Mirroring HLM7 in R

Lakshmi Lalchandani

12/14/2016

Contents

Purpose

The purpose of this manual is to compare the procedures and analyses for hierarchical linear models from two different statistical software programs: HLM 7 and R. We will be using the High School and Beyond - 1982 data set, (Raudenbush & Bryk, 2002). In R, this data set can be found in the `mlmRev` package. Installing this particular package will also load the required package `lme4`. `lme4` contains the function `lmer` that we will use to run the hierarchical linear models in R. You will also have to install the package `lmerTest` so that we can extract the p-values.

```
library('lmerTest')
library('mlmRev')
```

The data `Hsb82`

`Hsb82` is a data set that comes from a 1982 study. The data frame has 7185 observations of 8 variables:
school an ordered factor designating the school ID
minrty a yes/no factor designting minority status
sx a male/female factor designating sex
ses numeric scores designating the socio-economic status of the student
mAch numeric scores of Math Achievement of the student
meanses numeric scores of the mean socio-economic status for each school

sector a factor designating Public or Catholic school
cses numeric scores where socio-economic scores are group centered by school

Student level variables

In this data set, the student level variables are *school*, *minrty*, *sx*, *ses*, *mAch*, and *cses*.

School level variables

The school level variables are , *school*, *meanses*, and *sector*.

Organizing the data

In R, the data is already organized in the manner needed for `lmer`. That is, there is one row per student, with student nested within school. We need only to add a column of student id's.

```
student <- 1:nrow(Hsb82)
df <- data.frame(student,Hsb82)
```

In HLM 7, you must begin with two separate files: one of the student level variables and one of the school level variables. The following is the first 10 observations of each file.

```
##    school minrty     sx    ses   mAch         ces
## 1    1224     No Female -1.528  5.876 -1.09361702
## 2    1224     No Female -0.588 19.708 -0.15361702
## 3    1224     No   Male -0.528 20.349 -0.09361702
## 4    1224     No   Male -0.668  8.781 -0.23361702
## 5    1224     No   Male -0.158 17.898  0.27638298
## 6    1224     No   Male  0.022  4.583  0.45638298
## 7    1224     No Female -0.618 -2.832 -0.18361702
## 8    1224     No   Male -0.998  0.523 -0.56361702
## 9    1224     No Female -0.888  1.527 -0.45361702
## 10   1224     No   Male -0.458 21.521 -0.02361702
```

```
##      id   sector meanses
## 1  1224   Public  -0.428
## 2  1288   Public   0.128
## 3  1296   Public  -0.420
## 4  1308 Catholic   0.534
## 5  1317 Catholic   0.351
## 6  1358   Public  -0.014
## 7  1374   Public  -0.007
## 8  1433 Catholic   0.718
## 9  1436 Catholic   0.596
## 10 1461   Public   0.683
```

In HLM 7, you will then create an MDM file and an MDMT which will compile the data from both files in two levels.

The models

We will run 4 different types of hierarchical linear models as defined by Raudenbush & Bryk (2002)
1. The unconditional model (also known as the One-Way ANOVA model)
2. Regression with means-as-outcomes

3. Random-coefficient model
4. Intercepts and slopes as outcomes model

I will include both the Mixed model and the Hierarchical model before the analysis. I will also indicate possible research questions that each model can address. However, please note that many of the research questions can be answered only dependent on the centering decision made prior to the analysis.

The unconditional model

In this model, the simplest model, $\gamma_{0,0}$ estimates the weighted grand mean of Y. The estimate for $u_{0,j}$ is the amount that mean varies.

Research questions

1. What is the overall mean math achievement?
2. What proportion of the variance in student math achievement is at the school level?

Equations

Hierarchical Model

Level 1 (student level)

$$mAch_{i,j} = \beta_{0j} + r_{ij}$$

Level 2 (school level)

$$\beta_{0,j} = \gamma_{00} + u_{0j}$$

Mixed Model

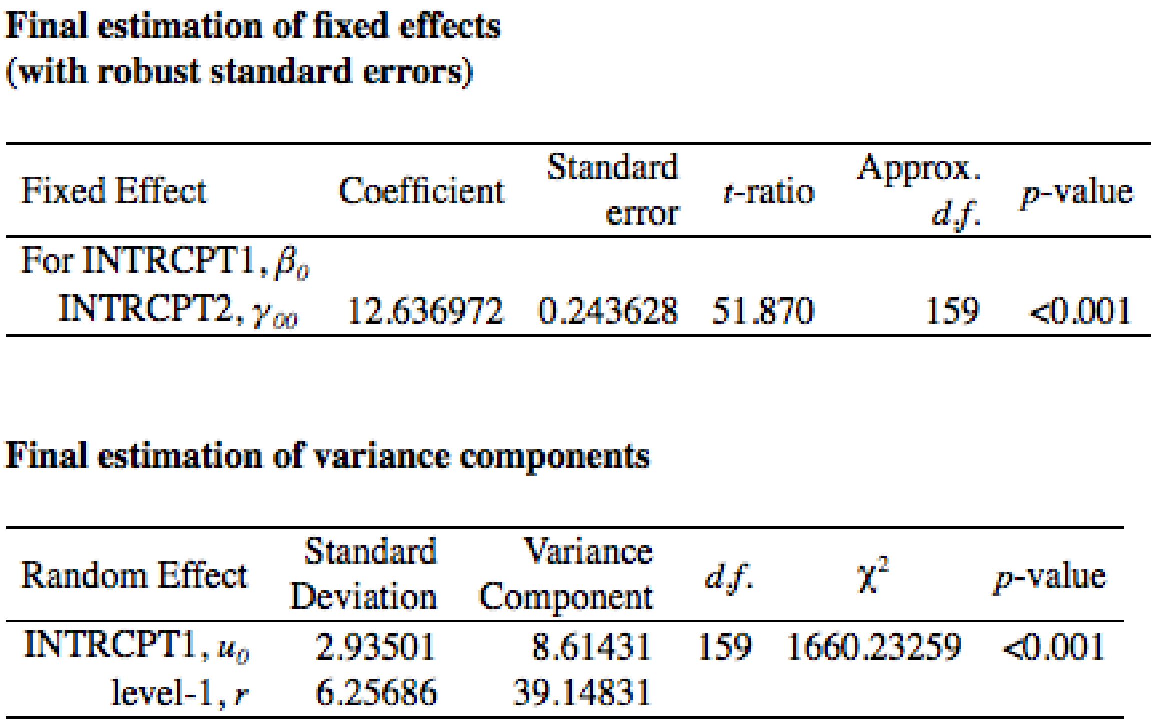
$$mAch_{i,j} = \gamma_{00} + u_{0,j} + r_{ij}$$

Analysis

In HLM 7, in order to run an unconditional model, all that needs to be specified is the outcome variable. In this case, the outcome variable is mAch, then click *Run Analysis*.

```
unconditional <- lmer(mAch ~ 1 +( 1|school), data = df)
summary(unconditional)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: mAch ~ 1 + (1 | school)
##    Data: df
##
## REML criterion at convergence: 47116.8
##
```



Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	12.636972	0.243628	51.870	159	<0.001

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	2.93501	8.61431	159	1660.23259	<0.001
level-1, r	6.25686	39.14831			

Figure 1: The relevant output from HLM 7.

```
## 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
```

The formula in R states that it is predicting the intercept and the random component varies due to school.

For the unconditional model, R will not produce p-values. But there are a couple of hacks to get some information.

```
# To extract the p-values for the estimate of the intercept:

coefs <- data.frame(coef(summary(unconditional)))
coefs$p.value <- 2 * (1-pnorm(abs(coefs$t.value)))
coefs
```

```
##             Estimate Std..Error       df  t.value      Pr...t.. p.value
## (Intercept) 12.63697  0.2443936 156.6473 51.70747 2.344845e-100       0
```

```r
# To extract the chi square for the random effects:

# First we will run a model without a random component.
# Then we will compare the two models

reg.unconditional <- lm(mAch ~ 1, data = df)
anova(unconditional,reg.unconditional)
```

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

```
## Data: df
## Models:
## reg.unconditional: mAch ~ 1
## unconditional: mAch ~ 1 + (1 | school)
##                   Df   AIC   BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## reg.unconditional  2 48104 48117 -24050    48100
## unconditional      3 47122 47142 -23558    47116 983.92      1  < 2.2e-16
##
## reg.unconditional
## unconditional     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimation of the χ^2 statistic does not line up quite as well because we are forced to fit the models using maximum likelihood (ML) as opposed to restricted maximum likelihood (REML). Maximum likelihood can be better for unbalanced data, but can produce biased results. However, the deviace is relatively close.

Answers to research questions

1. What is the overall mean math achievement?

   The estimated fixed effect, γ_{00}, is the weighted estimate of the grand mean math achievement (12.637).

2. What proportion of the variance in student math achievement is at the school level?

   We can use the break down of the variance components to decipher how much variance is explained at each level.

```r
student_var <- 39.148
school_var <- 8.614
(proportion_school_var <- school_var/(school_var+student_var))
```

```
## [1] 0.1803526
```

Approximately 18.04% of the variance in math achievement is at the school level.

Regression with means-as-outcomes

In this type of model, the equation now predicts the mean math achievement for each $school_j$ dependent upon the mean socio-economic status of $school_j$.

Research questions

1. Is there an association between school mean socio-economic status and mean math achievement?
2. How much of the between-school variance in math achievement is explained by mean ses?

Hierarchical Model

Level 1 (student level)

$$mAch_{i,j} = \beta_{0j} + r_{ij}$$

Level 2 (school level)

$$\beta_{0,j} = \gamma_{00} + \gamma_{01}(meanses) + u_{0j}$$

Mixed Model

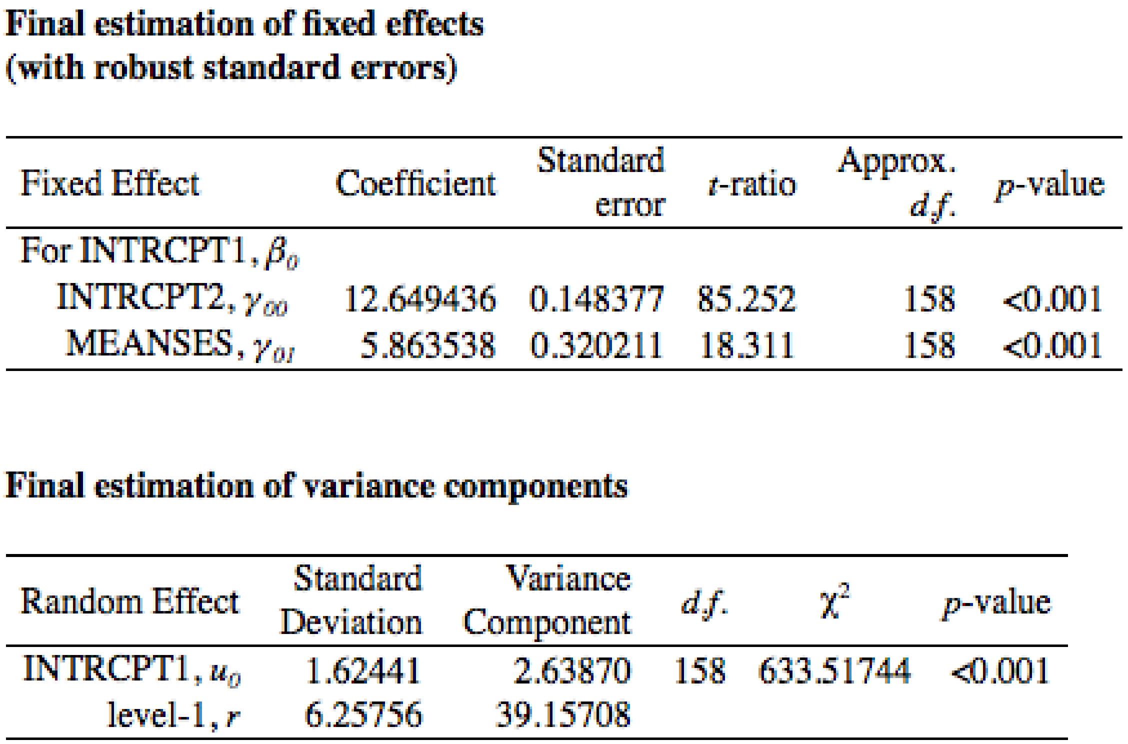
$$mAch_{i,j} = \gamma_{00} + \gamma_{01}(meanses) + u_{0,j} + r_{ij}$$

Analysis

In HLM 7, in order to run this model, all that needs to be specified is the outcome variable at level 1 and the level 2 variable, meanses. *Run Analysis.*

```r
mean_outcome <- lmer(mAch ~ 1 + meanses + (1|school), data = df)
summary(mean_outcome)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: mAch ~ 1 + meanses + (1 | school)
##    Data: df
##
## REML criterion at convergence: 46961.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.13493 -0.75254  0.02413  0.76766  2.78515
##
## 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)  12.6846     0.1493 153.7039   84.97   <2e-16 ***
## meanses       5.8635     0.3615 153.4105   16.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Final estimation of fixed effects
(with robust standard errors)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	12.649436	0.148377	85.252	158	<0.001
MEANSES, γ_{01}	5.863538	0.320211	18.311	158	<0.001

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	1.62441	2.63870	158	633.51744	<0.001
level-1, r	6.25756	39.15708			

Figure 2: The relevant output from HLM 7.

```
##
## Correlation of Fixed Effects:
##         (Intr)
## meanses 0.010
```

```
ranova(mean_outcome)
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## mAch ~ meanses + (1 | school)
##              npar logLik   AIC    LRT Df Pr(>Chisq)
## <none>          4 -23481 46969
## (1 | school)    3 -23601 47207 239.97  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The formula in R states that it is estimating the intercept (mean math achievement) as associated with mean ses and the random component varies due to school.

As long as `lmerTest` is installed, the `summary()` function should provide the p values for the fixed effects (the intercept and slope of mean ses). The function `ranova()` in `lmerTest` will provide the chi square test of the random effects.

Again, there are some expected differences in the output between HLM 7 and R. HLM 7 uses a Bayesian shrinkage estimation for the parameter estimates whereas R uses restricted maximum likelihood. The parameter estimates for the fixed effects will generally be very similar to each other, however they tend to differ on the tests of the random effects, χ^2.

Answers to research questions

1. Is there an association between school mean socio-economic status and mean math achievement? There is a significant association between the school mean ses and the mean math achievement such that schools that have higher mean ses also tend to have higher mean math achievement. As mean ses increases by 1 unit, math achievement is predicted to increase by 5.86 points.
2. How much of the between-school variance in math achievement is explained by mean ses? To answer this question, we can compare the school level variance component from this model to the residual variance in the unconditional model. This will tell us how much of the residual error was reduced by accounting for the mean ses of the school.

```
orig_var <- school_var
meanses_var <- 2.639

(var_explained <- (orig_var-meanses_var)/orig_var)
```

```
## [1] 0.6936383
```

The estimated proportion of the variance in math achievement scores that is accounted for my mean ses is 69.36%.

Random-coefficient model

In this type of model, we will remove the school level predictor and incorporate a student level predictor. We will use ses again, however it is now at the student level. In this model, each school has a unique intercept (mean math achievement) and a unique slope associated with ses.

Research questions

1. Is there evidence that schools vary in their association between math achievement and ses?
2. What is the range of the relationship between ses and mean math achievement expected to be for 95% of the schools?

Hierarchical Model

Given that I am interested the how the mean achievement varies across schools, I have opted to group mean center the ses variable. To do this in HLM 7, when you add ses to the level 1 equation, the drop down menu provides the option to group center. In R, the variable **cses** is group mean centered. Otherwise, you would have to create an additional column of ses scores that are centered around the school mean.

Level 1 (student level)

$$mAch_{i,j} = \beta_{0j} + \beta_{1j}(ses_{ij} - s\bar{e}s_{.j}) + r_{ij}$$

Level 2 (school level)

$$\beta_{0,j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$

Mixed Model

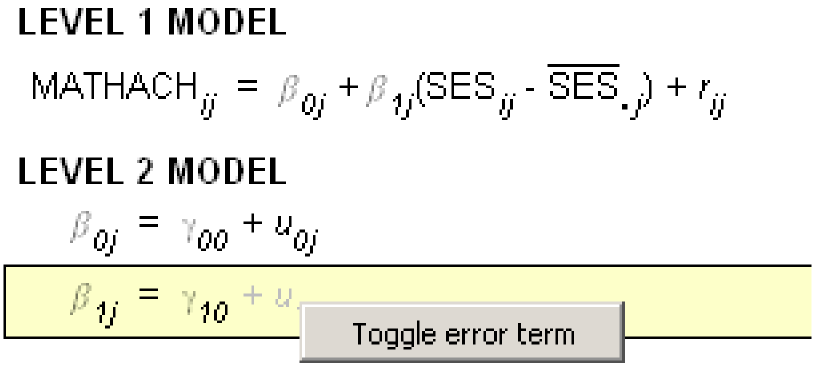
$$mAch_{i,j} = \gamma_{00} + \gamma_{10}(ses_{ij} - s\bar{e}s_{.j}) + u_{0j} + u_{1j} * (ses_{ij} - s\bar{e}s_{.j}) + r_{ij}$$

Analysis

In HLM 7, in order to run this model, all that needs to be specified is the outcome variable and the level 1 predictor (group centered as discussed). In order to run the above specified model, you must also toggle on the variance component for β_{1j}.

```r
randomCoef <- lmer(mAch ~ 1 + cses + (1 + cses|school), data = df)
summary(randomCoef)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: mAch ~ 1 + cses + (1 + cses | school)
##    Data: df
##
## REML criterion at convergence: 46714.2
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.09680 -0.73193  0.01855  0.75386  2.89924
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  school   (Intercept)  8.681   2.9463
```



LEVEL 1 MODEL

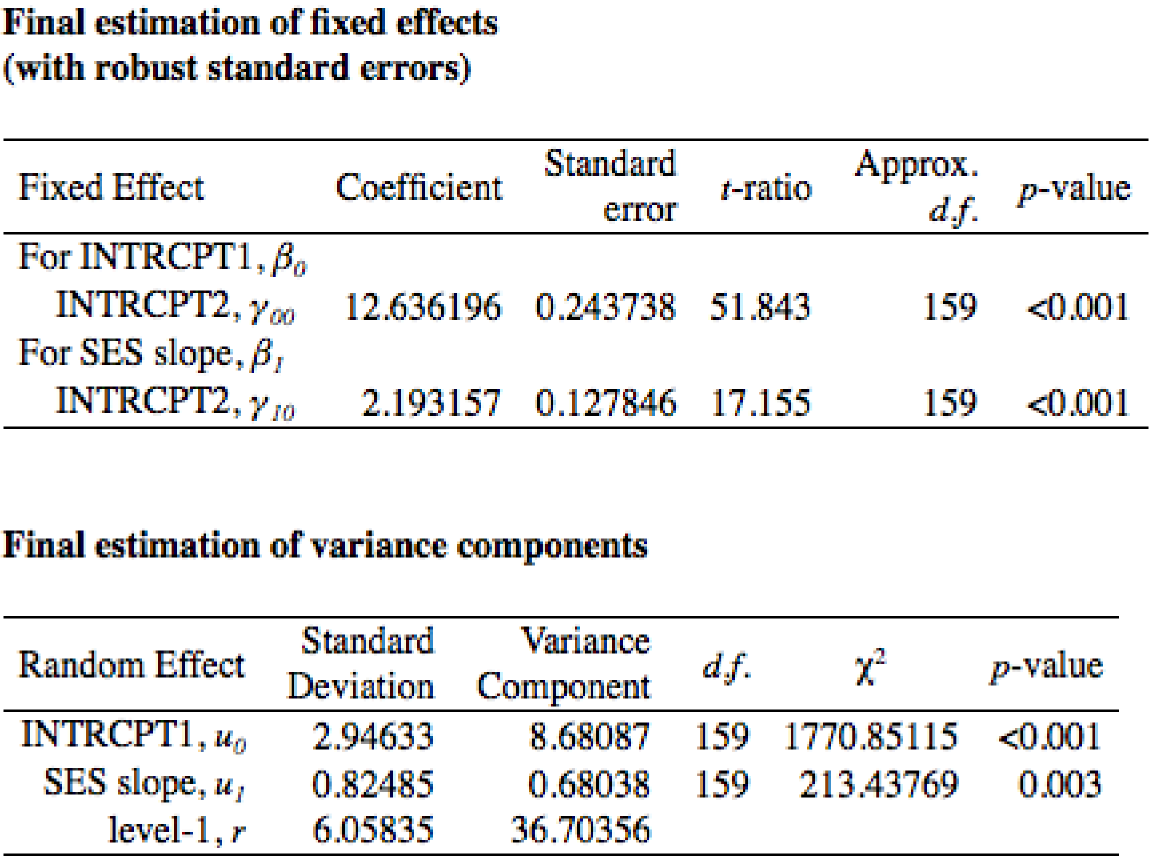
$$\text{MATHACH}_{ij} = \beta_{0j} + \beta_{1j}(\text{SES}_{ij} - \overline{\text{SES}}_{\bullet j}) + r_{ij}$$

LEVEL 2 MODEL

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u$$

Figure 3: Toggle on the variance



Final estimation of fixed effects (with robust standard errors)

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	12.636196	0.243738	51.843	159	<0.001
For SES slope, β_1					
INTRCPT2, γ_{10}	2.193157	0.127846	17.155	159	<0.001

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	2.94633	8.68087	159	1770.85115	<0.001
SES slope, u_1	0.82485	0.68038	159	213.43769	0.003
level-1, r	6.05835	36.70356			

Figure 4: The relevant output from HLM 7.

```
##           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)  12.6362     0.2445 156.7529   51.68   <2e-16 ***
## cses          2.1932     0.1283 155.2167   17.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## cses 0.009
```

```
ranova(randomCoef)
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## mAch ~ cses + (1 + cses | school)
##                              npar logLik   AIC    LRT Df Pr(>Chisq)
## <none>                          6 -23357 46726
## cses in (1 + cses | school)    4 -23362 46732 9.7617  2   0.007591 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If you decided that you did not want to permit the slope to vary, in HLM 7 you would toggle off the error variance. In R, you would constrict the random effect to be `(1|school)` like the previous models.

Answers to research questions

1. Is there evidence that schools vary in their association between math achievement and ses? Yes, there is evidence that schools vary signficantly in their association between math achievement and ses. This is evidenced by the χ^2 test. Again, the χ^2's don't match between HLM 7 and R, but they are both significant, which makes me comfortable with this response.
2. For this answer, we will create a 95% plausible value range around the slope for ses.

$$\gamma_{10} \pm 1.96 * \sqrt{u_i}$$

$$2.19 \pm 1.96 * \sqrt{.694}$$

```
## [1] "0.557 to 3.823"
```

Intercepts and slopes as outcomes model

In this type of model, we will keep the same level 1 model as the previous example and we will add a level 2 predictor as well, sector.

Research questions

1. Does the relationship between student ses and math achievement depend on whether the school is public or Catholic?

Hierarchical Model

I will be using a dummy code for sector (+1 if Catholic, 0 if public). R will automatically create dummy codes (1 being the first alphabetically) unless otherwise specified. For the purpose of this research question, I have left sector uncentered.

Level 1 (student level)

$$mAch_{i,j} = \beta_{0j} + \beta_{1j}(ses_{ij} - s\bar{e}s_{.j}) + r_{ij}$$

Level 2 (school level)

$$\beta_{0,j} = \gamma_{00} + \gamma_{01}(Catholic) + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}(Catholic) + u_{1j}$$

Mixed Model

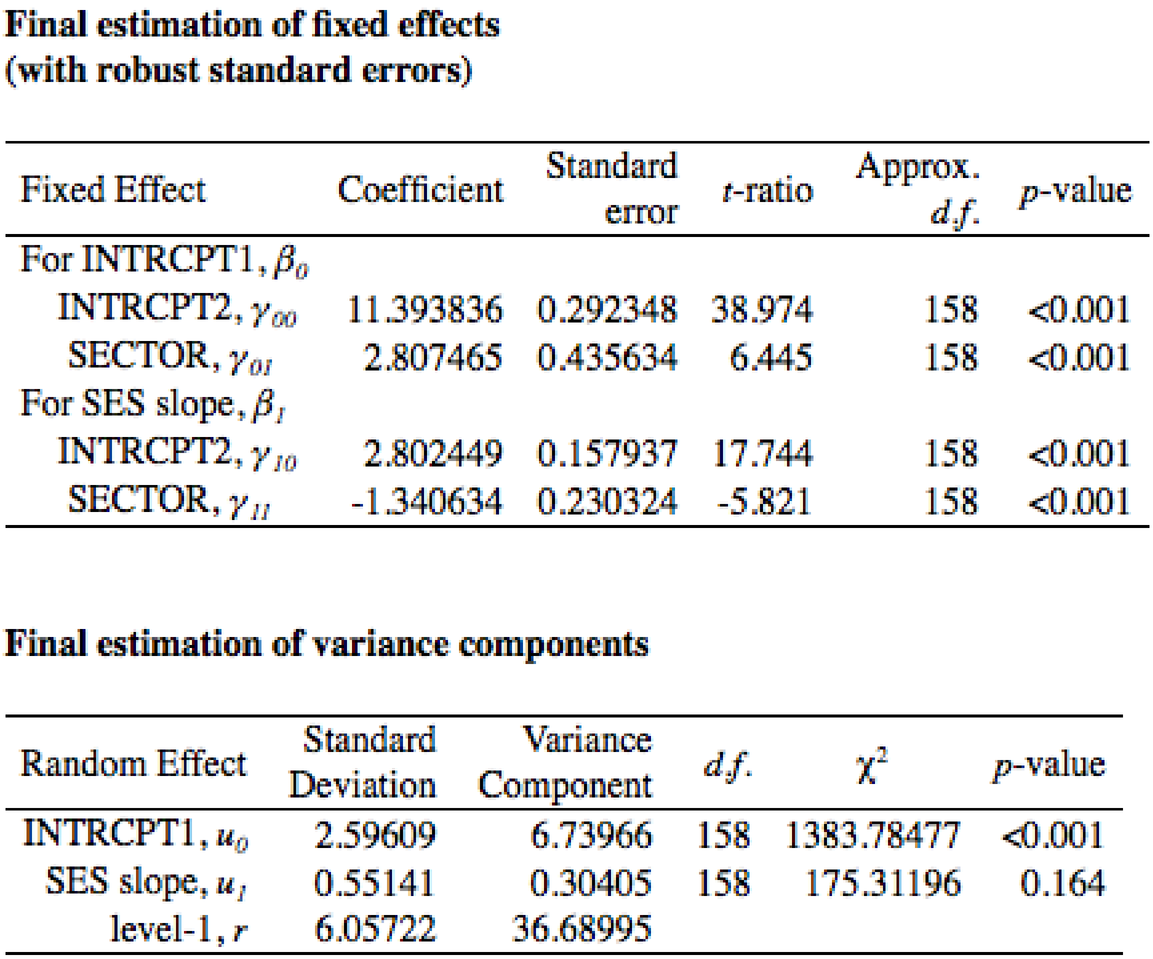
$$mAch_{i,j} = \gamma_{00} + \gamma_{01}(Catholic) + \gamma_{10}(ses_{ij} - s\bar{e}s_{.j}) + \gamma_{11}(Catholic) * (ses_{ij} - s\bar{e}s_{.j}) + u_{0j} + u_{1j} * (ses_{ij} - s\bar{e}s_{.j}) + r_{ij}$$

Analysis

Again, it is important to decide what you want to permit to vary and what you want to constrict. In HLM 7, you will have to toggle on the error terms for each slope. I have opted to let everything vary.

```
int_slope <- lmer(mAch ~ cses * sector + (1 + cses|school), data = df)
summary(int_slope)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: mAch ~ cses * sector + (1 + cses | school)
##    Data: df
##
## REML criterion at convergence: 46638.6
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.06490 -0.73237  0.01564  0.75373  2.94191
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  school   (Intercept)  6.7378  2.5957
##           cses         0.2657  0.5155   0.78
##  Residual             36.7056  6.0585
## Number of obs: 7185, groups:  school, 160
##
## Fixed effects:
##                      Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)           11.3939     0.2928 158.4233  38.919  < 2e-16 ***
## cses                   2.8028     0.1550 141.6652  18.087  < 2e-16 ***
```



Final estimation of fixed effects
(with robust standard errors)

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. $d.f.$	p-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	11.393836	0.292348	38.974	158	<0.001
SECTOR, γ_{01}	2.807465	0.435634	6.445	158	<0.001
For SES slope, β_1					
INTRCPT2, γ_{10}	2.802449	0.157937	17.744	158	<0.001
SECTOR, γ_{11}	-1.340634	0.230324	-5.821	158	<0.001

Final estimation of variance components

Random Effect	Standard Deviation	Variance Component	$d.f.$	χ^2	p-value
INTRCPT1, u_0	2.59609	6.73966	158	1383.78477	<0.001
SES slope, u_1	0.55141	0.30405	158	175.31196	0.164
level-1, r	6.05722	36.68995			

Figure 5: The relevant output from HLM 7.

```
## sectorCatholic        2.8075     0.4392 153.6914   6.393 1.85e-09 ***
## cses:sectorCatholic  -1.3411     0.2338 151.5416  -5.737 5.08e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) cses   sctrCt
## cses         0.259
## sectorCthlc -0.667 -0.173
## css:sctrCth -0.172 -0.663  0.261
```

```
ranova(int_slope)
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## mAch ~ cses + sector + (1 + cses | school) + cses:sector
##                             npar logLik   AIC    LRT Df Pr(>Chisq)
## <none>                         8 -23319 46655
## cses in (1 + cses | school)    6 -23325 46663 11.946  2   0.002547 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice that in the formula instead of using a + between cses and sector, a * will ensure to include the interaction and all the components of the interaction.

The R output can be a little confusing when there are predictors at both levels because it puts all interactions at the end rather than the order of the levels in the models. I find it easiest to decipher the γ's.

The following are in the order of the R output: $\gamma_{00} \Rightarrow$ Intercept of the intercepts (mean math achievement accounting for ses)
$\gamma_{10} \Rightarrow$ Intercept of the slopes (mean slope due to ses) $\gamma_{01} \Rightarrow$ Influence of level 2 predictor on intercept (influence of Catholic on mean math achievement) $\gamma_{11} \Rightarrow$ Influence of level 2 predictor on slopes (how the relationship between ses and math achievement changes due to Catholic)

Answers to research questions

1. Does the relationship between student ses and math achievement depend on whether the school is public or Catholic? Yes, examining γ_{11} in HLM 7 or the estimate for cses:Catholic we see that the influence of ses on math achievement is reduced by 1.34. That is, overall, one can expect a 2.8 point increase in math achievement for an increase in ses by 1 unit. This influence of ses on math achievement is reduced by 1.34 points. So that in a Catholic school each unit increase in ses is only associated with a 1.46 point increase in math achievement.

References

Douglas Bates, Martin Maechler and Ben Bolker (2014). mlmRev: Examples from Multilevel Modelling Software Review. R package version 1.0-6. https://CRAN.R-project.org/package=mlmRev

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Alexandra Kuznetsova, Per Bruun Brockhoff and Rune Haubo Bojesen Christensen (2016). lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-33. https://CRAN.R-project.org/package=lmerTest

Raudenbush, Stephen and Bryk, Anthony (2002), Hierarchical Linear Models: Applications and Data Analysis Methods, Sage (chapter 4).