The HSB Example

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Multilevel Regression Modeling, 2009

The HSB Example

Introduction

We take a quick look at the High School & Beyond example, the introductory example in the HLM manual and the Raudenbush & Bryk (2002) textbook.

The HSB Study

The data for this example are a subsample from the 1982 High School & Beyond Survey, and include information on 7185 students nested within 160 schools, 90 of which were public schools, 70 Catholic. Samples were on the order of 45 students per school.

The outcome variable Y_{ij} is math achievement. There is one potential level-1 predictor, SES of an individual student. At level 2, there were two potential (school-level) predictors: SECTOR (1 = Catholic, 0 = Public), and MEAN SES, the average SES of students at that school.

Key Research Questions

Raudenbush & Bryk (2002, p. 69) describe the key questions motivating their analyses:

1. How much do U.S. high schools vary in their mean math achievement?
2. Does a high level of SES in a school predict high math achievement?
3. Is the connection between student SES and math achievement similar across schools? Or does the relationship show substantial variation?
4. How do public and Catholic schools compare in terms of mean math achievement and in terms of the strength of association between SES and math achievement, after we control for the mean SES level at the schools?

Key Research Questions

Raudenbush & Bryk (2002, p. 69) describe the key questions motivating their analyses:

1. How much do U.S. high schools vary in their mean math achievement?
2. Does a high level of SES in a school predict high math achievement?
3. Is the connection between student SES and math achievement similar across schools? Or does the relationship show substantial variation?
4. How do public and Catholic schools compare in terms of mean math achievement and in terms of the strength of association between SES and math achievement, after we control for the mean SES level at the schools?

Key Research Questions

Raudenbush & Bryk (2002, p. 69) describe the key questions motivating their analyses:

1. How much do U.S. high schools vary in their mean math achievement?
2. Does a high level of SES in a school predict high math achievement?
3. Is the connection between student SES and math achievement similar across schools? Or does the relationship show substantial variation?
4. How do public and Catholic schools compare in terms of mean math achievement and in terms of the strength of association between SES and math achievement, after we control for the mean SES level at the schools?

Key Research Questions

Raudenbush & Bryk (2002, p. 69) describe the key questions motivating their analyses:

1. How much do U.S. high schools vary in their mean math achievement?
2. Does a high level of SES in a school predict high math achievement?
3. Is the connection between student SES and math achievement similar across schools? Or does the relationship show substantial variation?
4. How do public and Catholic schools compare in terms of mean math achievement and in terms of the strength of association between SES and math achievement, after we control for the mean SES level at the schools?

Connecting the Substantive and the Statistical

On the basis of the radon example we worked through in the last lecture, you should already have a few hunches about how to address the substantive research questions with multilevel statistical models. Let's work through the examples, replicating them in R as we go.

Combining Level-1 and Level-2 Data

Before we start, let's create the R file we need. HLM gives us two SPSS .SAV files, one for each level.

We need to add the level-2 variables to the level-1 file to create a file that R can use.

We start by reading in the two files. Make sure that `Hmisc` and `foreign` libraries are loaded, along with `arm` and `lme4`.

Combining Level-1 and Level-2 Data

Combining the files takes several steps:

- Read in the level-1 file and attach it so that the ID variable is visible.
- Read in the level-2 file.
- The level-2 file variables are replicated by referencing them to the (visible) ID variable at the student level.
- After creating expanded versions of all the level-2 variables, we create a new data frame with all the variables.

Combining Level-1 and Level-2 Data

Combining the files takes several steps:

- Read in the level-1 file and attach it so that the ID variable is visible.
- Read in the level-2 file.
- The level-2 file variables are replicated by referencing them to the (visible) ID variable at the student level.
- After creating expanded versions of all the level-2 variables, we create a new data frame with all the variables.

Combining Level-1 and Level-2 Data

Combining the files takes several steps:

- Read in the level-1 file and attach it so that the ID variable is visible.
- Read in the level-2 file.
- The level-2 file variables are replicated by referencing them to the (visible) ID variable at the student level.
- After creating expanded versions of all the level-2 variables, we create a new data frame with all the variables.

Combining Level-1 and Level-2 Data

Combining the files takes several steps:

- Read in the level-1 file and attach it so that the ID variable is visible.
- Read in the level-2 file.
- The level-2 file variables are replicated by referencing them to the (visible) ID variable at the student level.
- After creating expanded versions of all the level-2 variables, we create a new data frame with all the variables.

Combining Level-1 and Level-2 Data

```
> hsb1 ← spss.get("hsb1.sav")
> hsb2 ← spss.get("hsb2.sav")
> attach(hsb1)
> SIZE ← hsb2$SIZE[ID]
> SECTOR ← hsb2$SECTOR[ID]
> PRACAD ← hsb2$PRACAD[ID]
> DISCLIM ← hsb2$DISCLIM[ID]
> HIMINTY ← hsb2$HIMINTY[ID]
> MEANSES ← hsb2$MEANSES[ID]
> hsb.all ← data.frame(ID,MINORITY,FEMALE,
+ SES,MATHACH,SIZE,SECTOR,PRACAD,DISCLIM,
+ HIMINTY,MEANSES)
```

We can then write this data frame for safe-keeping.

```
> write.table(hsb.all,"HSBALL.TXT",
+ col.names = T, row.names = F)
```

One-Way ANOVA

The analysis of variance model provides us with useful preliminary information about how much total variation in math achievement occurs within and between schools.

It also can provide useful information about the reliability of each school's sample mean as an estimate of its true population mean.

Preparing for Analysis

In this example, we shall be using the supplied data files *hsb1.sav* and *hsb2.sav* as, respectively, the level-1 and level-2 files. See if you can execute the following steps on your own:

- Start up HLM
- Load *hsb1.sav* as the level-1 file, and select `MATHACH` as the outcome variable, and `SES` as a potential predictor. `ID` is the ID variable.
- Load *hsb2.sav* as the level-2 file, and select `ID` as the ID variable and include `SECTOR` and `MEANSES` as potential level-2 predictors
- Enter *hsb1.mdm* as the MDM file name, and save the MDMT file, entering the name *HSB1* when asked (Remember, there is no need for an extension on the MDMT file name, but there IS a need for an extension on the MDM file name!)
- Make the MDM file.

Checking the Statistics

After creating the MDM file, check the statistics. They should look like this:

LEVEL-1 DESCRIPTIVE STATISTICS

VARIABLE NAME	N	MEAN	SD	MINIMUM	MAXIMUM
SES	7185	0.00	0.78	-3.76	2.69
MATHACH	7185	12.75	6.88	-2.83	24.99

LEVEL-2 DESCRIPTIVE STATISTICS

VARIABLE NAME	N	MEAN	SD	MINIMUM	MAXIMUM
SECTOR	160	0.44	0.50	0.00	1.00
MEANSES	160	-0.00	0.41	-1.19	0.83

Now, create and analyze a 1-way random-effects ANOVA. Save the model as *OneWayAnova.hlm*.

Basic Output

The basic output consists of estimates of the fixed-effects coefficient γ_{00} and the variances τ_{00} and σ^2, respectively, of the random variables u_{0j} (representing variance across schools) and r_{ij} representing within school variance.

```
 The outcome variable is  MATHACH

 Final estimation of fixed effects:
 ---------------------------------------------------
                                      Standard             Approx.
    Fixed Effect         Coefficient   Error      T-ratio   d.f.     P-value
 ---------------------------------------------------
 For       INTRCPT1, B0
    INTRCPT2, G00          12.636972   0.244412    51.704       159    0.000
 ---------------------------------------------------
Final estimation of variance components:
 ---------------------------------------------------
 Random Effect           Standard      Variance     df    Chi-square  P-value
                         Deviation     Component
 ---------------------------------------------------
 INTRCPT1,       U0        2.93501       8.61431   159    1660.23259    0.000
  level-1,       R         6.25686      39.14831
 ---------------------------------------------------

 Statistics for current covariance components model
 ------------------------------------
 Deviance                        = 47116.793477
 Number of estimated parameters = 2
```

Interpreting Basic Output

The estimate for the grand mean of high school achievement is 12.64. The estimated standard error is .244412. In the Raudenbush & Bryk (2002) text, a 95% confidence interval on γ_{00} is calculated using a normal approximation as

$$12.64 \pm 1.96(0.24)$$

resulting in limits of 12.17 and 13.11.

Since this coefficient is tested for significance with a t-statistic with 159 degrees of freedom, it is not clear why the t-distribution was not used to construct the confidence interval, or why the standard error was rounded off from .244 to .24. In any case, it doesn't make much difference.

Interpreting Basic Output

Under the assumptions of the model, the population of school *population* means is normally distributed around γ_{00} with variance τ_{00}.

So 95% of the school population means should be within $\gamma_{00} \pm 1.96(\tau_{00})^{1/2}$. Raudenbush and Bryk (2002, p. 71) refer to this as the *plausible values range.*

In this case, we estimate the plausible values range as

$$
\begin{aligned}
\hat{\gamma}_{00} &\pm 1.96(\hat{\tau}_{00})^{1/2} \quad (1)\\
12.64 &\pm 1.96(8.61)^{1/2} \quad (2)\\
12.64 &\pm 2.94 \quad (3)
\end{aligned}
$$

which yields endpoints of 6.89 and 18.39.

That's a very substantial range!

A Statistical Side-Question

If we calculated sample means on math achievement for each of the 160 schools, would we expect the range of the sample means to be greater or less than the bounds shown? Why?

Intraclass Correlation

The *intraclass correlation* is the proportion of total variance in math achievement that is between schools. This is estimated as

$$\hat{\rho} = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}^2} = \frac{8.61}{8.61 + 39.15} = 0.18 \tag{4}$$

Reliability of Sample Means

The reliability of an estimate is the proportion of total variance that is "true score variance" as opposed to "error variance." As we learned in Psychology 310, the sample mean $\overline{Y}_{\bullet j}$ can be written as

$$\overline{Y}_{\bullet j} = \mu_j + \epsilon_j \tag{5}$$

What are the variances of each of these terms?

Reliability of Sample Means

That's right, according to the model, the means were taken from a population such that the population means across j actually have a variance, τ_{00}, and from basic theory, we know that a sample mean $\overline{Y}_{\bullet j}$ varies around its population mean with variance σ^2/n_j, so

$$\hat{\lambda}_j = \text{reliability}(\overline{Y}_{\bullet j}) = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}^2/n_j} \tag{6}$$

An "overall measure of reliability" can be obtained by averaging these sample estimates.

```
 ----------------------------------
  Random level-1 coefficient   Reliability estimate
 ----------------------------------
  INTRCPT1, B0                       0.901
 ----------------------------------
```

Replicating the Analysis with R

Examine your mixed model, and, before looking at the input and output on the next slide, see if you can recall how to get the output from R.

Replicating the Analysis with R

```
> one.way.fit ← lmer(MATHACH ~ 1 + (1|ID))
> summary(one.way.fit)
```

```
Linear mixed model fit by REML
Formula: MATHACH ~ 1 + (1 | ID)
   AIC   BIC logLik deviance REMLdev
 47123 47143 -23558    47116   47117
Random effects:
 Groups   Name        Variance Std.Dev.
 ID       (Intercept)  8.61    2.93
 Residual             39.15    6.26
Number of obs: 7185, groups: ID, 160

Fixed effects:
            Estimate Std. Error t value
(Intercept)   12.637      0.244    51.7
```

Introduction

In this model, we predict overall level of math achievement within a school from the overall SES level at that school. We do this by introducing a level-2 predictor, `MEANSES`. while continuing to model student variation around the school mean as random.

Model Setup

We are going to continue to use the same MDM file we created before.

Simply add `MEANSES` as a predictor at level 2.

Save your model as *HSBMODEL1.hlm* and analyze it.

Basic Output

The key output looks like this:

```
The outcome variable is  MATHACH

Final estimation of fixed effects:
---------------------------------------------------
                                        Standard             Approx.
   Fixed Effect         Coefficient     Error      T-ratio   d.f.     P-value
---------------------------------------------------
For       INTRCPT1, B0
   INTRCPT2, G00          12.649436    0.149280    84.736      158     0.000
    MEANSES, G01           5.863538    0.361457    16.222      158     0.000
---------------------------------------------------
 Final estimation of variance components:
---------------------------------------------------
Random Effect           Standard      Variance     df    Chi-square  P-value
                        Deviation     Component
---------------------------------------------------
INTRCPT1,       U0        1.62441       2.63870   158     633.51744    0.000
 level-1,       R         6.25756      39.15708
---------------------------------------------------


Statistics for current covariance components model
---------------------------------
Deviance                          = 46959.446959
Number of estimated parameters = 2
```

Interpreting the Output

There is a highly significant association between `MEANSES` and math achievement, as the t statistic of 16.22 indicates. Note also that the residual variance between schools, estimated as 2.64, is much smaller than before (8.61).

We can compute a "range of plausible values" for school means *given a mean SES of zero* as $12.65 \pm (2.64)^{1/2}$ which computes as (9.47, 15.83).

Variance Explained at Level 2

By comparing estimates of τ_{00} for the two models, we can estimate the proportional reduction of variance explained in the β_{0j}. This is

$$\frac{8.61 - 2.64}{8.61} \tag{7}$$

Conditional Intraclass Correlation

After removing the effect of school mean SES, the correlation between pairs of scores in the same school, which was estimated previously at .18, is now estimated as

$$\begin{aligned} \hat{\rho} &= \hat{\tau}_{00}/(\hat{\tau}_{00} + \hat{\sigma}^2) && (8) \\ &= 2.64/(2.64 + 39.16) && (9) \\ &= .06 && (10) \end{aligned}$$

This measures the degree of dependence among observations within schools that are of the same mean SES.

Summing it Up

This analysis demonstrates that the overall level of SES within a school is significantly (positively) related to mean achievement in the school. Nonetheless, even after controlling for this important factor, there is still substantial variation across schools in their average achievement level.

Replicating in R

Using the principles we discussed in class, take the mixed model specification from HLM and write the equivalent model to be fit by `lmer()` in R. Check your input and output against the next page.

Replicating in R

```
> fit.2 <- lmer(MATHACH ~ MEANSES + (1|ID))
> summary(fit.2)
```

```
Linear mixed model fit by REML
Formula: MATHACH ~ MEANSES + (1 | ID)
   AIC   BIC logLik deviance REMLdev
 46969 46997 -23481    46959   46961
Random effects:
 Groups   Name        Variance Std.Dev.
 ID       (Intercept)  2.64    1.62
 Residual             39.16    6.26
Number of obs: 7185, groups: ID, 160

Fixed effects:
            Estimate Std. Error t value
(Intercept)   12.649      0.149    84.7
MEANSES        5.864      0.361    16.2

Correlation of Fixed Effects:
        (Intr)
MEANSES -0.004
```

The Random-Coefficients Model

We now conceptualize each school as having a school-specific regression line (slope and intercept) relating a student's achievement to SES relative to that school's norm.

We conceptualize these slopes and intercepts varying around central values according to a bivariate normal distribution that allows the slopes and intercepts to be correlated, and to have different variances. Some questions to be addressed include:

- What is the meaning of the slope within a school? The intercept?
- What is the average of the 160 group regression equations?
- How much do the regression equations vary across schools? The slopes? The intercepts?
- What is the correlation between slopes and intercepts across schools?

The Random-Coefficients Model

We now conceptualize each school as having a school-specific regression line (slope and intercept) relating a student's achievement to SES relative to that school's norm.

We conceptualize these slopes and intercepts varying around central values according to a bivariate normal distribution that allows the slopes and intercepts to be correlated, and to have different variances. Some questions to be addressed include:

- What is the meaning of the slope within a school? The intercept?
- What is the average of the 160 group regression equations?
- How much do the regression equations vary across schools? The slopes? The intercepts?
- What is the correlation between slopes and intercepts across schools?

The Random-Coefficients Model

We now conceptualize each school as having a school-specific regression line (slope and intercept) relating a student's achievement to SES relative to that school's norm.

We conceptualize these slopes and intercepts varying around central values according to a bivariate normal distribution that allows the slopes and intercepts to be correlated, and to have different variances. Some questions to be addressed include:

- What is the meaning of the slope within a school? The intercept?
- What is the average of the 160 group regression equations?
- How much do the regression equations vary across schools? The slopes? The intercepts?
- What is the correlation between slopes and intercepts across schools?

The Random-Coefficients Model

We now conceptualize each school as having a school-specific regression line (slope and intercept) relating a student's achievement to SES relative to that school's norm.

We conceptualize these slopes and intercepts varying around central values according to a bivariate normal distribution that allows the slopes and intercepts to be correlated, and to have different variances. Some questions to be addressed include:

- What is the meaning of the slope within a school? The intercept?
- What is the average of the 160 group regression equations?
- How much do the regression equations vary across schools? The slopes? The intercepts?
- What is the correlation between slopes and intercepts across schools?

The Level 1 Model

At level 1, our model is

$$\text{MATHACH}_{ij} = \beta_{0j} + \beta_{1j}(\text{SES}_{ij} - \overline{\text{SES}}_{\bullet j}) + r_{ij} \tag{11}$$

Each school has its own slope and intercept.

The Level 2 Model

At level 2, we simply model random variation. There are no level-2 predictors.

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{12}$$

$$\beta_{0j} = \gamma_{10} + u_{1j} \tag{13}$$

We assume that β_{0j} and β_{1j} are bivariate normal, with covariance matrix $\boldsymbol{T}$ with non-redundant elements $\tau_{00} = \mathrm{Var}(\beta_{0j})$, $\tau_{11} = \mathrm{Var}(\beta_{1j})$, and $\tau_{10} = \mathrm{Cov}(\beta_{0j}, \beta_{1j})$

HLM Setup

Most of this should be pretty routine for you by now. Don't forget that, when you add `SES` as a predictor at level 1, make sure to specify that it is centered around its own group mean.

```
The outcome variable is  MATHACH

 Final estimation of fixed effects:
 ----------------------------------------------------
                                     Standard             Approx.
    Fixed Effect         Coefficient   Error      T-ratio   d.f.     P-value
 ----------------------------------------------------
 For       INTRCPT1, B0
    INTRCPT2, G00          12.636196   0.244503    51.681       159    0.000
 For      SES slope, B1
    INTRCPT2, G10           2.193157   0.127879    17.150       159    0.000
 ----------------------------------------------------

 Final estimation of variance components:
 ----------------------------------------------------
 Random Effect           Standard      Variance     df    Chi-square  P-value
                         Deviation     Component
 ----------------------------------------------------
 INTRCPT1,       U0        2.94633       8.68087   159    1770.85115    0.000
      SES slope, U1        0.82485       0.68038   159     213.43769    0.003
  level-1,       R         6.05835      36.70356
 ----------------------------------------------------


 Statistics for current covariance components model
 --------------------------------
 Deviance                        = 46712.398927
```

Interpreting Output

Can you construct a 95% interval of "feasible values" for the group-specific intercept?

How about the group specific slope?

What do these values suggest?

Introduction

Having established that the regression relationship between achievement and SES varies considerably across schools, we now seek to further understand the factors associated with this variation. We expand the model to predict slopes and intercepts at level 1 from mean SES and sector (Catholic or public) at level 2.

The Model

The level 1 model stays the same.

At level 2, our model now becomes

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\text{MEANSES}_j + \gamma_{02}\text{SECTOR}_j + u_{0j} \tag{14}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\text{MEANSES}_j + \gamma_{12}\text{SECTOR}_j + u_{0j} \tag{15}$$

HLM Setup

The model is the same as its predecessor, except at level 2 we need to add the two predictors, uncentered. Save your model as *HSB3.hlm*

Output

```
The outcome variable is  MATHACH

Final estimation of fixed effects:
---------------------------------------------------
                                  Standard             Approx.
   Fixed Effect         Coefficient   Error      T-ratio   d.f.     P-value
---------------------------------------------------
For       INTRCPT1, B0
   INTRCPT2, G00          12.096006   0.198734    60.865       157     0.000
     SECTOR, G01           1.226384   0.306272     4.004       157     0.000
    MEANSES, G02           5.333056   0.369161    14.446       157     0.000
For      SES slope, B1
   INTRCPT2, G10           2.937981   0.157135    18.697       157     0.000
     SECTOR, G11          -1.640954   0.242905    -6.756       157     0.000
    MEANSES, G12           1.034427   0.302566     3.419       157     0.001
---------------------------------------------------
Final estimation of variance components:
---------------------------------------------------
Random Effect           Standard      Variance     df    Chi-square  P-value
                        Deviation     Component
---------------------------------------------------
INTRCPT1,       U0        1.54271       2.37996   157     605.29503    0.000
     SES slope, U1        0.38590       0.14892   157     162.30867    0.369
 level-1,       R         6.05831      36.70313
---------------------------------------------------


Statistics for current covariance components model
---------------------------------
Deviance                       = 46501.875643
Number of estimated parameters = 4
```