

Module 12: Cross-Classified Multilevel Models

Stata Practical

George Leckie
Centre for Multilevel Modelling

Pre-requisites

- Modules 1-5,11

Contents

Introduction to the Scotland Neighbourhood Study.....	3
P12.1 Examining and Describing the Data.....	5
P12.1.1 Exploring the cross-classified data structure	6
P12.1.2 Summarising the response and predictor variables	12
P12.2 A Cross-Classified Model of Attainment.....	14
P12.2.1 Specifying and fitting the cross-classified model	14
P12.2.2 Interpretation of the model output	18
P12.2.3 Calculating coverage intervals, variance partition coefficients (VPCs) and intraclass correlation coefficients (ICCs).....	20
P12.2.4 Predicting and examining school and neighbourhood effects	23
P12.3 An Alternative Parameterisation.....	29
P12.4 Adding a Random School-by-Neighbourhood Interaction Classification	31
P12.5 Adding Predictor Variables	35
P12.5.1 Adding student level predictor variables	35
P12.5.2 Adding neighbourhood level predictor variables	42
P12.5.3 Adding school level predictor variables	45
P12.6 Adding Random Coefficients	46
P12.6.1 Adding school and neighbourhood level random coefficients	46
P12.6.2 Adding cross-level interactions	49
Further Reading.....	50
References	52

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

Leckie, G. (2013). Cross-Classified Multilevel Models - Stata Practical. LEMMA VLE Module 12, 1-52.
<http://www.bristol.ac.uk/cmm/learning/course.html>

Address for correspondence:

George Leckie
Centre for Multilevel Modelling
University of Bristol
2 Priory Road
Bristol, BS8 1TX
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), as an exercise dataset in the excellent *Multilevel and Longitudinal Modelling Using Stata* book (Rabe-Hesketh and Skrondal, 2012a), and as the illustrative application in the *Cross-Classified Multilevel Models Using Stata* book chapter by Leckie (2012).

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).¹ 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 closely follow Leckie (2012) and in doing so we will also replicate many of the results presented for these data by Raudenbush (1993), Raudenbush and Bryk (2002) and Rabe-Hesketh and Skrondal, 2012a.

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.

¹ 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.

P12.1 Examining and Describing the Data

Load '12.1.dta' into memory and open the do-file for this lesson

From within the LEMMA learning environment

- Go to **Module 12: Cross-Classified Multilevel Models**, and scroll down to **Stata files**
- Click '**12.1.dta**' to open the dataset

and use the `describe` command to produce a summary of the dataset

```
. describe
```

Contains data from 12.1.dta

obs:	2,310			
vars:	12		29 Aug 2011 16:18	
size:	85,470	(99.9% of memory free)		

variable name	storage type	display format	value label	variable label
schid	byte	%8.0g		School ID
neighid	int	%8.0g		Neighbourhood ID
studid	int	%9.0g		Student ID
attain	float	%3.2f		Attainment (age 16)
p7vrq	float	%3.2f		Verbal reasoning (age 12)
p7read	float	%3.2f		Reading (age 12)
dadocc	float	%3.2f		Father's occupation
dadunemp	byte	%8.0g		Father unemployed
daded	byte	%8.0g		Father stayed in school beyond 15
momed	byte	%8.0g		Mother stayed in school beyond 15
male	byte	%8.0g		Male
deprive	float	%3.2f		Neighbourhood deprivation

The data consist of 2,310 observations on 12 variables and each variable has been given a variable label.

Standard univariate summary statistics can be requested using the `summarize` command

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
schid	2310	10.00866	6.269943	0	20
neighid	2310	495.3398	267.4553	26	1098
studid	2310	1155.5	666.9839	1	2310
attain	2310	.0933957	1.002091	-1.3276	2.4151
p7vrq	2310	.5057663	10.64822	-27.028	42.972
p7read	2310	-.0443549	13.88751	-31.866	28.134
dadocc	2310	-.464169	11.78157	-23.454	29.226
dadunemp	2310	.1090909	.3118207	0	1
daded	2310	.2151515	.4110164	0	1
momed	2310	.2484848	.432228	0	1
male	2310	.4800866	.4997115	0	1
deprive	2310	.0216658	.6218811	-1.082	2.959

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 using the `list` command to list the data on the school (**schid**), neighbourhood (**neighid**) and student (**studid**) identifier variables and for the attainment score response variable (**attain**) for the first 10 students in the data.

```
. list schid neighid studid attain in 1/10
```

	schid	neighid	studid	attain
1.	0	675	1	0.74
2.	0	647	2	0.26
3.	0	650	3	-1.33
4.	0	650	4	0.74
5.	0	648	5	-0.13
6.	0	648	6	0.56
7.	0	665	7	-0.36
8.	0	661	8	0.74
9.	0	675	9	-0.36
10.	0	664	10	0.91

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.

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.