (outcome)

Q2: How different the relationship is across schools? - the variability of the effect of ‘standlrt’ (prior attainment) on ‘normexam’ (outcome) across schools

Data source: MLwiN tutorial sample

The variation between schools
Can ordinary linear regression method estimate the variation between schools?

It is possible that “The variation between schools could be modelled by incorporating separate terms for each school…”

(Rasbash, et al., 2005)

For example, to fit 64 school dummy variables in a model using school 1 as the reference school.

\[
\begin{align*}
\text{model} & = \alpha + \beta_{1,\text{stand1}} + \beta_{2,\text{school}_2} + \beta_{3,\text{school}_3} + \beta_{4,\text{school}_4} + \beta_{5,\text{school}_5} + \\
& \quad + \beta_{6,\text{school}_6} + \beta_{7,\text{school}_7} + \beta_{8,\text{school}_8} + \beta_{9,\text{school}_9} + \beta_{10,\text{school}_{10}} + \\
& \quad + \beta_{11,\text{school}_{11}} + \beta_{12,\text{school}_{12}} + \beta_{13,\text{school}_{13}} + \beta_{14,\text{school}_{14}} + \beta_{15,\text{school}_{15}} + \\
& \quad + \beta_{16,\text{school}_{16}} + \beta_{17,\text{school}_{17}} + \beta_{18,\text{school}_{18}} + \beta_{19,\text{school}_{19}} + \beta_{20,\text{school}_{20}} + \\
& \quad + \beta_{21,\text{school}_{21}} + \beta_{22,\text{school}_{22}} + \beta_{23,\text{school}_{23}} + \beta_{24,\text{school}_{24}} + \beta_{25,\text{school}_{25}} + \\
& \quad + \beta_{26,\text{school}_{26}} + \beta_{27,\text{school}_{27}} + \beta_{28,\text{school}_{28}} + \beta_{29,\text{school}_{29}} + \beta_{30,\text{school}_{30}} + \\
& \quad + \beta_{31,\text{school}_{31}} + \beta_{32,\text{school}_{32}} + \beta_{33,\text{school}_{33}} + \beta_{34,\text{school}_{34}} + \beta_{35,\text{school}_{35}} + \\
& \quad + \beta_{36,\text{school}_{36}} + \beta_{37,\text{school}_{37}} + \beta_{38,\text{school}_{38}} + \beta_{39,\text{school}_{39}} + \beta_{40,\text{school}_{40}} + \\
& \quad + \beta_{41,\text{school}_{41}} + \beta_{42,\text{school}_{42}} + \beta_{43,\text{school}_{43}} + \beta_{44,\text{school}_{44}} + \beta_{45,\text{school}_{45}} + \\
& \quad + \beta_{46,\text{school}_{46}} + \beta_{47,\text{school}_{47}} + \beta_{48,\text{school}_{48}} + \beta_{49,\text{school}_{49}} + \beta_{50,\text{school}_{50}} + \\
& \quad + \beta_{51,\text{school}_{51}} + \beta_{52,\text{school}_{52}} + \beta_{53,\text{school}_{53}} + \beta_{54,\text{school}_{54}} + \beta_{55,\text{school}_{55}} + \\
& \quad + \beta_{56,\text{school}_{56}} + \beta_{57,\text{school}_{57}} + \beta_{58,\text{school}_{58}} + \beta_{59,\text{school}_{59}} + \beta_{60,\text{school}_{60}} + \\
& \quad + \beta_{61,\text{school}_{61}} + \beta_{62,\text{school}_{62}} + \beta_{63,\text{school}_{63}} + \beta_{64,\text{school}_{64}} + \beta_{65,\text{school}_{65}}.
\end{align*}
\]
Can ordinary linear regression method estimate the variation between schools?

However, it is inefficient and inadequate “because it involves estimating many times coefficients...because it does not treat schools as a random sample...”

(Rasbash, et al., 2005)

Think about a national data set with hundreds of schools......

Multilevel modelling

A statistical technique that allows an analysis to take account of the levels of hierarchical structure in the population so that we can

- treat sample as random
- specify and fit a wide range of multilevel models
- understand where and how effects are occurring

(Rasbash, et al., 2005)
Statistical software packages

There are some statistical packages have the function

- **MLwiN** is one of them

Menu bar
Get started with creating a \textit{MLwiN worksheet}

\textit{MLwiN} can only input and output \textit{numerical} data
- code data numerically
- assign an identical numerical code to all missing data
- three ways of creating a \textit{MLwiN} worksheet:
  - input data into a \textit{MLwiN} worksheet
  - copy and paste data into a \textit{MLwiN} worksheet
  - import ASCII data from a text file

\textbf{Input data into a \textit{MLwiN} worksheet}

Input data directly into the \textit{MLwiN} worksheet

Each row represents one record

Data Manipulation/View or edit data
Paste data into an *MLwiN* worksheet

Copy text data onto the clipboard from other packages then click **Edit/Paste**.

Import ASCII data from a text file

Click **Locate the pathway of the file** and then enter to cover the columns in a ASCII file **File/ASCII text file Input**.
Name columns - variables

Data Manipulation/Names

eg Name column C2 (variable 2) then press Enter

Save the file as a *MLwiN* worksheet

Click

Name the new file and save it as a *MLwiN* worksheet

File/Save worksheet As...
Declare the missing data

Set a specific missing value code (be sure the same missing value code is used for every variable)

Options/Worksheet/Numbers

Sort a dataset to reflect its hierarchical structure

“before trying to fit a multilevel model to a dataset……the dataset must be sorted so that all records for the same highest-level unit are grouped together and within this group, all records for a particular lower level unit are contiguous”

(Rasbash, et al., 2005)
Sort the dataset

Select which columns to be sorted (those columns must be of the same length)

How many levels

Sort the dataset by which key variables

View hierarchical structure

View the hierarchical structure of a data set after fitting a model

(Note: to View after fitting a model)
Checklist

- All value codes are numerical? ✓
- An identical missing value code? ✓
- The dataset has been sorted? ✓
- The dataset is a MLwin worksheet? ✓

Understand the notation used in *MLwiN*

An example – linear regression with continuous variables $x$ and $y$ for one school with $i$ number of pupils

$$
\hat{y}_i = a + bx_i \quad i = 1, 2, 3\ldots\text{the number of pupils}
$$

$$
y_i = \hat{y}_i + e_i = a + bx_i + e_i
$$

- $e_i = \text{residual (or error) ie, the difference between } y_i \text{ and } \hat{y}_i$ - pupil level
- $a = \text{intercept (average across all pupils)}$
- $b = \text{slope (coefficient – the effect of } x)$

$a$ (intercept) and $b$ (slope of $x$) define the average line across all pupils in the school.
Understand the notation used in *MLwiN*

For one school

\[ y_i = a + bx_i + e_i \]

For a number of schools

\[
\begin{align*}
  y_{i1} &= a_1 + bx_{i1} + e_{i1} \\
  y_{i2} &= a_2 + bx_{i2} + e_{i2} \\
  \cdots \cdots \cdots \\
  y_{ij} &= a_j + bx_{ij} + e_{ij}
\end{align*}
\]

There is

\[ a_j = a + u_j \text{ - school level} \]

Thus

\[ y_{ij} = a + bx_{ij} + u_j + e_{ij} \]

Understand the notation used in *MLwiN*

For a number of schools

\[ y_{ij} = a + bx_{ij} + u_j + e_{ij} \]

Introduce \( x_0 (=1) \) and symbols \( \beta_0 \) and \( \beta_1 \) to denote \( a, b \)

\[
\begin{align*}
  y_{ij} &= \beta_0 x_0 + \beta_1 x_{ij} + u_{0j} x_0 + e_{0ij} x_0 \\
  x_0 \text{ called cons in } MLwiN
\end{align*}
\]

\[ y_{ij} = \beta_{0ij} x_0 + \beta_{1ij} x_{ij} \]

\[ \beta_{0ij} = \beta_0 + u_{0j} + e_{0ij} \]

\( \beta_0 \) and \( \beta_1 \) define the average line across all pupils in all schools.

\[ i = \text{pupil level}, \ j = \text{school level} \]
The fixed and random parts in *MLwiN*

The fixed part of the model:

\[ \beta_0, \beta_1 \] – multilevel modelling regression coefficients

– explained in the model

The random part of the model:

\[ \sigma^2_{u_{ij}} \] – the variance of the school level random effects \( u_{ij} \)

\[ \sigma^2_{e_{ij}} \] – the variance of the pupil level random effects \( e_{ij} \)

– unexplained in the model

(http://tramss.data-archive.ac.uk/documentation/MLwiN/chapter1.pdf)

WJ Peng

Fit a multilevel model in *MLwiN*

- start with simple models -

“Multilevel modelling is like any other type of statistical modelling and a useful strategy is to start by fitting simple models and slowly increase the complexity.”

“It is important...to know as much as possible about your data and to establish what questions you are trying to answer.”

(Rasbash, et al., 2005)
Research questions

We are interested in exploring – via data modelling – the size, nature and extent of the school effect on progress in normexam.

Q1 – What the relationship between the outcome attainment measure normexam and the intake ability measure standlrt would be?

Q2 – How this relationship varies across schools (what the proportions of the overall variability shown in the plot attributable to schools and to student)?


Considering the following 3 models

Cons Model – a random intercepts Null model with ‘normexam’ as the response variable, no predictor/explanatory variables apart from the Constant (ie representing the intercept) which is allowed to vary randomly across schools and with the levels defined as pupils (level 1) in schools (level 2)

Model A – a random intercepts/variance components model
   – Cons Model with also an explanatory variable (standlrt)

Model B – a random intercepts/slopes model
   – Model A with also the parameter associated with standlrt being allowed to vary randomly across schools (ie random slopes as well as intercepts)

Fitting Cons Model

Select Model/Equations from Menu bar

$$y \sim N(XB, \Omega)$$

Notice the red colour parts? – indicating that the variable and the parameter associated with it has not yet been specified

“The response vector has a mean specified in matrix notation by the fixed part $XB$, and a random part consisting of a set of random variables described by the covariance matrix $\Omega$.”

(youtube://tramss.data-archive.ac.uk/documentation/MLwiN/chapter1.pdf)

WJ Peng

Click either of the $y$’s to specify the response variable – normexam, assign $i$ and $j$ at pupil and school levels respectively.

Click either $x_0$ or $\beta_0$ to specify the explanatory variable – cons, assign $i$ and $j$ at pupil and school levels respectively to model the intercept.

Click Name to show the names of the variables.

Notice the variables and parameters have changed from red to black? – indicating that specification is completed.

(https://www.cmm.bristol.ac.uk/research/Lemma/two-level.pdf)
Cons Model has now been specified.

Click the + or - buttons to see the composition of $\beta_{yj}$.

Click Start on Menu bar to start estimation.

Completion of the parameters estimation

Note that the default method of estimation is iterative generalised least squares (IGLS).

The blue highlighted parameters in the Equations window change to green to indicate convergence.
What the parameter estimates tell us?

Note that normexam scores were normalised to have a approximately standard normal distribution

Overall mean -0.013 (approach to zero)

Total variance 0.169 + 0.848 = 1.017 (approach to 1)

(if children were taken from the whole population at random the variance would be)

Intra-school correlation 0.169 / (0.169 + 0.848) = 16.6%

(the proportion of the total variance attributable to the school)

Graphing prediction

In Prediction window, click $\beta_0$ to calculate the average predicted line produced from the intercept coefficient $\beta_0$ - this is the predicted overall mean normexam (= -0.013) line for all pupils in all schools.
Click also $u_{ij}$ to include the estimated school level intercept residuals in the prediction function and produce the predicted lines for all 65 schools. The line for school $j$ departs from the average prediction line by an amount $u_{ij}$.

Click $e_{ij}$ to include the estimated pupil level intercept residuals in the prediction function too. Plot shows identical predicted and observed normexam ($r = 1$). Pupil $i$ in school $j$ departs from the predicted line for school $j$ by an amount $e_{ij}$. 
Fitting Model A
- an random intercepts model -

Click Add Term to add an explanatory variable – standlrt.

(https://tramos.data-archive.ac.uk/documentation/MLwiN/chapter1.pdf)

\[ y_j - \text{N}(0, \Omega_j) \]
\[ y_j = \beta_0 + \beta_1 \text{standlrt}_j + u_0j + e_{0ij} \]
\[ u_0j - \text{N}(0, \Omega_0) : \Omega_0 = \Omega \]
\[ e_{0ij} - \text{N}(0, \Omega_0) : \Omega_0 = \Omega \]

MLwiN detects that cons is constant over the whole data set, whereas the values of standlrt change at both level 1 and level 2.

Click the +, -, and Name buttons to see how much the detail of the model is displayed.

\( \beta_0 \) (the intercept) and \( \beta_1 \) (the slope of standlrt) define the average line across all pupils in all schools.

“The model is made multilevel by allowing each school’s summary line to depart (be raised or lowered) from the average line by an amount \( u_{0j} \).” Pupil \( i \) in the school \( j \) departs from its school’s summary line by an amount \( e_{0ij} \).

(https://tramos.data-archive.ac.uk/documentation/MLwiN/chapter1.pdf)
In other words……

\[ u_{ij} \] – the level 2 or school level residuals (one for each school); distributed Normally with mean 0 and variance \( \sigma^2_u \)

\[ e_{ij} \] – the level 1 or pupil level residuals (one for each pupil); distributed Normally with mean 0 and variance \( \sigma^2_e \)

Estimate the parameters of the specified model

The parameters highlighted in blue are to be estimated.

Click Start on Menu bar to start estimation.
Completion of the parameters estimation

Slope – the slopes of the lines across schools are all the same, of which the common slope is 0.563 with SE = 0.012

Intercept – the intercepts of the lines vary across schools. Their mean is 0.002 with SE = 0.040. The intercept of school $j$ is $0.002 + u_{0j}$ with a variance of 0.092 and SE = 0.018.

What the parameter estimates tell us?

Total variance (0.092 + 0.566 = 0.658) – the sum of the level 2 and level 1 variances

Intra-school correlation (0.092/0.658 = 0.140) – measuring the extent to which pupils’ scores in the same school are more alike as compared with those from pupils at different schools
Graphing predication

\[ \hat{y} = \beta_0 \text{cons} + \beta_1 x \]
\[ \hat{y} = 0.002 + 0.563 \text{standlrt} \]

The average line across all pupils in all schools

Graphing predication

\[ \hat{y}_j = \beta_{0j} \text{cons} + \beta_{1j} x \]
\[ \beta_{0j} = \beta_0 + u_{0j} \]
\[ = 0.002 + u_{0j} \]

One line for each school
Graphing prediction

The line for school \( j \) departs from the average prediction blue line by an amount \( u_{0j} \).

\[
\hat{y} = \beta_0 \text{cons} + \beta_1 x
\]

\[
\beta_{0j} = \beta_0 + u_{0j} + e_{0ij}
\]

\[
\beta_{0j} = 0.002 + u_{0j} + e_{0ij}
\]

Pupil \( i \) in school \( j \) departs from the school \( j \) summary line by an amount \( e_{0ij} \).
What all these about – Model A?

In brief, by employing multilevel modelling approach…

The line for school 3

The average line

The line for school 64

\( u_{03} \) – residual for School 3

\( u_{064} \) – residual for School 64

\( e_{05,64} \) – residual for pupil 5 in school 64

\( e_{028,64} \) – residual for pupil 28 in school 64
Graphing residuals

Residuals for individual schools, of which their mean is 0 and their estimated variance of 0.092

Each vertical line represents a residual with 95% confidence interval estimated for each school.
What is meant by residual?

School residual – the departure of a school (grey) line from the average (blue) line

These school residuals might be regarded as school effect – expressed by the term ‘value added’ in school effectiveness and improvement research.

What is meant by value added?

In this case, value added (or residual) for each school represents the differences between the observed level of school performance (pupil normexam scores taken at age 16) and what would be expected on the basis of pupils’ prior attainment (pupil standlrt scores taken at age 11).

In other words “value added is a measure of the relative progress made by pupil in a school over a particular period of time (usually from entry to the school until public examinations in the case of secondary schools, or over particular years in primary schools – in this case, between age 11 and 16) in comparison to pupils in others schools in the same sample.”

(Thomas, 2005)

Were some schools doing better than others?

- A positive value added score (i.e. residual) indicating a school may be performing above expectation.
- A negative value added score indicating a school may be performing below expectation.

“However, information about the 95% confidence interval (CI) is required to evaluate whether an individual school’s value added performance is likely to have occurred by chance. In other words, the confidence interval is vital to judge whether a school’s performance above or below expectation is statistically significant.” (Thomas, 2005)

So...
Was School▲ doing better than school▼?
Was School▲ doing worse than school▼?
How about school▲ and school▼?
Compare raw (Cons Model) and value added residuals

On average, how were these schools performing in their raw normexam and value added (VA) scores as compared to other schools?

As compared to other schools:
- ▲ performed higher than expected in both scores
- ▲ performed as expected in both scores
- ▲ performed lower than expected in both scores
- ▲ performed as expected in raw but lower than expected in VA
- how about other schools?

Compare level 2 (and level 1) variance between two models

Variance between schools
- which residual curve ~ is steeper
- what the implication of it

Variance within schools
- the likely bounds (95%CI) of variation on schools for raw residuals wider than the ones for value added
- What the implication of it
Fitting Model B
- random intercepts/slopes model -

Model A which we have just specified and estimated assumes that the only variation between schools is in their intercepts. “However, there is a possibility that the school lines have different slopes. This implies that the coefficient of standlrt will vary from school to school.”

Specifying Model B

Click $\beta_j$ to specify the coefficient of standlrt which is random at level 2.

“The terms $u_0j$ and $u_1j$ are random departures or ‘residuals’ at the school level from $\beta_0$ and $\beta_1$. They allow the $j^{th}$ school's summary line to differ from the average line in both its slope and its intercept.”
Specifying Model B

To fit this new model we could click Start as before, but it will probably be quicker to use the estimates already obtained from the 1st model as initial values for the iterative calculations. Click More.

Completion of the parameters estimation

<table>
<thead>
<tr>
<th>mean (SE)</th>
<th>variance (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>individual school slopes vary</td>
<td>0.557 (0.020) 0.015 (0.004)</td>
</tr>
<tr>
<td>school line intercepts vary</td>
<td>-0.012 (0.040) 0.090 (0.018)</td>
</tr>
</tbody>
</table>
"The positive covariance between intercepts and slopes estimated as +0.018 (SE = 0.007) suggests that schools with higher intercepts tend to some extent to have steeper slopes and this corresponds to a correlation between the intercept and slope (across schools) of 0.49."

(https://tramss.data-archive.ac.uk/documentation/MLwiN/chapter1.pdf)

"The pupils' individual scores vary around their schools' lines by quantities $e_{ij}$, the level 1 residuals, whose variance is estimated as 0.554 (SE = 0.012)."

(https://tramss.data-archive.ac.uk/documentation/MLwiN/chapter1.pdf)
**Graphing prediction**

The positive covariance between slopes and intercepts leading to a fanning out pattern when plotting the schools predicted lines (the average line = blue line).

**Graphing residual**

One residual plot for the intercepts of individual school lines
One residual plot for the individual line slopes
Compare the two models

Model A

Model B

Which model is better fit?

Which model was better fit?

“A -2log-likelihood value is the probability of obtaining the observed data if the model were true and can be used in the comparison of two different models.” (Bashsheh, et al., 2005)
Which model is better fit – the likelihood ratio test

- the change of the two \(-2\log\)-likelihood values
  
  \[ 9357.2 - 9316.9 = 40.3. \]

- the change in the \(-2\log\)-likelihood value (which is also the change in deviance) has a chi-squared distribution on 2 degrees of freedom under the null hypothesis that the extra parameters have population values of zero.

- two extra parameters involved in the 2nd model
  
  (1) the variance of the slope residuals \(u_{1j}\)
  
  (2) their covariance with the intercept residuals \(u_{0j}\)

  \((\text{Rasbash, et al., 2005})\)

The change is very highly significant, confirming the better fit of the 2nd model, a more elaborate model to the data.

---

Examples of other modelling

Gender effects
- Do girls make more progress than boys? (F)
- Are boys more or less variable in their progress than girls? (R)

Contextual effects
- Are pupils in key schools less variable in their progress? (R)
- Do pupils do better in urban schools (or key schools)? (F)
- Does gender gap vary across schools? (R)

Cross-level interaction
- Do boys learn more effectively in a boys’ or mixed sex school? (F)
- Do low ability pupils fare better when educated alongside higher ability pupils? (F)

\((\text{Jones, 2007; Rasbash, et al., 2005})\)
Examples of other hierarchical structures in education settings

<table>
<thead>
<tr>
<th>Level 1</th>
<th>level 2</th>
<th>level 3</th>
<th>level 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pupils</td>
<td>classes</td>
<td>schools</td>
<td></td>
</tr>
<tr>
<td>Pupils</td>
<td>schools</td>
<td>regions</td>
<td></td>
</tr>
<tr>
<td>Pupils</td>
<td>schools</td>
<td>regions</td>
<td>countries</td>
</tr>
<tr>
<td>Pupils</td>
<td>neighbourhoods</td>
<td>schools</td>
<td>regions</td>
</tr>
</tbody>
</table>

NB:
What happens if we have other types of data (eg ordered/unordered categorical data, binomial/multinomial data, repeated data) or non-hierarchical structure (eg pupils changing schools)?

Useful references/links

Getting start with the concept of value added in school effectiveness and improvement research:


Getting start with how to fit a model in MLwiN:

Centre for Multilevel Modelling, Graduate School of Education, University of Bristol
http://www.cmm.bristol.ac.uk/MLwiN/index.shtml

Teaching Resources and Materials for Social Scientists, ESRC
http://tramss.data-archive.ac.uk/documentation/MLwiN/what-is.asp