

SC705: Advanced Statistics
Instructor: Natasha Sarkisian

Introduction to Hierarchical Linear Modeling

Hierarchical models (also known as multilevel models or mixed models) are used to handle nested data structures, e.g.:

Students – classrooms – schools

Workers – firms

Individuals – neighborhoods

Households – countries

Repeated measures over time – individuals

Traditional regression methods only allow analyzing such data by focusing on one level – e.g. focusing on students and ignoring that they are nested within classrooms and schools – but that leads to errors.

Using relationships between variables on the level of the individuals to make conclusions about groups → atomistic fallacy

[Note: We can aggregate the information collected from individuals to characterize groups – but the analysis should be conducted on the level of groups.]

Using relationships between variables describing groups to make conclusions about relationships between variables on the level of individuals → ecological fallacy

Traditional regression methods assume that the relationships among individual-level variables are the same in all groups (they assume homogeneity of regression). Therefore, they do not allow to assess how the relationships between variables at the individual level might vary according to the group context, or to study what it is about group context that may cause such variation.

In addition, traditional regression methods produce biased results in multilevel samples – they assume independent observations, and therefore misestimate standard errors – fail to take into account the dependence among observations that belong to the same group.

HLM regression models resolve all these problems:

- They allow simultaneously estimating relationships at individual and group levels
- They allow for the relationships at the individual level to vary across groups
- They allow to model cross-level interactions – i.e., examine how group-level factors shape individual-level relationships
- They allow to correctly estimate standard errors by dividing the unexplained variance into two components – group-level (random effects of groups) and individual level:
 $Y_{ij} = \alpha + X\beta + u_j + e_{ij}$ where u_j is the effect of membership in group j , and e_{ij} is the residual effect for individual i within this group.

Such model with random effects can also be interpreted as a two-level model, where:

Level 1 model is: $Y_{ij} = \alpha + X\beta + e_{ij}$

Level 2 model is: $\alpha = \mu + u_i$,

In this kind of model, we treat the intercept α as a random variable. This is called conditional model with a random intercept. We will learn more about this and other types of HLM models in this course. The basic point is: When studying complex, multilevel processes, we should use theories and analytic techniques that are also multilevel – hence we need HLM.

Learning the basics of HLM program

To run an analysis in HLM, four steps are required:

1. Select the general type of model to be fitted.
2. Create or open an MDM file.
3. Specify the model and various statistical options and output options.
4. Run the model; after that, model-based graphs can be obtained.

Types of models

The HLM program has 5 modules that may be used to fit different types of models:

1. HLM2 module -- two-level linear and non-linear (HGLM) models
2. HLM3 module -- three-level linear and non-linear (HGLM) models
3. The HMLM module -- estimation of longitudinal models with a variety of covariance structures
4. HMLM2 – three-level HMLM
5. The HCM2 module -- two-level cross-classified models, where lower-level units simultaneously belong to 2 higher-level units

In this beginning of this course, we will primarily work with HLM2 so for now we select HLM2.

Creating an MDM dataset

Next, we need to construct the Multivariate Data Matrix (MDM) from raw data or from a dataset created by some statistical package. It is usually more convenient to import data from a statistical package (HLM accommodates a range of them). Your data could be stored as either one file or two files (one for each level).

We need to select the Make new MDM file option on the File menu, and read in the data file.

The procedure to create an MDM file consists of the following steps:

1. Specify the input file type.
2. Give name to your new .MDM file
3. Specify level 1 and level 2 data files and variables
4. Specify whether there are missing data, and how you want to deal with these
5. Once everything is set up, give a name to your template file (.MDTM)
6. Create the file and check the stats (descriptive statistics on all variables included in an MDM file are saved to a file automatically placed in the same folder as the MDM file, with a .STS file extension; this file can be opened in Notepad, Wordpad, etc.).

We'll be working with "High School and Beyond" data (included with HLM – see Examples, Chapter 2 in HLM folder: HSB1.sav and HSB2.sav). I also placed these data files on the course website as Stata data files: hsb1.dta and hsb2.dta – one file for each level, and hsb.dta – the two levels combined.

The level-1 file (hsb1.dta) has 7185 cases (students) and 5 variables:

id – school number

minority – an indicator of students' ethnicity (1=minority, 0=other)

female – an indicator of students' gender (1=female, 0=male)

ses – a standardized scale constructed from measures of parental occupation, education, and income

mathach – a measure of mathematics achievement

The level-2 file (hsb2.dta) has 160 cases (schools) and 7 variables:

id – school number

size – school enrollment

sector – 1=Catholic, 0=public

pracad – proportion of students in the academic track

disclim – a scale measuring disciplinary climate

himnty – 1=more than 40% minority enrollment, 0=less than 40%

meanses – mean of the SES values for the students in each school (generated as group means from level 1 file).

It is essential that both files contain group identifier (in this case, school ID).

Important: Datasets must be sorted by the level-2 ID (if using 3 levels, also the level-3 ID)!

The combined file (hsb.dta) has 7185 cases (students) and 11 variables – same as above. Note that school-level variables now have 7185 observations, but they are the same for all students within each school. This is called the disaggregation of level 2 predictors.

Let's construct the MDM file.

LEVEL-1 DESCRIPTIVE STATISTICS

| VARIABLE NAME | N | MEAN | SD | MINIMUM | MAXIMUM |
|---------------|------|-------|------|---------|---------|
| MINORITY | 7185 | 0.27 | 0.45 | 0.00 | 1.00 |
| FEMALE | 7185 | 0.53 | 0.50 | 0.00 | 1.00 |
| 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 |
|---------------|-----|---------|--------|---------|---------|
| SIZE | 160 | 1097.83 | 629.51 | 100.00 | 2713.00 |
| SECTOR | 160 | 0.44 | 0.50 | 0.00 | 1.00 |
| PRACAD | 160 | 0.51 | 0.26 | 0.00 | 1.00 |
| DISCLIM | 160 | -0.02 | 0.98 | -2.42 | 2.76 |
| HIMINTY | 160 | 0.28 | 0.45 | 0.00 | 1.00 |
| MEANSES | 160 | -0.00 | 0.41 | -1.19 | 0.83 |

We want to carefully examine these statistics and make sure that they correspond to what we would expect on the basis of our level 1 and level 2 files. You cannot explore your data in HLM so you should familiarize yourself with your data using another statistical package before importing the data into HLM. Once the MDM file is constructed, all subsequent analyses will be computed using the MDM file as input – if you want to change the data, you’ll need to construct another MDM file.

Specifying the model

Let’s try to estimate the simplest possible model. We just need to specify the outcome variable, and the basic model appears – this is the simplest possible model. It is known as the fully unconditional model (FUM) -- no predictors are specified at either level 1 or level 2.

Note: To get more information on the formulas used, go to File→ Preferences and select “Show mixed model” and “Use subscripts.”

Model 0. Unconditional model with random intercept (a.k.a. intercept-only model, or one way ANOVA with random intercept):

LEVEL 1 MODEL

$$\text{MATHACH}_{ij} = \beta_{0j} + r_{ij}$$

LEVEL 2 MODEL

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

MIXED MODEL

$$\text{MATHACH}_{ij} = \gamma_{00} + u_{0j} + r_{ij}$$

[Technical note: you can save the equations as a picture file – go to File→Save model as .emf]

γ_{00} is the grand mean (i.e. average intercept) – this is the fixed component of the model (fixed effect)

The two random components are:

$$r_{ij} \sim N(0, \sigma^2)$$

$$u_{0j} \sim N(0, \tau_{00})$$

Note that this model is very similar to a one-way ANOVA model utilizing the grouping variable as a nominal-level variable. What distinguishes the two is the random variable nature of u_{0j} -- in a regular one-way ANOVA, each u_{0j} is a fixed number, in a sense it is the value of a dummy-variable indicator for that specific group. In HLM models, however, u_{0j} is modeled as a random variable rather than a set of fixed coefficients.

Estimating the fully unconditional model is useful as a preliminary step in a hierarchical data analysis. Its most important function is to provide the information about outcome variability at each of the two levels. Sigma (σ) will provide the information about level-1 (within-group)

variability, and tau (τ) will provide the information on level-2 (between-group) variability. Running this model allows us to decompose the variance in the dependent variable into variance components for each hierarchical level -- into within-group and between-group variance. This model does not explain anything, but it allows us to evaluate whether there is variation across groups, and how much of it. That's why it is always a good idea to run this basic model when starting the analyses -- it's the null model of our regression analysis. If we find that there is no significant between-group variation, then there is no need for a hierarchical model.

The proportion of variance due to group-level variation in means can be calculated as

$$\rho = \tau_{00} / (\sigma^2 + \tau_{00})$$

and it represents the *intra-class correlation coefficient*. It can be interpreted as the proportion of variance explained by the grouping structure in the population.

Running the model:

```
-----
Module:          HLM2.EXE (6.02.25138.2)
-----
SPECIFICATIONS FOR THIS HLM2 RUN
  Problem Title: no title
  The data source for this run = C:\Program Files\HLM6\Examples\Chapter2\HSB.MDM
  The command file for this run = whlmtmp.hlm
  Output file name           = C:\Program Files\HLM6\Examples\Chapter2\hlm2.txt
  The maximum number of level-1 units = 7185
  The maximum number of level-2 units = 160
  The maximum number of iterations = 100
  Method of estimation: restricted maximum likelihood

Weighting Specification
-----
                Weight
                Variable
Level 1      Weighting?  Name      Normalized?
Level 2      no
Precision    no

The outcome variable is  MATHACH

The model specified for the fixed effects was:
-----
Level-1      Level-2
Coefficients Predictors
-----
INTRCPT1, B0  INTRCPT2, G00

The model specified for the covariance components was:
-----
Sigma squared (constant across level-2 units)

Tau dimensions
INTRCPT1
Summary of the model specified (in equation format)
-----
Level-1 Model
  Y = B0 + R

Level-2 Model
  B0 = G00 + U0
```

Iterations stopped due to small change in likelihood function
 ***** ITERATION 4 *****

Sigma_squared = 39.14831

Tau
 INTRCPT1,B0 8.61431

Tau (as correlations)
 INTRCPT1,B0 1.000

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

The value of the likelihood function at iteration 4 = -2.355840E+004

The outcome variable is MATHACH

Final estimation of fixed effects:

| Fixed Effect | Coefficient | Standard Error | T-ratio | Approx. d.f. | P-value |
|------------------|-------------|----------------|---------|--------------|---------|
| For INTRCPT1, B0 | | | | | |
| INTRCPT2, G00 | 12.636972 | 0.244412 | 51.704 | 159 | 0.000 |

The outcome variable is MATHACH

Final estimation of fixed effects
 (with robust standard errors)

| Fixed Effect | Coefficient | Standard Error | T-ratio | Approx. d.f. | P-value |
|------------------|-------------|----------------|---------|--------------|---------|
| For INTRCPT1, B0 | | | | | |
| INTRCPT2, G00 | 12.636972 | 0.243628 | 51.870 | 159 | 0.000 |

Final estimation of variance components:

| Random Effect | | Standard Deviation | Variance Component | df | Chi-square | P-value |
|---------------|--|--------------------|--------------------|-----|------------|---------|
| 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

Only one fixed effect is estimated in this model – that’s the average value of the outcome across all individuals – here, the average math achievement is estimated to be 12.64.

The main thing we have to conclude from examining this output is that there is a significant amount of school-level variation in math achievement. The intra-class correlation is:

$$\rho = \tau_{00} / (\tau_{00} + \sigma^2) = 8.61431 / (8.61431 + 39.14831) = .18035673$$

We also use variance components to estimate the reliability of the sample mean for any school as an estimate of its population mean. Such reliability for a particular group is calculated as:

$\lambda_j = \tau_{00} / (\tau_{00} + \sigma^2/n_j)$ where n_j is the sample size for group j . For example, school #1224 has 47 students, therefore:

$$\lambda_{1224} = 8.61431 / (8.61431 + 39.14831/47) = .91183228$$

Reliability ranges from 0 to 1 so .91 is pretty high.

School #1308 has only 20 students, therefore:

$$\lambda_{1308} = 8.61431 / (8.61431 + 39.14831/20) = .81484427$$

HLM output includes an average of such reliabilities:

$$\lambda = \sum \lambda_j / j = 0.901.$$

This number indicates whether estimated differences across schools are reliable indicators of real differences among schools' population means.