Module 12: Cross-Classified Multilevel Models

MLwiN Practical[1]

George Leckie and Andrew Bell
Centre for Multilevel Modelling

Pre-requisites

- Modules 1-5,11

Contents

[1] This MLwiN practical is adapted from the corresponding Stata practical: Leckie, G. (2013). Cross-Classified 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 Bell, A. (2013). Cross-Classified Multilevel Models – MLwiN Practical. LEMMA VLE Module 12, 1-56.
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 Scotland Neighbourhood Study

We will analyse data from the Scotland Neighbourhood Study (Garner and Raudenbush, 1991). This study set out to test the hypothesis that a neighbourhood's level of social deprivation has a negative effect on a student's educational attainment even after controlling for the student's prior attainment and family background. The data were subsequently restudied by Raudenbush (1993) and were also used as one of the examples in the classic *Hierarchical Linear Models* textbook (Raudenbush and Bryk, 2002).

The data relate to a single education authority in Scotland and consist of 2,310 students who attended 17 secondary schools and resided in 524 neighbourhoods. Secondary schools teach students from age 11-12 to the end of compulsory schooling (age 15-16). The neighbourhoods are defined as the enumeration districts within which students lived. (The education authority in this study corresponds to a school district in the U.S., while the secondary schools correspond to high schools and the neighbourhoods are similar in size to U.S. census tracts.) The data are not, however, strictly hierarchical. Not all students from the same neighbourhood attend the same school and so the data do not form a three-level hierarchy of students (level 1) within neighbourhoods (level 2) within schools (level 3).[2] Similarly, not all students from the same school live in the same neighbourhood and so neither do the data form a three-level hierarchy of students (level 1) within schools (level 2) within neighbourhoods (level 3). Rather, students are nested within the cells of a two-way cross-classification of schools-by-neighbourhoods.

In the current analyses, we will explore this non-hierarchical cross-classified data structure and we will fit cross-classified multilevel models to examine the relative importance of schools and neighbourhoods as sources of variation in student educational attainment. The analyses will replicate many of the results presented for these data by Raudenbush (1993) and Raudenbush and Bryk (2002).

The response variable is a total attainment score, based on a series of national examinations taken at the end of compulsory secondary schooling in Scotland (age 16). Successful performance in these examinations is a crucial factor in decisions regarding employment or further post compulsory education possibly leading to entrance to universities. Higher scores indicate higher attainment. Predictor variables include student level verbal reasoning and reading prior attainment scores on entering secondary education, student gender, a range of family level background characteristics, and a neighbourhood level deprivation score.

[2] See Module 11 for an introduction to multilevel models for three- and higher-level hierarchical data structures.

The dataset contains the following variables

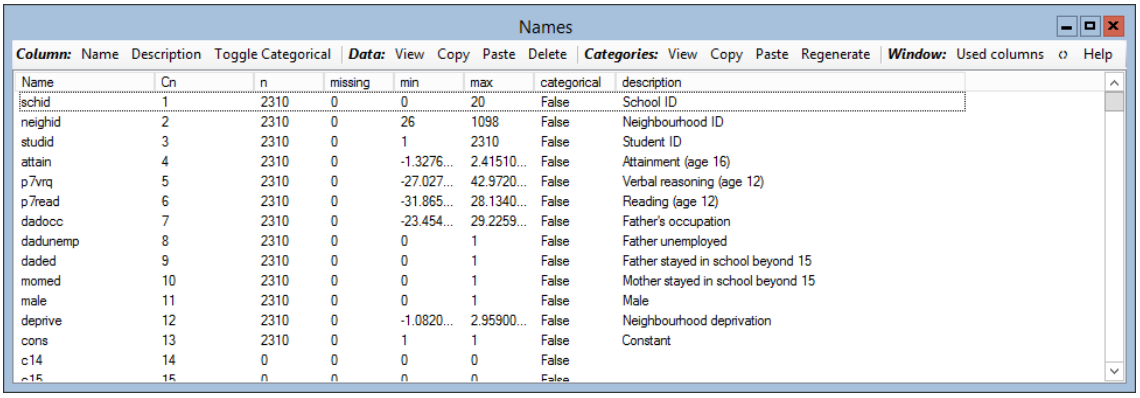
Variable name	Description and codes
schid	School ID
neighid	Neighbourhood ID
studid	Student ID
attain	Total attainment, based on a series of national examinations taken at the end of compulsory secondary schooling in Scotland (age 16). The variable is approximately standardised. Scores range from -1.328 to 2.415.
p7vrq	Verbal reasoning at the end of primary schooling (age 12). The variable is centred on the mean for the study area. Scores range from -27.028 to 42.972.
p7read	Reading attainment at the end of primary schooling (age 12). The variable is centred on the mean for the study area. Scores range from -31.866 to 28.134.
dadocc	Father's occupation, a proxy for social class. The variable is centred on the mean for the study area. Scores range from -23.454 to 29.226.
dadunemp	Father unemployed (0 = employed, 1 = unemployed).
daded	Father stayed in school beyond 15 (0 = left school, 1 = stayed in school).
momed	Mother stayed in school beyond 15 (0 = left school, 1 = stayed in school).
male	Male (0 = female, 1 = male).
deprive	Neighbourhood deprivation, with higher scores indicating neighbourhoods with higher concentrations of poverty, worse health and poorer housing stock than neighbourhoods with lower scores. The measure is designed to have a mean of 0 and a standard deviation of 1 for all Scotland. The sample mean and standard deviation are 0.037 and 0.622 and scores range from -1.082 to 2.959. Thus, the education authority under study has a similar mean level of deprivation to the average for Scotland, but is more homogenous than the country as a whole.
cons	A column of ones. This variable will be included as an explanatory variable in all models and its coefficient will be the intercept.

P12.1 Examining and Describing the Data

Open the worksheet '12.1.wsz'

From within the LEMMA learning environment
- Go to **Module 12: Cross-Classified Multilevel Models**, and scroll down to **MLwiN Datafiles**
- Click '**12.1.wsz**' to open the worksheet

The **Names** window will appear.



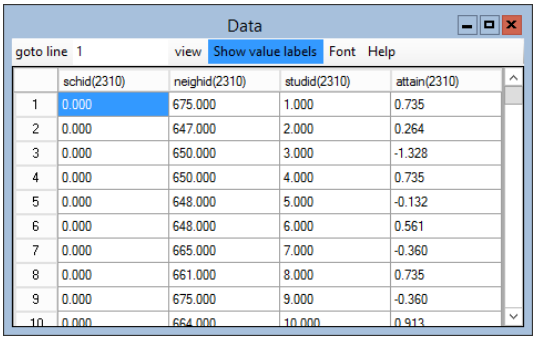
Name	Cn	n	missing	min	max	categorical	description
schid	1	2310	0	0	20	False	School ID
neighid	2	2310	0	26	1098	False	Neighbourhood ID
studid	3	2310	0	1	2310	False	Student ID
attain	4	2310	0	-1.3276...	2.41510...	False	Attainment (age 16)
p7vrq	5	2310	0	-27.027...	42.9720...	False	Verbal reasoning (age 12)
p7read	6	2310	0	-31.865...	28.1340...	False	Reading (age 12)
dadocc	7	2310	0	-23.454...	29.2259...	False	Father's occupation
dadunemp	8	2310	0	0	1	False	Father unemployed
daded	9	2310	0	0	1	False	Father stayed in school beyond 15
momed	10	2310	0	0	1	False	Mother stayed in school beyond 15
male	11	2310	0	0	1	False	Male
deprive	12	2310	0	-1.0820...	2.95900...	False	Neighbourhood deprivation
cons	13	2310	0	1	1	False	Constant
c14	14	0	0	0	0	False	
c15	15	0	0	0	0	False	

The data consist of 2,310 observations on 13 variables and each variable has been given a variable label. We see, for example, that the response variable **attain** ranges from -1.328 to 2.415. We shall describe a range of summary statistics for the response and predictor variables in P12.1.2.

P12.1.1 Exploring the cross-classified data structure

We start by looking in more detail at the structure of the data, specifically at the school (**schid**), neighbourhood (**neighid**) and student (**studid**) identifier variables, and the attainment score response variable (**attain**).

- In the **Names** window, select the variables **schid**, **neighid**, **studid** and **attain** (use the **Shift** button on the keyboard to select multiple variables)
- Under the **Data** toolbar of the **Names** window, click **View**



	schid(2310)	neighid(2310)	studid(2310)	attain(2310)
1	0.000	675.000	1.000	0.735
2	0.000	647.000	2.000	0.264
3	0.000	650.000	3.000	-1.328
4	0.000	650.000	4.000	0.735
5	0.000	648.000	5.000	-0.132
6	0.000	648.000	6.000	0.561
7	0.000	665.000	7.000	-0.360
8	0.000	661.000	8.000	0.735
9	0.000	675.000	9.000	-0.360
10	0.000	664.000	10.000	0.913

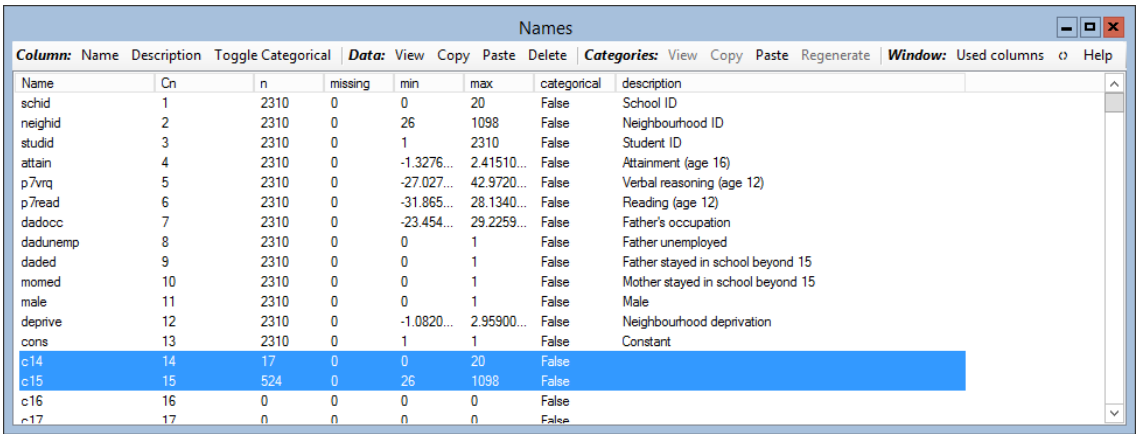
We see, for example, that student 1 attended school 0, resided in neighbourhood 675 and scored 0.74 in their national examinations. Note that ID variables are typically defined as consecutive integers starting at a value of one and so the 0 value in the above output appears somewhat peculiar. While this is how we received the data, there is nothing to stop us recoding the variable along more conventional lines.

Next, we use the **Command interface** window to confirm that the number of schools and neighbourhoods in the data are 17 and 524, respectively. Specifically, we use the **UNIQ** command to generate new 'short' versions of the school and neighbourhood identifier variables which take one record per group.

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

```
UNIQ 'schid' c14
UNIQ 'neighid' c15
```

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



Name	Cn	n	missing	min	max	categorical	description
schid	1	2310	0	0	20	False	School ID
neighid	2	2310	0	26	1098	False	Neighbourhood ID
studid	3	2310	0	1	2310	False	Student ID
attain	4	2310	0	-1.3276...	2.41510...	False	Attainment (age 16)
p7vrq	5	2310	0	-27.027...	42.9720...	False	Verbal reasoning (age 12)
p7read	6	2310	0	-31.865...	28.1340...	False	Reading (age 12)
dadocc	7	2310	0	-23.454...	29.2259...	False	Father's occupation
dadunemp	8	2310	0	0	1	False	Father unemployed
daded	9	2310	0	0	1	False	Father stayed in school beyond 15
momed	10	2310	0	0	1	False	Mother stayed in school beyond 15
male	11	2310	0	0	1	False	Male
deprive	12	2310	0	-1.0820...	2.95900...	False	Neighbourhood deprivation
cons	13	2310	0	1	1	False	Constant
c14	14	17	0	0	20	False	
c15	15	524	0	26	1098	False	
c16	16	0	0	0	0	False	
c17	17	0	0	0	0	False	

The new variable **c14** now contains a single record for each unique school, whilst the new variable **c15** contains a single record for each unique neighbourhood. The number of records for each of these new variables, 17 and 524, confirms that there are indeed 17 schools and 524 neighbourhoods in the data.

The low number of schools in the data is not ideal for pursuing a multilevel analysis that includes school level random effects. To reliably estimate the between-school variance (and potentially other school level random part parameters), we would ideally have a higher number of units. The high number of neighbourhoods, on the other hand, makes it likely that any neighbourhood level random part parameters that we choose to include in the model will be reliably estimated.

To obtain a first impression of the structure of the cross-classified data, we present the data as a two-way cross-tabulation of neighbourhoods (**neighid**) by schools