

Module 5: Introduction to Multilevel Modelling

R Practical

Camille Szmaragdand George Leckie¹
Centre for Multilevel Modelling

Pre-requisites

- Modules 1-4

Contents

Introduction to the Scottish Youth Cohort Trends Dataset.....	1
P5.1 Comparing Groups using Multilevel Modelling.....	3
P5.1.1 A multilevel model of attainment with school effects	3
P5.1.2 Examining school effects (residuals).....	6
P5.2 Adding Student-level Explanatory Variables: Random Intercept Models.....	11
P5.3 Allowing for Different Slopes across Schools: Random Slope Models	16
P5.3.1 Testing for random slopes.....	17
P5.3.2 Interpretation of random cohort effects across schools	18
P5.3.3 Examining intercept and slope residuals for schools	18
P5.3.4 Between-school variance as a function of cohort.....	21
P5.3.5 Adding a random coefficient for gender (dichotomous x)	23
P5.3.6 Adding a random coefficient for social class (categoricalx).....	25
P5.4 Adding Level 2 Explanatory Variables	29
P5.4.1 Contextual effects.....	30
P5.4.2 Cross-level interactions.....	32
P5.5 Complex Level 1 Variation	35
P5.6 References	36

¹ This R practical is adapted from the corresponding MLwiN practical: Steele, F. (2008) Module 5: Introduction to Multilevel Modelling. LEMMA VLE, Centre for Multilevel Modelling. Accessed at <http://www.cmm.bris.ac.uk/lemma/course/view.php?id=13>.

Some of the sections within this module have online quizzes for you to test your understanding. To find the quizzes:

EXAMPLE

From within the LEMMA learning environment

- Go down to the section for **Module 5: Introduction to Multilevel Modelling**
- Click "[5.1 Comparing Groups Using Multilevel Modelling](#)" to open Lesson 5.1
- Click [Q1](#) to open the first question

Introduction to the Scottish Youth Cohort Trends Dataset

You will be analysing data from the Scottish School Leavers Survey (SSLS), a nationally representative survey of young people. We use data from seven cohorts of young people collected in the first sweep of the study, carried out at the end of the final year of compulsory schooling (aged 16-17) when most sample members had taken Standard grades.²

In the practical for Module 3 on multiple regression, we considered the predictors of attainment in Standard grades (subject-based examinations, typically taken in up to eight subjects). In this practical, we extend the (previously single-level) multiple regression analysis to allow for dependency of exam scores within schools and to examine the extent of between-school variation in attainment. We also consider the effects on attainment of several school-level predictors.

The dependent variable is a total attainment score. Each subject is graded on a scale from 1 (highest) to 7 (lowest) and, after recoding so that a high numeric value denotes a high grade, the total is taken across subjects. The analysis dataset contains the student-level variables considered in Module 3 together with a school identifier and three school-level variables:

Variable name	Description and codes
caseid	Anonymised student identifier
schoolid	Anonymised school identifier
score	Point score calculated from awards in Standard grades taken at age 16. Scores range from 0 to 75, with a higher score indicating a higher attainment
cohort90	The sample includes the following cohorts: 1984, 1986, 1988, 1990, 1996 and 1998. The cohort90 variable is calculated by

² We are grateful to Linda Croxford (Centre for Educational Sociology, University of Edinburgh) for providing us with these data. The dataset was constructed as part of an ESRC-funded project on Education and Youth Transitions in England, Wales and Scotland 1984-2002.

Further analyses of the data can be found in Croxford and Raffe(2006).

	subtracting 1990 from each value. Thus values range from -6 (corresponding to 1984) to 8 (1998), with 1990 coded as zero
female	Sex of student (1 = female, 0 = male)
sclass	Social class, defined as the higher class of mother or father (1 = managerial and professional, 2 = intermediate, 3 = working, 4 = unclassified)
sctype	School type, distinguishing independent schools from state-funded schools (1 = independent, 0 = state-funded)
schurban	Urban-rural classification of school (1 = urban, 0 = town or rural)
schdenom	School denomination (1 = Roman Catholic, 0 = non-denominational)

There are 33,988 students in 508 schools.

P5.1 Comparing Groups using Multilevel Modelling

Download the Rdataset for this lesson:

From within the LEMMA Learning Environment

- Go to **Module 5: Introduction to Multilevel Modelling**, and scroll down to **Rdatasets and Rfiles**
- Right click “5.1.txt” and select **Save Link As ...** to save the dataset to your computer

Read the dataset into R using the `read.table` command and create a dataframe object named `mydata`³:

```
> mydata<- read.table(file = "5.1.txt", sep = ",", header = TRUE)
```

and use the `str` command to produce a summary of the dataset:

```
> str(mydata)
'data.frame': 33988 obs. of 9 variables:
 $ caseid : int 18 17 19 20 21 13 16 14 15 12 ...
 $ schoolid: int 1 1 1 1 1 1 1 1 1 1 ...
 $ score : int 0 10 0 40 42 4 0 0 14 27 ...
 $ cohort90: int -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 ...
 $ female : int 1 1 1 1 1 1 1 1 1 1 ...
 $ sclass : int 2 2 4 3 2 2 3 4 3 2 ...
 $ schtype : int 0 0 0 0 0 0 0 0 0 0 ...
 $ schurban: int 1 1 1 1 1 1 1 1 1 1 ...
 $ schdenom: int 0 0 0 0 0 0 0 0 0 0 ...
```

P5.1.1 A multilevel model of attainment with school effects

We will start with the simplest multilevel model which allows for school effects on attainment, but without explanatory variables. This ‘null’ model may be written:

$$\text{score}_{ij} = \beta_0 + u_{0j} + e_{ij}$$

where score_{ij} is the attainment of student i in school j , β_0 is the overall mean across schools, u_{0j} is the effect of school j on attainment, and e_{ij} is a student-level residual. The school effects u_{0j} , which we will also refer to as school (or level 2) residuals, are assumed to follow a normal distribution with mean zero and variance $\sigma_{u_0}^2$.

R’s main command for fitting multilevel models is part of the additional `lme4`⁴ library which can be installed through the R Packages menu; select Install Package(s)

³At the beginning of your R session, you will need to set R’s working directory to the file location where you saved the dataset. This can be done using the command line and the function `setwd`:

```
> setwd("C:/userdirectory")
```

Or through selecting Change dir... on the File menu.

⁴`lme4` is a package developed by Douglas Bates and Martin Maechler for fitting linear and generalized linear mixed-effect models.

and then select the correct Mirror and package from the scroll-down menus. As you will see, there is a variety of additional packages that can be installed with R. You only need to install a package once to your own computer. If you then want to use the package, you simply need to call it from within R prior to using the command for the first time in each R session. This can be done with the `library()` function and in this case `library(lme4)`.

```
> library(lme4)
Loading required package: Matrix
```

The output informs us that R has loaded the additional package `Matrix` which is required by the `lme4` package to work.

We will use the `lmer()` function from the `lme4` library to fit the above model. The syntax for this function is very similar to the syntax used for the `lm()` function for multiple regression which we introduced in Module 3.⁵ Below we choose to store the model as a new object called **nullmodel**:

```
> nullmodel <- lmer(score ~ (1|schoolid), data = mydata, REML = FALSE)
```

The response variable (**score**) follows the command which is then followed by a `~` and then by a list of fixed part explanatory variables (excluding the constant as this is included by default)⁶. The above model contains only an intercept and so no fixed part explanatory variables are specified. The level 2 random part of the model is specified in brackets by the list of random part explanatory variables (the constant has to be explicitly specified by `1`, followed by a single vertical bar `|` and then by the level 2 identifier (**schoolid**). The `data` option specifies the dataframe being used to fit the model. The `REML=FALSE` option is used to request maximum likelihood estimation (as opposed to the default of restricted maximum likelihood estimation).

We then display the results using the `summary` command, which gives the following output:

```
> summary(nullmodel)
Linear mixed model fit by maximum likelihood ['lmerMod']
Formula: score ~ (1 | schoolid)
Data: mydata

           AIC          BIC       logLik   deviance  df.resid
286545.1  286570.4 -143269.5  286539.1    33985

Scaled residuals:
   Min       1Q   Median       3Q      Max
-2.9763 -0.7010  0.1017  0.7391  3.0817

Random effects:
 Groups   Name                Variance Std.Dev.
schoolid (Intercept)    61.02      7.812
Residual                258.36     16.073
Number of obs: 33988, groups: schoolid, 508
```

⁵To obtain details of the different options available for the `lmer()` function, just type `help("lmer")`

⁶Note, to omit the constant you need to add `-1` to the right-hand side of the `"~"` sign.

```
Fixed effects:
              Estimate Std. Error t value
(Intercept)  30.6006      0.3694   82.83
```

Before interpreting the model, we will discuss the estimation procedure that `lmer` uses.⁷ The estimation procedure used by both MLE and REML is based on optimizing a function of the log likelihood using penalized iteratively re-weighted least squares. The log-likelihood is evaluated using an adaptive Gauss-Hermite approximation, which, when using the default value of one, reduces to the Laplacian approximation. This default approximation can be changed by using the `nAGQ=n` option, where `n` is an integer greater than one, representing the number of points used for evaluating the adaptive Gauss-Hermite approximation. The greater the value of `n`, the more accurate the evaluation of the log-likelihood, but the longer it takes to fit the model.

The output of `lmer` consists of three parts. The first part under `Formula:` and `Data:` reports a range of summary statistics (AIC, BIC, LogLik,...). The second part under `Random effects:` summarises the variance and standard deviation of each random effect (including the level 1 model residuals). Underneath the random effects table, the total number of observations is provided along with the number of units (or groups) for each higher level in the model. Here, schools are our only higher level and the output reports that we have 508 different schools. The final part of the output is the `Fixed effects:` table which reports the parameter estimate (`Estimate`) standard error (`Std. Error`) and t-value (`t value`), for each parameter in the model. For models with more than one fixed part explanatory variable (including the intercept), a correlation table between these variables is also provided underneath the table of parameter estimates (see later examples).

The overall mean attainment (across schools) is estimated as 30.60. The mean for school j is estimated as $30.60 + \hat{u}_{0j}$, where \hat{u}_{0j} is the school residual which we will estimate in a moment. A school with $\hat{u}_{0j} > 0$ has a mean that is higher than average, while $\hat{u}_{0j} < 0$ for a below-average school. (We will obtain confidence intervals for residuals to determine whether differences from the overall mean can be considered 'real' or due to chance.)

Partitioning variance

The between-school (level 2) variance `schoolid (Intercept)` in attainment is estimated as $\hat{\sigma}_{u_0}^2 = 61.02$, and the within-school between-student (level 1) variance `Residual` is estimated as $\hat{\sigma}_e^2 = 258.36$. Thus the total variance is $61.02 + 258.36 = 319.38$.

The variance partition coefficient (VPC) is $61.02/319.38 = 0.19$, which indicates that 19% of the variance in attainment can be attributed to differences between schools. Note, however, that we have not accounted for intake ability (measured by exams taken on entry to secondary school) so the school effects are not value-added.

⁷For further details see the PDF vignettes available on the `lme4` website <http://cran.r-project.org/web/packages/lme4>, in particular the vignette entitled "Computational Methods" which deals with the statistical theory.

This document is only the first few pages of the full version.

To see the complete document please go to learning materials and register:

<http://www.cmm.bris.ac.uk/lemma>

The course is completely free. We ask for a few details about yourself for our research purposes only. We will not give any details to any other organisation unless it is with your express permission.