Module 11: Three-Level Multilevel Models

MLwiN Practical[1]

George Leckie and Rob French
Centre for Multilevel Modelling

Pre-requisites

- Modules 1-5

Contents

[1] This MLwiN practical is adapted from the corresponding Stata practical: Leckie, G. (2013). Three-Level Multilevel Models – Stata Practical. LEMMA VLE Module 12, 1-52. Accessed at http://www.bristol.ac.uk/cmm/learning/course.html.

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

Leckie, G. and French, R. (2013). Three-Level Multilevel Models – MLwiN Practical. LEMMA VLE Module 11, 1-50. http://www.bristol.ac.uk/cmm/learning/course.html

Address for correspondence:

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.
cons	A column of ones. This variable will be included as an explanatory variable in all models and its coefficient will be the intercept.

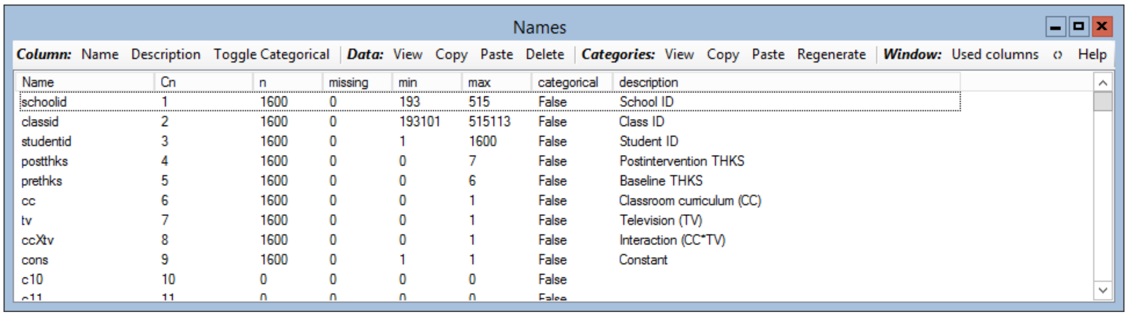
P11.1 Examining and Describing the Data

Open the worksheet ‘11.1.wsz’

From within the LEMMA Learning Environment

- Go to **Module 11: Three-Level Multilevel Models**, and scroll down to **MLwiN Datafiles**
- Click ‘**11.1.wsz**’ to open the worksheet

The **Names** window will appear.


Names

Column: Name Description Toggle Categorical | Data: View Copy Paste Delete | Categories: View Copy Paste Regenerate | Window: Used columns Help

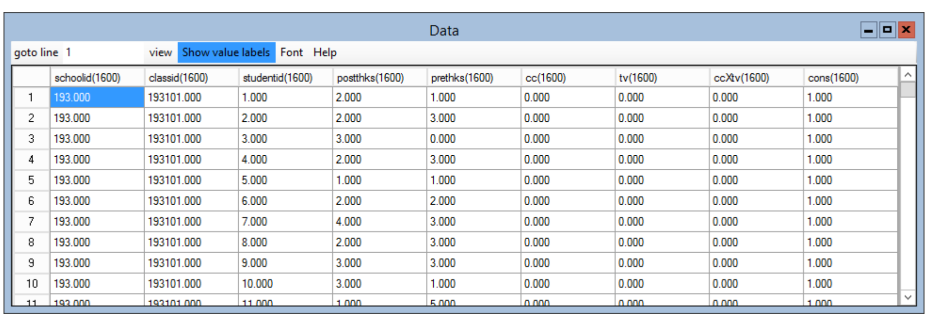
Name	Cn	n	missing	min	max	categorical	description
schoolid	1	1600	0	193	515	False	School ID
classid	2	1600	0	193101	515113	False	Class ID
studentid	3	1600	0	1	1600	False	Student ID
postthks	4	1600	0	0	7	False	Postintervention THKS
prethks	5	1600	0	0	6	False	Baseline THKS
cc	6	1600	0	0	1	False	Classroom curriculum (CC)
tv	7	1600	0	0	1	False	Television (TV)
ccXtv	8	1600	0	0	1	False	Interaction (CC*TV)
cons	9	1600	0	1	1	False	Constant
c10	10	0	0	0	0	False	

The data consist of 1,600 observations on 9 variables and each variable has been given a variable label. 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 looking in more detail at the structure of the data for the first 10 students.

- In the **Names** window, select all nine variables **schoolid** through to **cons** (use the **Shift** button on the keyboard to select multiple variables)
- Under the **Data** toolbar of the **Names** window, click **View**


Data

goto line 1 view Show value labels Font Help

	schoolid(1600)	classid(1600)	studentid(1600)	postthks(1600)	prethks(1600)	cc(1600)	tv(1600)	ccXtv(1600)	cons(1600)
1	193.000	193101.000	1.000	2.000	1.000	0.000	0.000	0.000	1.000
2	193.000	193101.000	2.000	2.000	3.000	0.000	0.000	0.000	1.000
3	193.000	193101.000	3.000	3.000	0.000	0.000	0.000	0.000	1.000
4	193.000	193101.000	4.000	2.000	3.000	0.000	0.000	0.000	1.000
5	193.000	193101.000	5.000	1.000	1.000	0.000	0.000	0.000	1.000
6	193.000	193101.000	6.000	2.000	2.000	0.000	0.000	0.000	1.000
7	193.000	193101.000	7.000	4.000	3.000	0.000	0.000	0.000	1.000
8	193.000	193101.000	8.000	2.000	3.000	0.000	0.000	0.000	1.000
9	193.000	193101.000	9.000	3.000	3.000	0.000	0.000	0.000	1.000
10	193.000	193101.000	10.000	3.000	1.000	0.000	0.000	0.000	1.000

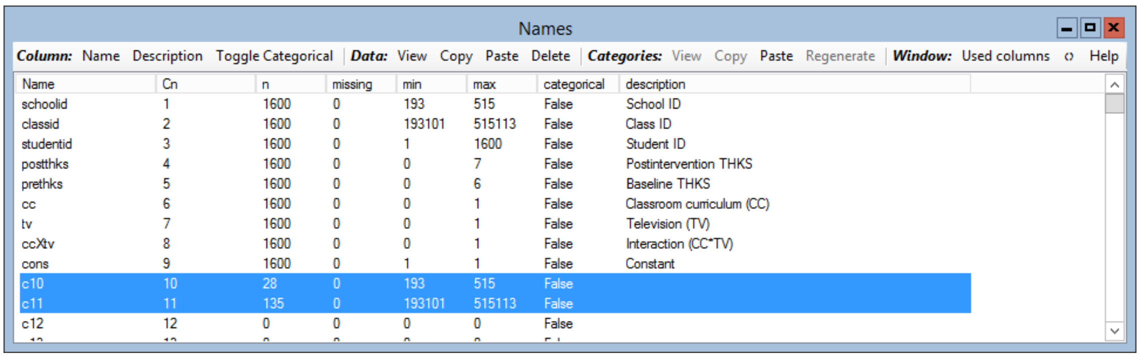
We see, for example, that student 1 was taught in class 193101 within school 193. The student scored 1 out of 7 on the THKS at baseline (**prethks**) and 2 out of 7 at postintervention (**postthks**). The variables **cc** and **tv** (and therefore **ccXtv**) are both zero and so school 193 received neither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

Next, we use the **Command interface** window to confirm that the number of schools and classrooms in the data are 28 and 135, respectively. Specifically, we use the **UNIQ** command to generate new 'short' versions of the school and classroom identifier variables which take one record per group.

- From the **Data Manipulation** menu, select **Command interface**
- Type the following into the bottom pane of the window and press **Enter** after typing each command

```
UNIQ 'schoolid' c10
UNIQ 'classid' c11
```

The **Names** window should update and show the following.



Name	Cn	n	missing	min	max	categorical	description
schoolid	1	1600	0	193	515	False	School ID
classid	2	1600	0	193101	515113	False	Class ID
studentid	3	1600	0	1	1600	False	Student ID
postthks	4	1600	0	0	7	False	Postintervention THKS
prethks	5	1600	0	0	6	False	Baseline THKS
cc	6	1600	0	0	1	False	Classroom curriculum (CC)
tv	7	1600	0	0	1	False	Television (TV)
ccXtv	8	1600	0	0	1	False	Interaction (CC*TV)
cons	9	1600	0	1	1	False	Constant
c10	10	28	0	193	515	False	
c11	11	135	0	193101	515113	False	
c12	12	0	0	0	0	False	

The new variable **c10** now contains a single record for each unique school, while the new variable **c11** contains a single record for each unique classroom. The number of records for each of these new variables, 28 and 135, confirms that there are indeed 28 schools and 135 classrooms in the data.

Next, we will explore the distribution of schools, classrooms and students across the four study conditions outlined in our introduction to the data: (1) Neither intervention; (2) CC only; (3) TV only; and (4) CC and TV.

Tabulating **cc** by **tv** at the school level shows seven schools were assigned to each condition.[2] The data are therefore balanced, at the school level, across conditions. Note, however, that balance at level 3, or any other level, is by no means a requirement when fitting three-level, or any other, multilevel models.

		Television-based programme (TV)		
		No	**Yes**	**Total**
Classroom Curriculum (CC)	**No**	7	7	14
	Yes	7	7	14
	Total	14	14	28

[2] We do not present the step-by-step instructions to replicate this or subsequent cross-tabulations shown in this section as they are somewhat involved. Such higher-level cross-tabulations are more easily carried out in standard statistical software packages such as R, SPSS or Stata.