PSY9185 - Multilevel models

Cross-classified data

Espen Moen Eilertsen

2023-05-04

Hierarchical data

The multilevel models discussed so far have dealt with hierarchical data structures where units are classified by some factor into higher level clusters (2-level models), which may again be classified by some other factor into higher level clusters (3-level models or more)

- `Students` (level-1) nested in `classes` (level-2) which are again nested in `schools` (level-3)

The classification factors (`student`, `class`, `school`) are **nested** because a lower-level cluster can only belong to one higher level cluster

- All `students` in the same `class` also go to the same `school`

Hierarchical data

student	class	school
1	1	1
2	1	1
3	2	1
4	2	1
5	3	2
6	3	2
7	4	2
8	4	2

- A `student` can only belong to one `class` which can only belong to one `school`

Cross-classified data

When units are classified by multiple grouping factors that cannot be arranged into hierarchies, the data is instead **cross-classified**

- Students cross-classified by primary and secondary school
- Patients cross-classified by doctors and nurses
- Reaction times cross-classified by experimental condition and subject
- Children cross-classified by mothers and fathers

In designed experiments the factors `condition` and `subject` can often be **fully crossed** as every subject receive all conditions, whereas the factors `doctor` and `nurse` can often be **partially crossed** as not every every patient see all doctors and all nurses

Observational designs will often have partially crossed data, whereas experimental designs will often have fully crossed data

A split-plot design

All **subjects** (1-4) received both levels (A, B) of experimental factor **F1** (within-subjects factor). Half of the subjects received level A of experimental factor **F2**, the other half received level B (between-subjects factor)

F1	F2	subject
A	A	1
B	A	1
A	A	2
B	A	2
A	B	3
B	B	3
A	B	4
B	B	4

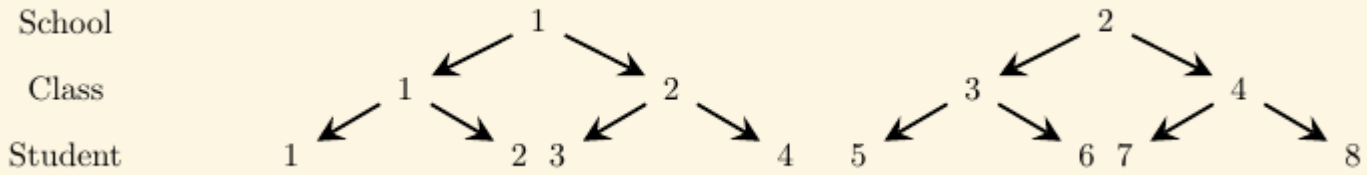
```
   F2
F1  A B
  A 2 2
  B 2 2
```

```
   subject
F1  1 2 3 4
  A 1 1 1 1
  B 1 1 1 1
```

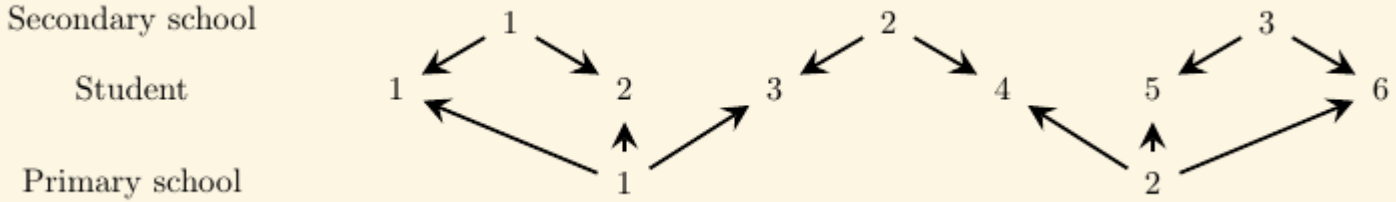
```
   subject
F2  1 2 3 4
  A 2 2 0 0
  B 0 0 2 2
```

Cross-classified data

Nested data structure



Cross-classified data structure



Students from the same primary school can go to different secondary schools and students from the same secondary school can come from different primary schools

Exercises 1

https://github.com/espenmei/PSY9185

Open the file `cross_classified/day3_cross_classified_exercises.R`. We will work with data from the MLwiN program that consist of test scores from students at age 16. Each row represents a student

- `attain` - exam attainment score
- `pid` - primary school identifier (age 5 - 12)
- `sid` - secondary school identifier (age 12 - 16)

1. How many students, primary schools and secondary schools are in the data? (hint: `length(unique(xc1$pid))`)
2. Cross-tabulate the grouping factors primary and secondary school. What does the table show us? (hint: `xtabs(~pid + sid, xc1)`)
3. Can we determine from the table if primary and secondary school are partially or fully crossed?

Multilevel models for cross-classified data

If the classification factors may contribute to the outcome that is under study, multilevel models can be used to model those effects - both primary and secondary school may influence educational achievement

Multilevel model for achievements y_{ijk} for student i from secondary school j and primary school k

$$y_{ijk} = \beta + \eta_{1j} + \eta_{2k} + \epsilon_{ijk}$$

- β is a fixed intercept
- η_{1j} is a random intercept for secondary school j
- η_{2k} is a random intercept for primary school k
- ϵ_{ijk} is the residual deviation for each student

Multilevel models for cross-classified data

Similar to other multilevel models, individual covariates can be added, possibly with random coefficients varying over primary and/or secondary school. The random coefficients can be explained by primary and/or secondary school variables

Random intercepts and slopes for secondary school and random intercepts for primary school

$$y_{ijk} = \beta_1 + \beta_2 x_{ijk} + \eta_{1j} + \eta_{2j} x_{ijk} + \eta_{3k} + \epsilon_{ijk}$$

Random intercepts for primary and secondary school with covariate for primary school

$$y_{ijk} = \beta_1 + \beta_2 x_{k} + \eta_{1j} + \eta_{2k} + \epsilon_{ijk}$$

lme4 with crossed random effects

Some estimation methods/software packages are restricted to models with nested random effects

With **lme4**, crossed random effects are specified the same way as nested random effects

```
lmer(y ~ 1 + (1|factor1) + (1|factor2))
```

lme4 doesn't need to know whether factors are crossed or nested - that is a property of the data

Warning

Don't code data like this

score	student	school
0.5	1	1
-0.5	1	1
0.4	2	1
-0.2	2	1
-2.0	1	2
1.8	1	2
-0.3	2	2
-0.2	2	2

if what you mean is this

score	student	school
0.5	1	1
-0.5	1	1
0.4	2	1
-0.2	2	1
-2.0	3	2
1.8	3	2
-0.3	4	2
-0.2	4	2

- `lmer(score ~ 1 + (1|student) + (1|school))` will treat student and school as crossed in first case and `lmer(score ~ 1 + (1|student:school) + (1|school))` is necessary. Both give nested model in second case
- Not obvious in large datasets - worth checking carefully

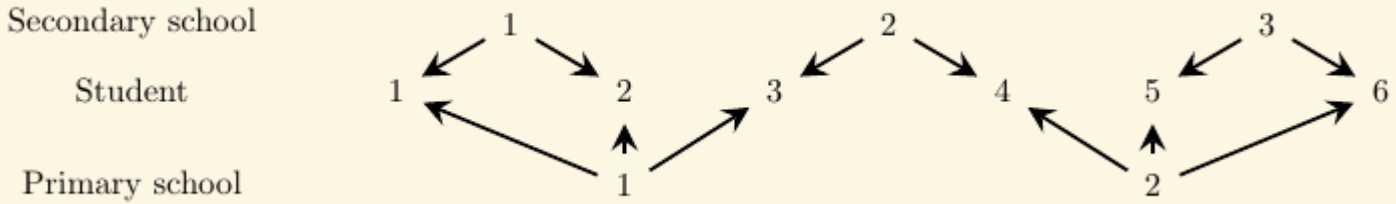
Exercises 2

Continue working with the file `day3_cross_classified_template.R` with test scores from students at age 16

- `attain` - educational attainment score
- `pid` - primary school identifier (age 5 - 12)
- `sid` - secondary school identifier (age 12 - 16)

1. Fit a model for educational attainment with a fixed intercept and random intercepts for secondary school
    - Interpret the parameters of the model
    - Try to interpret what these "random school effects" represent
2. Extend the model to have random intercepts for both primary and secondary school
    - Interpret the parameters of the model
    - Does primary or secondary school appear to be most important for attainment?

Dependence among responses



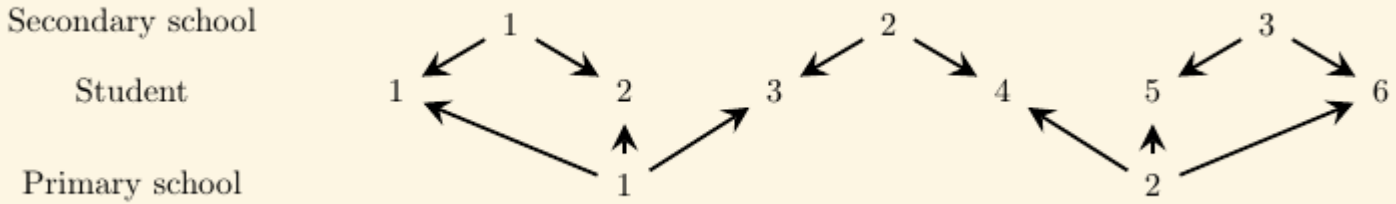
General formulation $y_{ijk} = \beta + \eta_{1j} + \eta_{2k} + \epsilon_{ijk}$

student 1 $y_{111} = \beta + \eta_{11} + \eta_{21} + \epsilon_{111}$

student 2 $y_{211} = \beta + \eta_{11} + \eta_{21} + \epsilon_{211}$

student 3 $y_{321} = \beta + \eta_{12} + \eta_{21} + \epsilon_{321}$

Dependence among responses



Denote the variance of the random intercepts for secondary school with ψ_1, the variance for the random intercepts for primary school with ψ_2, and the residual variance with θ. The total variance in test scores is then

$$\text{Var}(y_{ijk}) = \psi_1 + \psi_2 + \theta$$

Covariance between student 1 and 2

$$\text{Cov}(y_{111}, y_{211}) = \psi_1 + \psi_2$$

Covariance between student 1 and 3

$$\text{Cov}(y_{111}, y_{321}) = \psi_2$$

Intraclass correlations

Intraclass correlations measures degree of dependence between units because they belong to the same group

The correlation between students attending the same primary but different secondary school is

$$\frac{\psi_2}{\psi_1 + \psi_2 + \theta}$$

Students attending the same secondary but different primary school

$$\frac{\psi_1}{\psi_1 + \psi_2 + \theta}$$

Students attending the same primary *and* secondary school

$$\frac{\psi_1 + \psi_2}{\psi_1 + \psi_2 + \theta}$$

Random effects are residuals

If we included for example a student level covariate x_{ijk} in the school model

$$y_{ijk} = \beta_1 + \beta_2 x_{ijk} + \eta_{1j} + \eta_{2k} + \epsilon_{ijk},$$

then η_{1j} and η_{2k} measure differences in test scores between secondary and primary schools that can **not** be explained by x_{ijk}. In other words, ψ_1 and ψ_2 measure the variance in test scores between schools when we compare students with the same level of the covariate (for example parent involvement)

Therefore, intraclass correlations are also conditional on covariates and should be interpreted as **residual** intraclass correlations in models with covariates

Exercises 3

Continue working with the file `day3_cross_classified_template.R` with test scores from students at age 16

- `attain` - educational attainment score
- `pid` - primary school identifier (age 5 - 12)
- `sid` - secondary school identifier (age 12 - 16)

Use the model from exercise 2

1. Compute the intraclass correlation for students from the same primary but different secondary school
2. Compute the intraclass correlation for students from the same secondary school but different primary school
3. Compute the intraclass correlation for students from the same primary *and* secondary school

Random interactions

In cross-classified designs, we can model random interaction effects among factors

- The effect of going to different secondary schools differ depending on which primary schools the student went to
- The effect of going to different primary schools differ depending on which secondary schools the student went to

Multilevel model for achievements y_{ijk} for student i from secondary school j and primary school k

$$y_{ijk} = \beta + \eta_{1j} + \eta_{2k} + \eta_{3jk} + \epsilon_{ijk}$$

η_{3jk} takes different values for each combination of primary and secondary school, and therefore allows deviations from their main effects - *there is something about the particular combination of primary and secondary schools*

- For example some secondary schools may be more beneficial when paired with students who attended a primary school with similar teaching practices

Random interactions

Crosstabulation for 6 first secondary schools (rows) and 6 first primary schools (columns)

```
# 6 x 6 sparse Matrix of class "dgCMatrix"
#    pid
# sid 1 2 3 4  5  6
#   1 8 . . . 53  1
#   2 . . . .  .  .
#   3 . . . .  .  1
#   4 . . . .  .  .
#   5 . . 3 .  . 52
#   6 . . . 1  .  .
```

- The additive model fits a random main effect for each row η_{1j} and each column η_{2k}
- The interaction model additionally fits a random effect for each non-zero cell η_{3jk} - allowing each combination of primary/secondary school to deviate from the main effects

Intraclass correlations

Denote the variance of the random intercepts for secondary school with ψ_1, the variance for the random intercepts for primary school with ψ_2, the variance of the interaction with ψ_3 and the residual variance with θ

The correlation between students attending the same primary but different secondary school is

$$\frac{\psi_2}{\psi_1 + \psi_2 + \psi_3 + \theta}$$

Students attending the same secondary but different primary school

$$\frac{\psi_1}{\psi_1 + \psi_2 + \psi_3 + \theta}$$

Students attending the same primary *and* secondary school

$$\frac{\psi_1 + \psi_2 + \psi_3}{\psi_1 + \psi_2 + \psi_3 + \theta}$$

Random interaction in lme4

$$y_{ijk} = \beta + \eta_{1j} + \eta_{2k} + \eta_{3jk} + \epsilon_{ijk}$$

primary	secondary	prim_sec	student
1	1	1	1
1	1	1	2
1	2	2	3
2	2	3	4
2	3	4	5
2	3	5	6

```
lmer(score ~ 1 + (1|secondary)
+ (1|primary) +
(1|primary:secondary))
```

```
lmer(score ~ 1 + (1|secondary)
+ (1|primary) + (1|prim_sec))
```

Require >1 student for combinations of primary and secondary school or else η_{3jk} is confounded with the residuals ϵ_{ijk}

Exercises 4

Continue working with the file `day3_cross_classified_template.R` with test scores from students at age 16

- `attain` - educational attainment score
- `pid` - primary school identifier (age 5 - 12)
- `sid` - secondary school identifier (age 12 - 16)

1. Make a new variable `pid_sid` that codes the combination of primary and secondary school for each student
    - How many levels does this variable have?
    - How many students are there in average for each combination of primary and secondary school?
    - What is the standard deviation?
2. Expand the main effects model from previous exercises to include interaction between primary and secondary school
3. Perform a likelihood ratio test to evaluate the null-hypothesis that there is no interaction

Cross-classified data

Grouping factor **A** is *nested* within grouping factor **B** if each *level* of **A** occurs within only one *level* of **B** - if not, they are *crossed*

If all *levels* of grouping factor **A** occurs within all *levels* of grouping factor **B**, the factors are *fully* crossed, if not, they are *partially* crossed

Nested		Fully crossed		Partially crossed	
A	B	A	B	A	B
1	1	1	1	1	1
2	1	2	1	2	1
3	1	3	1	3	1
4	2	1	2	1	2
5	2	2	2	2	2
6	2	3	2	1	3

References

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*. Springer New York.

Goldstein, H. (2011). *Multilevel statistical models*. Vol. 922. John Wiley & Sons.

Rabe-Hesketh, S. and A. Skrondal (2008). *Multilevel and longitudinal modeling using Stata*. STATA press.

Jokes and raters

Go to https://github.com/espenmei/PSY9185

copy the code in `jokes/jokes.R` to an R script on your computer and run it