Multilevel linear models

Alexander McLain

1 High School and Beyond Survey

The data file is a subsample from the 1982 High School and Beyond Survey and is used extensively in Hierarchical Linear Models by Raudenbush and Bryk. The data file consists of 7185 students nested in 160 schools. The outcome variable of interest is student-level math achievement score (MATHACH). Variable SES is social-economic-status of a student and therefore is a student-level variable. Variable MEANSES is the group mean of SES and therefore is a school-level variable. Both SES and MEANSES are centered at the grand mean (they both have means of 0). Variable SECTOR is an indicator variable indicating if a school is public or catholic and is therefore a school-level variable. There are 90 public schools (SECTOR=0) and 70 catholic schools (SECTOR=1) in the sample.

```
library(sas7bdat)
library(tidyverse)
library(lme4)
library(lmerTest)
Hsb12 <- read.sas7bdat("Hsb12.sas7bdat")
glimpse(Hsb12)
```

```
## Rows: 7,185
## Columns: 11
## $ SCHOOL   <chr> "1224", "1224", "1224", "1224", "1224", "1224", "1224", "1...
## $ MINORITY <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0...
## $ FEMALE   <dbl> 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0...
## $ SES      <dbl> -1.528, -0.588, -0.528, -0.668, -0.158, 0.022, -0.618, -0....
## $ MATHACH  <dbl> 5.876, 19.708, 20.349, 8.781, 17.898, 4.583, -2.832, 0.523...
## $ SIZE     <dbl> 842, 842, 842, 842, 842, 842, 842, 842, 842, 842, 842, 842...
## $ SECTOR   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ PRACAD   <dbl> 0.35, 0.35, 0.35, 0.35, 0.35, 0.35, 0.35, 0.35, 0.35, 0.35...
## $ DISCLIM  <dbl> 1.597, 1.597, 1.597, 1.597, 1.597, 1.597, 1.597, 1.597, 1....
## $ HIMINTY  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ MEANSES  <dbl> -0.428, -0.428, -0.428, -0.428, -0.428, -0.428, -0.428, -0...
```

```
head(Hsb12[ , 1:9], 8)
```

SCHOOL	MINORITY	FEMALE	SES	MATHACH	SIZE	SECTOR	PRACAD	DISCLIM
1224	0	1	-1.528	5.876	842	0	0.35	1.597
1224	0	1	-0.588	19.708	842	0	0.35	1.597
1224	0	0	-0.528	20.349	842	0	0.35	1.597
1224	0	0	-0.668	8.781	842	0	0.35	1.597
1224	0	0	-0.158	17.898	842	0	0.35	1.597
1224	0	0	0.022	4.583	842	0	0.35	1.597
1224	0	1	-0.618	-2.832	842	0	0.35	1.597
1224	0	0	-0.998	0.523	842	0	0.35	1.597

1.1 Unconditional Means Model

This model is referred as a one-way ANOVA with random effects and is the simplest possible random effect linear model and is discussed in detail by Raudenbush and Bryk. The motivation for this model is the question on how much schools vary in their mean mathematics achievement. In terms of regression equations, we have the following, where $e_ij \sim N(0, \sigma^2)$ and b_{0j} $N(0, \tau^2)$,

```
formu <- MATHACH ~ (1|SCHOOL)
Mod1 <- lmer( formu, data = Hsb12)
summary(Mod1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: formu
##    Data: Hsb12
##
## REML criterion at convergence: 47116.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0631 -0.7539  0.0267  0.7606  2.7426
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  SCHOOL   (Intercept)  8.614   2.935
##  Residual             39.148   6.257
## Number of obs: 7185, groups:  SCHOOL, 160
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)  12.6370     0.2444 156.6473   51.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.2 Including Effects of School Level (level 2) Predictors — predicting mathach from meanses

This model is referred as regression with Means-as-Outcomes by Raudenbush and Bryk. The motivation of this model is the question on if the schools with high MEANSES also have high math achievement. In other words, we want to understand why there is a school difference on mathematics achievement. In terms of regression equations, we have the following.

$$MATHACH_{ij} = \beta_0 + \beta_1 MEANSES_{ij} + b_{0j} + e_{ij}$$

At least, that's how I would write it. It is common in multilevel modeling to write equations using a method some refer to as THNE (the hardest notation ever). Writing the above in THNE looks like this

$$\begin{aligned} MATHACH_{ij} &= \beta_{0j} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} MEANSES_{ij} + b_{0j} \end{aligned}$$

which means exactly the same thing as above, only THNE-ier. (In all seriousness some people find this way to be more clear)

```
formu <- MATHACH ~ MEANSES + (1|SCHOOL)
Mod2 <- lmer( formu, data = Hsb12)
summary(Mod2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: formu
##    Data: Hsb12
##
## REML criterion at convergence: 46961.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.13480 -0.75256  0.02409  0.76773  2.78501
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  SCHOOL   (Intercept)  2.639   1.624
##  Residual             39.157   6.258
## Number of obs: 7185, groups:  SCHOOL, 160
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)  12.6494     0.1493 153.7425   84.74   <2e-16 ***
## MEANSES       5.8635     0.3615 153.4067   16.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr)
## MEANSES -0.004
```

1.3 Including Effects of Student-Level Predictors–predicting mathach from centered student-level ses, cses

This model is referred as a random-coefficient model by Raudenbush and Bryk. Pretend that we run regression of mathach on centered ses on each school, that is we are going to run 160 regressions.

1. What would be the average of the 160 regression equations (both intercept and slope)?
2. How much do the regression equations vary from school to school?
3. What is the correlation between the intercepts and slopes?

These are some of the questions that motivates the following model (in THNE form).

$$\begin{aligned} MATHACH_{ij} &= \beta_{0j} + \beta_{1j}(SES_{ij} - MEANSES_{ij}) + e_{ij} \\ \beta_{0j} &= \gamma_{00} + b_{0j} \\ \beta_{1j} &= \gamma_{10} + b_{1j} \end{aligned}$$

How I would write this model is

$$MATHACH_{ij} = \beta_0 + \beta_1(SES_{ij} - MEANSES_{ij}) + b_{0j} + b_{1j}(SES_{ij} - MEANSES_{ij}) + e_{ij},$$

which is the same.

```
Hsb12 <- Hsb12 %>% mutate( cses = SES - MEANSES)
formu <- MATHACH ~ cses + (1 + cses|SCHOOL)
Mod3 <- lmer( formu, data = Hsb12)
summary(Mod3)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: formu
##    Data: Hsb12
##
## REML criterion at convergence: 46714.2
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.09680 -0.73194  0.01858  0.75388  2.89928
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  SCHOOL   (Intercept)  8.682   2.9465
##           cses         0.694   0.8331   0.02
##  Residual             36.700   6.0581
## Number of obs: 7185, groups:  SCHOOL, 160
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)  12.6493     0.2445 156.7391   51.73   <2e-16 ***
## cses          2.1932     0.1283 155.2180   17.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## cses 0.012
```

1.4 Including Both Level-1 and Level-2 Predictors

Here, we will be predicting mathach from meanses, sector, cses and the cross level interaction of meanses and sector with cses

This model is referred as an intercepts and slopes-as-outcomes model by Raudenbush and Bryk. We have examined the variability of the regression equations across schools. Now we will build an explanatory model to account for the variability. That is we want to model the following: (in THNE form).

$$\begin{aligned} MATHACH_{ij} &= \beta_{0j} + \beta_{1j}(SES_{ij} - MEANSES_j) + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} MEANSES_j + \gamma_{02} SECTOR_j + b_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} MEANSES_j + \gamma_{12} SECTOR_j + b_{1j} \end{aligned}$$

How I would write this model is

$$
\begin{aligned}
MATHACH_{ij} &= \beta_0 + \beta_1(SES_{ij} - MEANSES_{ij}) + \beta_2 MEANSES_j + \beta_3 SECTOR_j + \\
&\quad \beta_4 MEANSES_j(SES_{ij} - MEANSES_{ij}) + \beta_5 SECTOR_j(SES_{ij} - MEANSES_{ij}) + \\
&\quad b_{0j} + b_{1j}(SES_{ij} - MEANSES_{ij}) + e_{ij},
\end{aligned}
$$

which is the same. The questions that we are interested in are:

```
1. Do MEANSES and SECTOR significantly predict the intercept?
2. Do MEANSES and SECTOR significantly predict the within-school slopes?
3. How much variation in the intercepts and the slopes is explained by MEANSES and SECTOR?
```

```
formu <- MATHACH ~ cses + MEANSES + SECTOR + cses*MEANSES + cses*SECTOR + (1 + cses|SCHOOL)
Mod4 <- lmer( formu, data = Hsb12)
summary(Mod4)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: formu
##    Data: Hsb12
##
## REML criterion at convergence: 46503.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.15921 -0.72319  0.01706  0.75439  2.95822
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  SCHOOL   (Intercept)  2.3819  1.5433
##           cses         0.1014  0.3184   0.39
##  Residual             36.7211  6.0598
## Number of obs: 7185, groups:  SCHOOL, 160
##
## Fixed effects:
##              Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)   12.1136     0.1988 159.8921  60.931  < 2e-16 ***
## cses           2.9388     0.1551 139.3043  18.948  < 2e-16 ***
## MEANSES        5.3391     0.3693 150.9689  14.457  < 2e-16 ***
## SECTOR         1.2167     0.3064 149.5994   3.971 0.000111 ***
## cses:MEANSES   1.0389     0.2989 160.5528   3.476 0.000656 ***
## cses:SECTOR   -1.6426     0.2398 143.3450  -6.850 2.01e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) cses   MEANSE SECTOR c:MEAN
## cses         0.080
## MEANSES      0.245  0.020
## SECTOR      -0.697 -0.056 -0.356
## css:MEANSES  0.019  0.282  0.079 -0.028
## cses:SECTOR -0.056 -0.694 -0.029  0.082 -0.351
```

Let's now test if the random "cses" variable is needed. We'll do this by fitting a model that removes that variable, then using the `anova` function.

```
formu <- MATHACH ~ cses + MEANSES + SECTOR + cses*MEANSES + cses*SECTOR + (1|SCHOOL)
Mod5 <- lmer( formu, data = Hsb12)
anova(Mod5,Mod4)
```

```
## refitting model(s) with ML (instead of REML)
```

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Mod5	8	46513.44	46568.48	-23248.72	46497.44	NA	NA	NA
Mod4	10	46516.43	46585.23	-23248.22	46496.43	1.003696	2	0.6054107

It appears that model 4 has a better fit. Let's look at the results of this model.

```
summary(Mod5)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: formu
##    Data: Hsb12
##
## REML criterion at convergence: 46504.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1701 -0.7249  0.0149  0.7543  2.9655
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  SCHOOL   (Intercept)  2.375   1.541
##  Residual             36.766   6.064
## Number of obs: 7185, groups:  SCHOOL, 160
##
## Fixed effects:
##               Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)    12.1138     0.1986  160.0306  60.981  < 2e-16 ***
## cses            2.9358     0.1507 7018.2624  19.481  < 2e-16 ***
## MEANSES         5.3429     0.3690  151.0757  14.480  < 2e-16 ***
## SECTOR          1.2146     0.3061  149.7022   3.968 0.000112 ***
## cses:MEANSES    1.0441     0.2910 7018.2625   3.587 0.000336 ***
## cses:SECTOR    -1.6421     0.2331 7018.2624  -7.045 2.04e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) cses   MEANSE SECTOR c:MEAN
## cses         0.005
## MEANSES      0.245  0.001
## SECTOR      -0.697 -0.003 -0.356
## css:MEANSES  0.001  0.284  0.005 -0.002
## cses:SECTOR -0.003 -0.694 -0.002  0.005 -0.351
```

Now let's take a further look at the results. To do this, we're going to look at the predicted values at "low" and "high" meanses schools. To define "low" and "high" we'll use the first and third quartile of the meansas data.

```
summary(Hsb12$MEANSES)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.188	-0.317	0.038	0.0061385	0.333	0.831

```
summary(Hsb12$cses)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.657	-0.454	0.01	-0.0059951	0.463	2.85

```
# Create dummy SES and SECTOR variables
CSES <- seq( -3.65, 2.85, 0.01)
SECTOR <- c(0,1)
MEANSES <- c(-0.317, 0, 1/3)
# Combine all values of the variables
dummy_data <- expand.grid( CSES = CSES, SECTOR = SECTOR, MEANSES = MEANSES)
# Make sure the data are correct
length(CSES)
```

```
## [1] 651
```

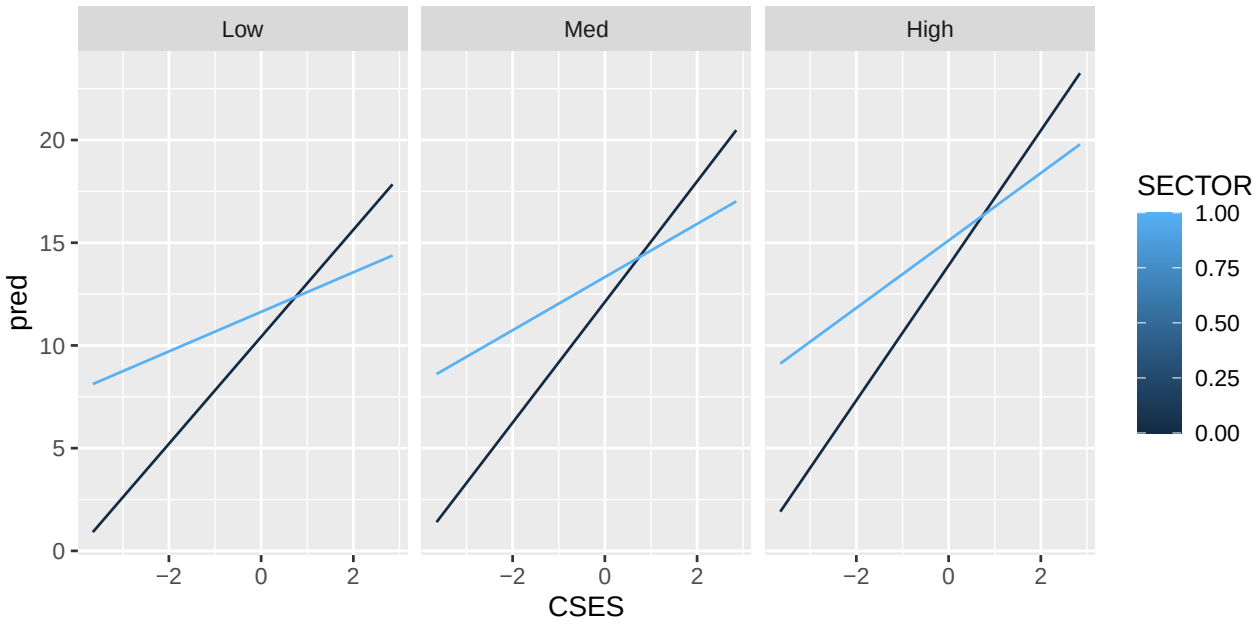
```
dim(dummy_data)
```

```
## [1] 3906    3
```

```
# Create variables for the predicted values
dummy_data <- dummy_data %>% mutate( pred = 12.1138 + 2.9358*CSES +
                                       5.3429*MEANSES + 1.2146*SECTOR  +
                                       1.0441*MEANSES*CSES - 1.6421*SECTOR*CSES) %>%
              mutate(MEANSES_cat = factor( MEANSES, levels = c(-0.317, 0, 1/3),
                                           labels = c("Low", "Med", "High")))
head(dummy_data)
```

CSES	SECTOR	MEANSES	pred	MEANSES_cat
-3.65	0	-0.317	0.9125066	Low
-3.64	0	-0.317	0.9385548	Low
-3.63	0	-0.317	0.9646030	Low
-3.62	0	-0.317	0.9906512	Low
-3.61	0	-0.317	1.0166994	Low
-3.60	0	-0.317	1.0427476	Low

```
p <- ggplot( data = dummy_data, aes(x = CSES, y = pred, group = SECTOR, color = SECTOR))
p + geom_line() + facet_grid(~MEANSES_cat)
```



1.5 Ranking

We can use the results from the final model to rank the schools to see which is doing best in their math achivement scores, after adjusting for the schools mean ses, the students ses and the sector of the school.

```
cbind( ranef(Mod1)$SCHOOL, ranef(Mod5)$SCHOOL)[1:20,]
```

	(Intercept)	(Intercept)
1224	-2.6639348	-0.0719965
1288	0.7394098	0.4525430
1296	-4.5684656	-1.6787527
1308	2.9485171	0.0479320
1317	0.4939461	-1.5249758
1358	-1.2425116	-0.5384325
1374	-2.5023497	-1.5014405
1433	6.2682432	1.7796806
1436	4.9621083	1.2979402
1461	3.6965749	0.7486655
1462	-1.9832816	0.5609992
1477	1.4828017	0.0180119
1499	-4.5835763	-1.5333542
1637	-4.8042049	-0.8522736
1906	3.0819229	-0.0783440
1909	1.5368925	0.6255618
1942	4.7323026	1.5276796
1946	0.2431299	0.5431479
2030	-0.5095143	-1.3553671
2208	2.5728142	-0.1635530