Module 11: Three-Level Multilevel Models

R Practical

Tamas Novak and George Leckie[1]
Centre for Multilevel Modelling

Pre-requisites

- Modules 1-5

Contents

[1] This Practical is adapted from the corresponding Stata practical: Leckie, G. (2013). Three-Level Multilevel Models – Stata Practical. LEMMA VLE Module 11, 1-52. Accessed at: https://www.bristol.ac.uk/cmm/learning/online-course/.

If you find this module helpful and wish to cite it in your research, please use the following citation:

Tamas, N. & Leckie, G. (2025). Three-Level Multilevel Models – R Practical. LEMMA VLE Module 11, 1-46. https://www.bristol.ac.uk/cmm/learning/online-course

Address for correspondence:

Tamas Novak
Centre for Multilevel Modelling
University of Bristol
35 Berkeley Square
Bristol, BS8 1JA
UK

tamas.novak@bristol.ac.uk

George Leckie
Centre for Multilevel Modelling
University of Bristol
35 Berkeley Square
Bristol, BS8 1JA
UK

g.leckie@bristol.ac.uk

Introduction to the Television School and Family Smoking Prevention and Cessation Project

We will analyse data from the Television School and Family Smoking Prevention and Cessation Project (TVSFP) (Flay *et al.*, 1989). The project was designed to test the effect of two different school-based interventions on student tobacco and health knowledge: (1) A social-resistance classroom curriculum (CC); and (2) A television-based programme.

The study sample involved schools with seventh-grade students (age 12 to 13 years) in Los Angeles and San Diego, California. Schools were randomized to one of the four study conditions formed by crossing the two interventions in a 2 × 2 design.

		Television-based programme (TV)	
		No	**Yes**
Classroom Curriculum (CC)	**No**	Neither intervention	TV only
	Yes	CC only	CC and TV

The two interventions were delivered to the seventh-grade students in these schools in spring 1986. Students were baselined in January 1986, completed an immediate postintervention questionnaire in April 1986, a one-year follow-up questionnaire in April 1987, and a two-year follow-up questionnaire in April 1988. At each time point, students' knowledge was assessed using a tobacco and health knowledge scale (THKS), constructed as the number of correct answers to seven binary questionnaire items.

The data were restudied by Hedeker *et al.* (1994) who used them to illustrate the importance of clustering in clinical and public health research and how multilevel models could be used to account for two-level and three-level hierarchical clustering structures. They concentrated on the sub sample of students who studied at 28 Los Angeles schools and only analysed data from the baseline and postintervention time points. Students who missed data at either time point were listwise deleted.

In this Module, we will explore the three-level hierarchical structure of the data: students (level 1) in classrooms (level 2) in schools (level 3). We will fit three-level multilevel models to examine the relative importance of schools and classrooms as influences on student tobacco and health knowledge and we will pay particular attention to assessing the possible causal effects of the CC and TV interventions.

There is good reason to expect both school and classroom effects on students' THKS scores. While schools were randomly assigned to the four study conditions, implementation of the CC and TV interventions were carried out at the classroom level. It seems very likely that some schools and teachers would have been more enthused about the interventions than others and this is likely to have had a direct effect on the success of the interventions. We therefore expect to see both between-school and within-school-between-classroom variation in students' THKS

scores, even after accounting for baseline differences in their tobacco and health knowledge.

We use the Hedeker *et al*. sub sample of the original data. The data consist of 1,600 students (level 1) nested within 135 classrooms (level 2) nested within 28 schools (level 3).

The response variable is students' postintervention THKS. We shall treat this score as a continuous response variable in our multilevel models, though we note that we could equally treat this response as ordinal and therefore fit ordinal response multilevel models (see Module 9). The predictor variables of key interest are the school level binary indicators of whether each school was randomly assigned to the CC or TV interventions. The predictor variables also include students' baseline THKS scores. We will include this predictor variable in our models to adjust for baseline variation in students' tobacco and health knowledge.

The dataset contains the following variables

Variable name	Description and codes
schoolid	School ID
classid	Class ID
studentid	Student ID
postthks	Postintervention THKS score. Scores range from 0 to 7, with a higher score indicating a higher tobacco and health knowledge
prethks	Baseline THKS score. Scores are measured on the same scale as **postthks**.
cc	Classroom curriculum (CC) (0 = no CC, 1 = CC)
tv	Television (TV) (0 = no TV, 1 = TV)
ccXtv	CC × TV, the interaction between CC and TV. The variable is constructed by multiplying the variables **cc** and **tv**. Note that **ccXtv** is also binary and 1 = both CC and TV and 0 otherwise.

P11.1 Examining and Describing the Data

From within the LEMMA learning environment
- Go to **Module 11: Three-Level Multilevel Models**, and scroll down to **R files**
- Click '**11.1.rds** to download the dataset

Open the R script for this lesson, read the dataset using the `readRDS()` function and save it to a variable called `mydata`.

```
> mydata <- readRDS('11.1.rds')
```

Use the `str()` function to produce a summary of the dataset. The `str()` function displays the structure of an R object in a compact way. Running `str()` on a data frame object (like `mydata`) will result in a console output that lists the dimensions of the data frame as well as the names, data types, and first ten of values of each column.

```
> str(mydata)
'data.frame':  1600 obs. of  8 variables:
$ schoolid : int  193 193 193 193 193 193 193 193 193 193 ...
$ classid  : int  193101 193101 193101 193101 193101 193101 193101 193101 193101
193101 ...
$ studentid: int  1 2 3 4 5 6 7 8 9 10 ...
$ postthks : int  2 2 3 2 1 2 4 2 3 3 ...
$ prethks  : int  1 3 0 3 1 2 3 3 3 1 ...
$ cc       : int  0 0 0 0 0 0 0 0 0 0 ...
$ tv       : int  0 0 0 0 0 0 0 0 0 0 ...
$ ccXtv    : int  0 0 0 0 0 0 0 0 0 0 ...
```

The data consist of 1,600 observations on 8 variables.

Standard univariate summary statistics can be requested using the `summary()` function.

```
> summary(mydata)
schoolid        classid          studentid
Min.   :193.0   Min.   :193101   Min.   :   1.0
1st Qu.:405.0   1st Qu.:405101   1st Qu.: 400.8
Median :415.0   Median :415105   Median : 800.5
Mean   :421.9   Mean   :422042   Mean   : 800.5
3rd Qu.:509.0   3rd Qu.:509106   3rd Qu.:1200.2
Max.   :515.0   Max.   :515113   Max.   :1600.0

postthks        prethks         cc
Min.   :0.000   Min.   :0.000   Min.   :0.0000
1st Qu.:2.000   1st Qu.:1.000   1st Qu.:0.0000
Median :3.000   Median :2.000   Median :0.0000
Mean   :2.662   Mean   :2.069   Mean   :0.4769
3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:1.0000
Max.   :7.000   Max.   :6.000   Max.   :1.0000

tv               ccXtv
Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000
Median :0.0000   Median :0.0000
Mean   :0.4994   Mean   :0.2394
3rd Qu.:1.0000   3rd Qu.:0.0000
```

```
Max.   :1.0000   Max.   :1.0000
```

We see, for example, that the response variable **postthks** ranges from 0 to 7. We shall describe a range of summary statistics for the response and predictor variables in P11.1.2.

P11.1.1 Exploring the three-level data structure

We start by using the `head()` function to list the data for the first couple of students in the data frame. We set the argument `n` to `10` to look at the first 10 students.

```
> head(mydata, n = 10)
   schoolid classid studentid postthks prethks cc tv ccXtv
1       193  193101         1        2       1  0  0     0
2       193  193101         2        2       3  0  0     0
3       193  193101         3        3       0  0  0     0
4       193  193101         4        2       3  0  0     0
5       193  193101         5        1       1  0  0     0
6       193  193101         6        2       2  0  0     0
7       193  193101         7        4       3  0  0     0
8       193  193101         8        2       3  0  0     0
9       193  193101         9        3       3  0  0     0
10      193  193101        10        3       1  0  0     0
```

We see, for example, that student 1 was taught in class 193101 within school 193. The student scored 1 out of 7 on the THKS at baseline (**prethks**) and 2 out of 7 at postintervention (**postthks**). The variables **cc** and **tv** (and therefore **ccXtv**) are both zero and so school 193 received neither intervention.

Next, we use the `unique()` and `length()` functions to confirm that the number of schools and classrooms in the data are 28 and 135, respectively. The `unique()` function removes duplicates from a vector, data frame, or array like objects. In our example it removes duplicate school IDs and class IDs. The `length()` function returns the length of an object. We can count how many schools and classes we have by running the `length()` function on the unique array of school and class IDs.

```
> length(unique(mydata$schoolid))
[1] 28
> length(unique(mydata$classid))
[1] 135
```

Let's explore the unbalanced nature of the three-level hierarchy: the number of classrooms varies across schools and the number of students varies across classrooms.

First, we count the number of classes within each school. We will use the `aggregate()` function for this. The `aggregate()` function groups data by a given variable and runs a function (e.g., `mean()`) on each of the subsets and returns the results per group in one data frame. In this case we subset `mydata$classid` by `mydata$schoolid`. We set the `x` and `by` arguments to named lists created from `mydata$classid` and `mydata$schoolid` respectively. The names `nclasses`