intro to multilevel modeling

April 17, 2024

1

What is this about?

- Imagine we are interested in the extent to which a pre-test (X; mean-centered!) predicts standardized math test scores (Y) in 5th graders.
- We collect data from one classroom and find:

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$
$$\hat{Y}_i = 70 + 0.2X_i$$

2

A complication

- Imagine that we collected more data for a second classroom and found this:

$$\hat{Y}_i = 60 + 0.2X_i$$

- Different intercept (maybe the class has a different overall level of ability)

3

What should we do?

- Three options, from least to most complex:
  1) Combine the data across classes and ignore that they come from different classes
  2) Acknowledge that the data come from different classes and include classrooms as a part of our regression model
  3) Multilevel modeling

4

Option 1

- Collapsing across classes
- This gives us:

$$\hat{Y}_i = 65 + 0.2X_i$$

5

Option 2

- Modeling the classroom, too
- Using a dummy-code (classroom 1 = 0)
- This gives us

$$\hat{Y}_i = \beta_0 + \beta_1 X_{pretest,i} + \beta_2 X_{class\ i}$$
$$\hat{Y}_i = 70 + 0.2X_{pretest,i} + (-10)X_{class,i}$$

6

Option 3

- Modeling not only the effect of the pretest at the subject level
- Also modeling the differences in classrooms

$$\hat{Y}_i = \hat{\beta}_{0,j} + \beta_1 X_{pretest,i} + e_i$$
$$\hat{\beta}_{0,j} = \gamma_0 + u_j$$

7

Option 3

- Modeling not only the effect of the pretest at the subject level
- Also modeling the differences in classrooms

$$\hat{Y}_i = \hat{\beta}_{0,j} + \beta_1 X_{pretest,i} + e_i$$
$$\hat{\beta}_{0,j} = \gamma_0 + u_j$$

- This is called a random-intercept model, and can be presented as one equation

$$\hat{Y}_i = [\gamma_0 + \beta_1 X_{pretest,i}] + [u_j + e_i]$$

8

What's up with the names?

- You'll hear many names for the same (or similar analyses)
  - linear mixed effects models; mixed linear models; linear mixed models
  - hierarchical linear modeling (HLM)
  - general linear mixed model
  - mixed models
  - nested growth curves
  - random effects modeling
  - random coefficient modeling
  - covariance components models

9

What's "mixed" about these models?

- They include a mix of fixed-effects and random-effects variables
- Fixed-effects variables
  - non-randomly selected
  - no desire to generalize to other levels
  - repeatable
  - get slope estimates
- Random-effects variables
  - randomly selected
  - wish to generalize to other levels
  - not repeatable
  - get variance estimates

10

What is multilevel data?

- Data that are somehow grouped in a way that leads to non-independent observations
  - That is, residuals at a/some low level(s) are correlated
- Some examples:
  - in educational research, students are nested within classrooms (& schools, districts, etc.)
  - in political science, legislators are nested within parties (& states, houses of Congress)
  - in public-health policy research, respondents are nested within cities, counties, etc.

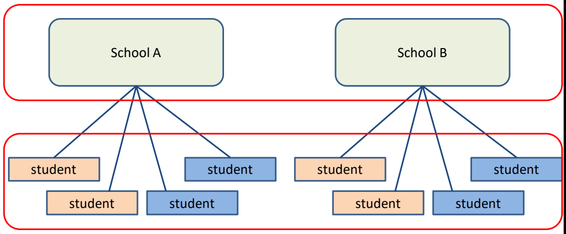
11

What is multilevel data?

- More examples:
  - A clinical psychology student here did a dissertation examining the client-therapist alliance, and got data from many clients who shared therapists
  - In repeated-measures designs, multiple observations are made from the same person
  - In dyad-based research, the two subjects in the dyad provide related observations

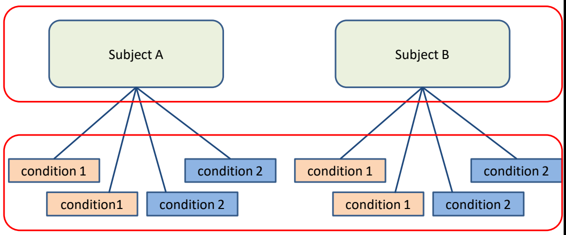
12

Some diagrams of multilevel data: students (within conditions) within schools



13

Some diagrams of multilevel data: observations (within conditions) within subjects



14

Why does this matter?

- regression (ANOVA *is* regression) assumes that residuals (unexplained influences on scores) are independent
- if they aren't, this can inflate the Type 1 or Type 2 error rate
- theoretically, ignoring the ways in which data are grouped ignores that context (however it's defined) matters
- practically, some of the higher-level grouping variables may be of interest themselves
- ecological fallacy (high-level relationships may mis-estimate low-level relationships)
- atomistic fallacy (low-level relationships may not scale up)

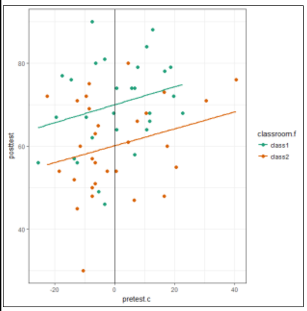
15

yet more introduction to MLM

16

exploring the twoclassrooms data

 # Section 1



- things to notice:
  - classes have similar slopes
  - class1 has an intercept of 70
  - class2 has an intercept of 60
- think of intercepts as DV means (when predictors are centered)

17

separate regressions

 # Section 2

- class 1 regression equation (rounded)

$$\hat{Y} = 70 + 0.2 * pretest$$

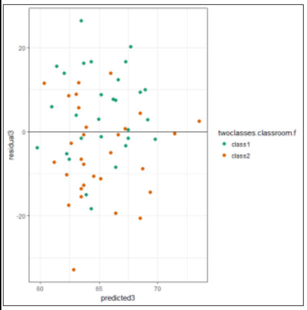
- class 2 regression equation (rounded)

$$\hat{Y} = 60 + 0.2 * pretest$$

18

analysis option: collapsing across classrooms

 # Section 3

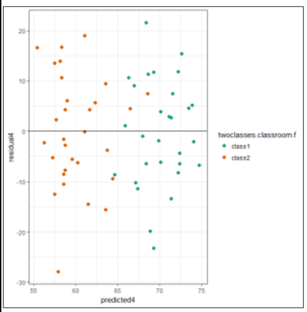


- R^2 = .055; s_{resid} = 11.9
  $\hat{Y} = 65 + 0.2 * pretest$
- intercept between 60 & 70
- pro: n = 60 better than n = 30
- con: independence assumption is in bad shape
  - residuals are clustered (class1 mostly > 0, class2 mostly < 0)
  - residuals M_{class1} = 4.9
  - residuals M_{class2} = -4.9

19

analysis option: modeling classrooms as a predictor

 # Section 4



- R^2 = .22; s_{resid} = 10.9
  $\hat{Y} = 70 + 0.2pre + -10class$
- intercept represents class1
- class slope represents class2 difference from class1
- pro: independence of residuals is in good shape
  - residuals not clustered
  - residuals M_{class1} = 0
  - residuals M_{class2} = 0
- con: scaling up (e.g., to 30 classes) costs parameters (which we probably don't care about)

20

How to run a multilevel model using `lmer()` in the lme4 package

```
lmer(data = twoclasses,
   posttest ~ (1 | classroom.d) + pretest.c )
```

21

How to run a multilevel model using `lmer()` in the lme4 package

```
lmer(data = twoclasses,
   posttest ~ (1 | classroom.d) + pretest.c )
```

22

How to run a multilevel model using `lmer()` in the lme4 package

```
lmer(data = twoclasses,
   posttest ~ (1 | classroom.d) + pretest.c )
```

- This indicates that the intercept (1) varies from classroom to classroom

23

How to run a multilevel model using `lmer()` in the lme4 package

```
lmer(data = twoclasses,
   posttest ~ (1 | classroom.d) + pretest.c )
```

24

Parsing part of the `summary()` of an `lmer()` call

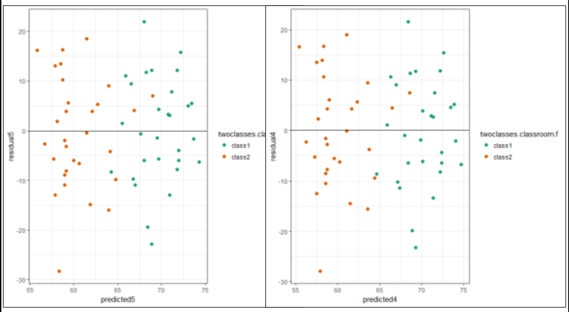
```
Random effects:
 Groups      Name        Variance Std.Dev.
 classroom.d (Intercept)  44.72    6.688
 Residual                118.52   10.887
Number of obs: 60, groups:  classroom.d, 2

Fixed effects:
            Estimate Std. Error t value
(Intercept)  65.0333     4.9333  13.182
pretest.c     0.2089     0.1036   2.017
```

25

Analysis option: multilevel modeling

 # Section 5



26