Getting Started with HLM 5

For Windows

Updated: August 2012

Table of Contents

Section 1: Overview

1.1 About this Document

This document is designed to introduce you to the HLM 5 for Windows software. In order to make use of this document, you should have a background in regression. A background in hierarchical or multilevel models is useful, but not necessary for understanding materials in this document. While the primary use of the document is to familiarize you with the use of the software, topics in hierarchical modeling are discussed in enough detail that you should be able to implement the techniques described here in your own data analyses.

This document makes use of some sample datasets that are available with the software, or available on the University of Texas Microsoft Windows terminal server. The primary datasets are contained in the *Examples* directory in the *HLM 5* directory.

1.2 Introduction to HLM

HLM stands for *hierarchical linear models*, which are a type of model used for analyzing data in a clustered or nested structure. An example of such data is students who are nested within classrooms, which are nested within schools; in this situation, we would expect that students within a cluster, such as a classroom or school, would share some similarities due to their common environment. Hierarchical linear models are also known as *multilevel models*, *random coefficient models*, or *random effects models*. HLM can be used to analyze a variety of questions with either categorical or continuous dependent variables.

1.3 Accessing HLM

You may access HLM in one of three ways:

1. License a copy from Scientific Software International for your own personal computer.
2. Download the free student version of HLM from Scientific Software International for your own personal computer. If your models are small, the free demonstration version may be sufficient to meet your needs. For larger models, you will need to purchase your own copy of HLM or access the ITS shared copy of the software through the campus network. The latter option is typically more cost effective, particularly if you decide to access the other software programs available on the server (e.g., SAS, SPSS, AMOS, etc.).

1.4 Getting Help with HLM

If you are a member of UT-Austin, you can schedule an appointment with a statistical consultant or send e-mail to stat.consulting@austin.utexas.edu . See stat.utexas.edu/consulting/ for more details about consulting services, as well as answers to frequently asked questions about hierarchical models, multilevel models, HLM and other topics. Non-UT and UT HLM users will find the HLM site to be a useful resource.

Section 2: Accessing Data in HLM

2.1 File Format Requirements

There are two methods for entering data into HLM: (1) importing ASCII data, or (2) importing files from one of the following statistical software packages: SAS, SPSS, or SYSTAT. The present example assumes that your data are saved in an SPSS format. If this not the case, you can consult our tutorial "SPSS for Windows: Getting Started" to convert your ASCII or Excel document into an SPSS format.

The first step to getting started with HLM is to create a SSM (sufficient statistics matrix) file, which is the file format that the HLM software uses. To create a file that can be used in HLM, you will first need a file for each level of your model. In the present example, a two-level model is illustrated and consequently, level-1 and level-2 datasets are needed. A level-1 model, as the name suggests, contains data on level-1 units whereas level-2 models contain information on level-2 units. The level-1 units are typically subject-level units, such as individual students. The level-2 units are typically units in which the level-1 units are nested, such as schools. If you have entered your data into SPSS or another statistical software package and were not originally anticipating using HLM, your data are likely to not be in these two separate files, so you will need to create these two separate files from your original file. In addition, there are some requirements that need to be met in order to use these files in HLM:

- Level-1 units should be grouped together by the ID of their level-2 unit's ID.
- ID variables should be in a string format.
- ID variables must not exceed 12 characters.
- ID variables should all be the same length. For example, if you had ID's between 1 and 20,000, you should format the value 1 as 00001 so that it is the same length as the largest ID value.

As your data are not likely to meet the above requirements, you will probably need to make some modifications to your data. These modifications can be done using a variety of statistical software packages. While you should use the package with which you are most familiar, a few brief suggestions are offered here for using SPSS to prepare your data. If you are using SAS or SYSTAT, the instructions below should provide some guidelines for general considerations for importing data using these file formats.

If your data are in a multivariate format, you will need to transform that data into a univariate format. A multivariate format is a data structure where there is more than one dependent variable per unit in the rows of your dataset. For example, if you had a dataset on married couples where you had a single row for each couple that contained each member of the couple's marital satisfaction score, this would be a multivariate format, as there is more than one dependent variable per row. In this situation, you would want to modify the dataset so that there is only a single dependent variable per row, which would mean that each member of the couple's score would be in a separate row. For an example of how to transform your data into a univariate format, see SPSS FAQ: Converting SPSS multivariate repeated measures data to univariate

format.

The next step is to sort your data on the basis of their level-2 ID. For instructions on how to do this, see the **Sorting Cases** section of our "SPSS for Windows: Getting Started" document. Next, you will need to convert your ID variables to a string format. There are general instructions on how to change the format of a variable in **Creating and Defining Data** section of the above mentioned document. That document illustrates the use of the *Variable Type* dialog box; to change the variable format, select the *String* option in the dialog box shown in that section of the document. If your ID variables are not all the same length, the next requirement is to make all of them the same length by placing leading zeros on values that are less that maximum length. The following SPSS syntax can be modified to perform this operation (if you have any questions about how to execute SPSS syntax, examine the **Syntax** section of the "SPSS for Windows: Data Manipulation and Advanced Topics" document):

IF (LENGTH(RTRIM(sch_id)) < 3) sch_id = LPAD(RTRIM(sch_id),3,'0') . EXECUTE .

The code above is designed to make all ID variables three characters in length. By understanding this example, you can make your ID variable any length that you like. The first step to modifying the above syntax is to replace all instances of the example variable, *sch_id*, with the name of your SPSS variable. The first part of the **IF** statement, **IF (LENGTH(RTRIM(sch_id)) = < 3)**, evaluates the length of the variable with trailing spaces deleted. There are two functions used in this statement: the **LENGTH** and the **RTRIM** functions. The **RTRIM** function removes any trailing space characters from the variable values. This is important because although a variable may only have one digit, if the maximum length of the variable is three, there will be two additional spaces following the character, making it three characters long although it may appear to be only one character in length. The other function used in the **IF** statement is the **LENGTH** function, which returns the number of characters of the variable (in the present example it is the number of characters stripped of trailing spaces). When the condition in the **IF** statement is true, the second part of the statement is executed. For example, when the value of *sch_id* is 12, which is less than three characters in length, the condition in the first line is true because the value 12, is 2 characters in length; thus, the second part of the statement is executed. The second part of the statement, **sch_id = LPAD(RTRIM(sch_id),3,'0')**, uses the **LPAD** function to concatenate zeros on the *sch_id* variable which is again striped of any trailing blank with the **RTRIM** function. The **LPAD** function has three arguments: (1) the variable name, which, in this case is the *sch_id*, is stripped of trailing spaces, (2) the length of the new variable which should be the maximum length of the ID variable, and (3) the character that is to be concatenated to the ID variable, which is a single zero in this example.

After you have created your ID variable, the next step is to create a new file for each of the levels in your analysis. You can use your complete data file as it is for the level-1 dataset, but will need to aggregate data for higher levels of the model. This can be done with the **AGGREGATE** procedure in SPSS that is documented under the heading, **Aggregating Data** in the "SPSS for Windows: Data Manipulation and Advanced Topics" document. There are two types of variables described in this document: break and aggregate variables. The *break variable(s)* is the variable that identifies the unit of analysis for that level of the model. For example, if your second-level ID variable is for schools, then this is your break variable. The *aggregate variables(s)* are any other variables that you wish to include in your file in a summarized form. For example, if one of your variables is the socioeconomic status of students, you may wish to create an aggregated

version of this variable that is the average socioeconomic status for each school.

2.2 Creating the SSM File for HLM

To create the SSM file, go to the *File* menu and select the following option:

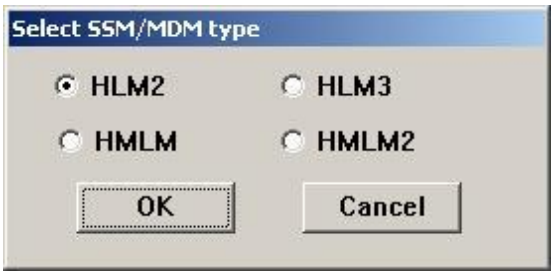
File

SSM...

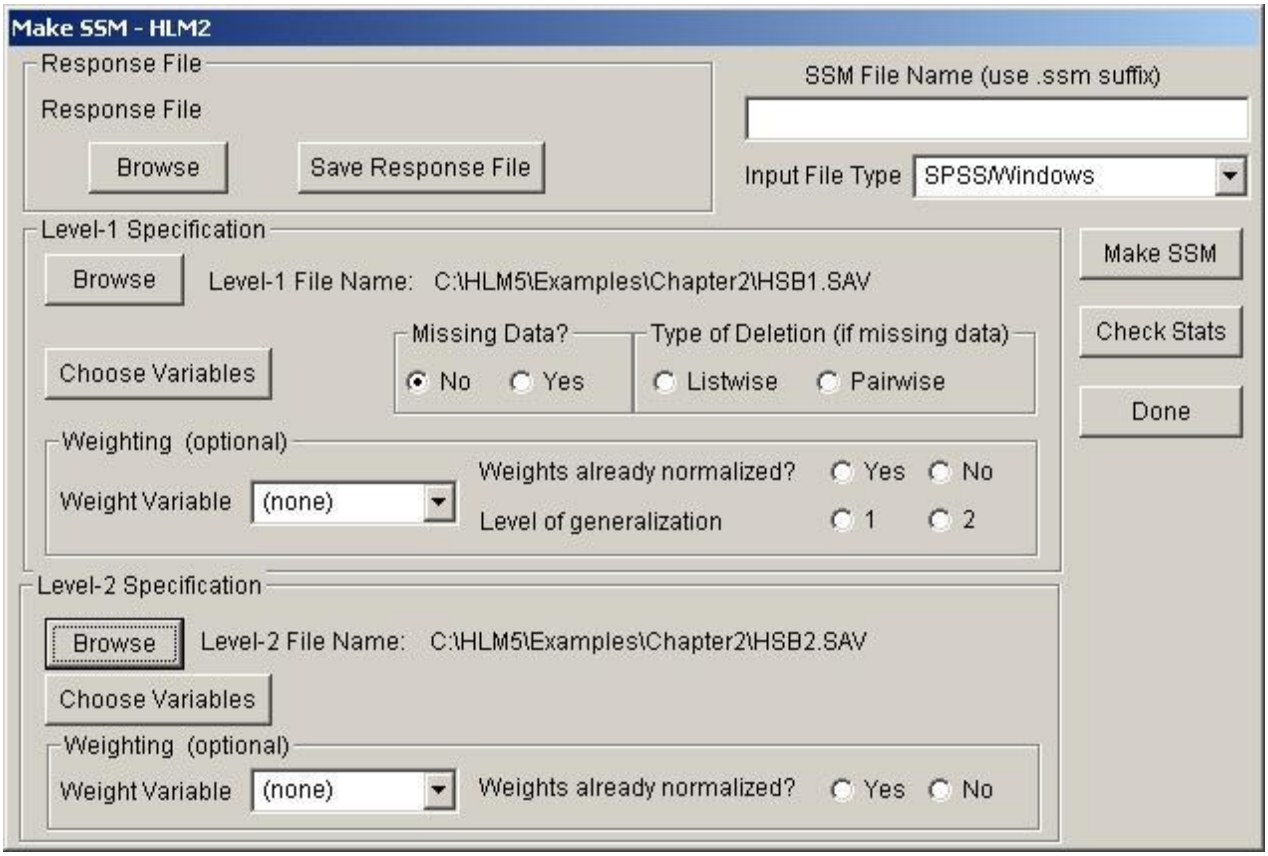
New...

Stat package input

Doing so will produce this dialog box:

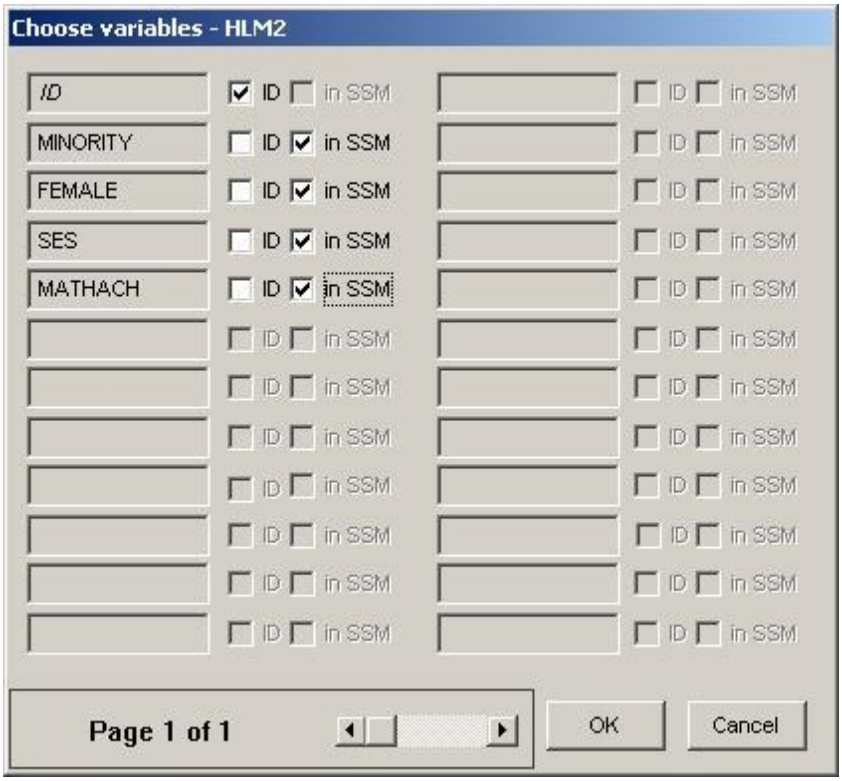


This dialog box gives you the option of creating one of the four file types that can be used in HLM. Each of the four file types is associated with a model type: (1) HLM2 is a two-level hierarchical linear model, (2) HLM3 is a three-level hierarchical linear model, (3) HMLM is a hierarchical multivariate linear model, and (4) HMLM2 is a multilevel, multivariate linear model. The present example uses a two-level model, and thus, the HLM2 option is the appropriate choice. After selecting this option, click **OK** to produce the following dialog box:



The present example creates an SSM file out of two SPSS files that are available with the HLM software. The level-1 file is *HSB1.SAV* and the level-2 file is *HSB2.SAV*. Both are located in the *Chapter2* subdirectory in the *Examples* subdirectory of the *HLM5* directory. The first step for importing files is to specify your level-1 and level-2 datasets. HLM can construct an SSM file out of first and second level datasets that are stored in one of several file formats, including SPSS, SYSTAT, and SAS 5 transport files, in addition to ASCII files. Start by selecting the file type of the files you are importing in the *Input File Type* box. After you have specified the file type, click the **Browse** button in the section labeled *Level-1 Specification* and select your level-1 dataset. Next, repeat this for the level-2 dataset by clicking the **Browse** button in the *Level-2 Specification* section.

After you have specified your datasets, you need to identify the variables that will be included in your SSM file. Click **Choose Variables**, which will produce the following dialog box:



The two columns of check boxes next to the variable names are used to select the variables you wish to include in your SSM file. Only one ID variable should be specified. The ID variable is used to match units in the level-1 file with their level-2 units. In the above example, the variable *ID* is the ID of the school rather than the ID for individual students. The school ID is the unit of analysis for the second level, and is used to link students with a particular school. After selecting the appropriate ID variable, choose any other variables that you wish to include in your SSM file by clicking the *in SSM* box to the right of the variable name. The same process is repeated for the level-2 model.

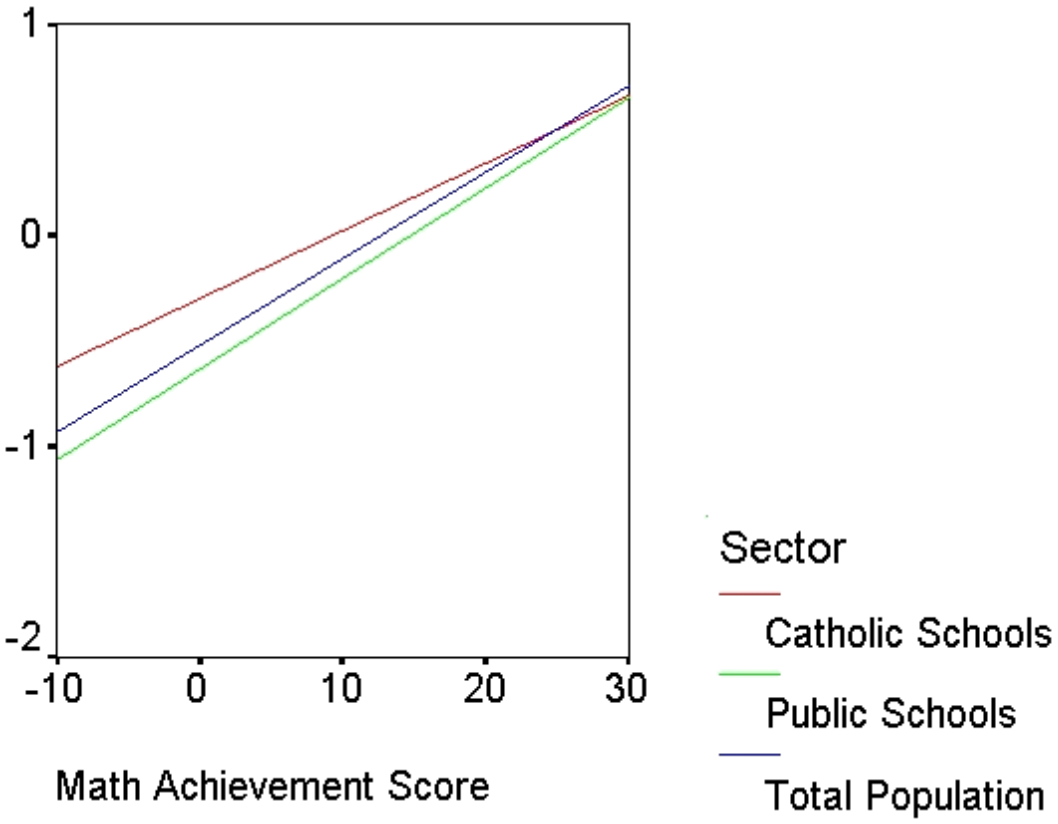
To complete the process, you need to create a *response file*, which is a file that contains the information you have entered in the dialog boxes to create the SSM file. Create the response file by clicking on the **Save Response File** button and assigning the file a name. You also need to assign a name to the SSM file by typing its name in the *SSM File Name* box. To verify that your SSM file contains the data you intended it to contain, click **Check Stats**. This will open a text file that contains statistics for the variables you have included in your SSM file. Finally, click **Make SSM** to create the SSM file. You can click **Done** to begin data analysis using HLM (although you will first be warned to check the .STS file which can be done by clicking **Check Stats**, and you will have to click **Done** a second time before you enter the HLM analysis environment).

Section 3: Introduction to Hierarchical Linear Models

3.1 Levels of a Model

Hierarchical linear models derive their name because they are designed to analyze data in which lower level units of analysis are nested in higher-level units of analysis. For example, students are nested within classrooms, which are nested within schools. While experimenters are often not interested in the effects of a particular classroom or school when they are examining the effects of a classroom intervention, these units potentially have an effect on the outcome of the study that should be accounted for in a statistical model. While designed experiments can counterbalance to control for the effect of variables in which the experimenter is not interested, it is not possible to counterbalance in studies conducted in naturalistic settings. For example, it is not possible to have one classroom situated within one school for half of the experiment and within the other school for the other half of the experiment; thus, the effect of a particular classroom cannot be counterbalanced. Hierarchical linear models are useful in these situations where the traditional experimental design cannot be used in its most ideal form.

Although it may be apparent why one would be concerned with the effects of variables such as classrooms and schools, it is likely not apparent why a special type of model is necessary. Why not just control for these variables by including them as predictors in a regression equation? On the surface, entering variables such as classroom and school in a regression equation appears to be a good solution: by entering these variables, a model can be constructed in which only the unique effect attributable to the classroom intervention is being measured. However, the key deficiency with this approach is that every student is nested within a particular classroom and each classroom is nested within a particular school, a situation that cannot be modeled with standard regression techniques. By treating students as independent, differences between schools can potentially be obscured. For example, the relationship between family income and grades may be different in public schools, where a wider range of socioeconomics is represented, as compared to private schools where all students are likely to minimally be of a fairly affluent socioeconomic status. In contrast to regression models, multilevel models first make predictions about higher-level parameters, which are then used to make predictions about lower level parameters. By modeling the levels of the model, the variance associated with levels of the model, such as the variance associated with school and the variance associated with individual students, can be separated for both the level-1 intercept and slope. An illustration of this can be seen in the figure below which plots the regression slopes of Catholic and public schools, illustrating that there are differences in the slopes and intercepts of two types of schools and that a single regression line may not be appropriate for modeling the relationship between SES and math achievement scores.



3.2 Fixed and Random Effects

Understanding the distinction between fixed and random effects is critical for the study of hierarchical models. *Fixed effects* are defined as being the only levels of a variable in which an experimenter is interested in studying. *Random effects* are effects that are a subset of the total possible levels of a variable where the experimenter is interested in generalizing to levels not observed. For example, consider a classroom intervention experiment in which there is a control group that receives no special treatment and an experimental group where teachers are given special training. A variable representing students' membership in one of the two groups would be considered a fixed effect because the two levels (control and experimental) are the only two possible levels of the variables. The same experiment would also include a random effect, as there is more than one classroom involved and these classrooms represent a subset of all possible classrooms to which the investigator would like to generalize the findings (i.e., the experimenter has sampled from a number of possible levels of the variable of interest).

As mentioned above, hierarchical models can model nested data, while traditional techniques such as standard regression cannot. Hierarchical models do this by predicting parameters using separate regression equations at each level of the model to predict parameters of variables at lower levels of the model. A hypothetical example illustrates how and why this is done. In the intervention example discussed previously, it is possible that the intervention has a much greater effect on disadvantaged children than it does on privileged children (a reasonable scenario as privileged children are likely to have had stimulating environments and therefore have less room for improvement). Accordingly, the differing effects of being in the intervention group would produce different parameters for the intercept and slope in a regression equation. The intercept

would be lower for children in the disadvantaged group, as their starting values for academic achievement are lower on average. Given that there is greater improvement for disadvantaged children, their slope, or increase in performance, is likely to be much greater. In other words, being a disadvantaged child in the experimental group results in a greater increase in the outcome relative to other disadvantaged children, as opposed to the privileged children, who have little room for improvement.

This example illustrates an important feature of hierarchical models: they take into account the fact that there are separate intercepts and slopes for higher levels in a model. Thus, information about higher levels, such as classroom and school, can be used to predict the slopes and intercept parameters of variables in lower levels in the model, such as individuals' SES.

3.3 Assumptions and Determining Sample Size

There are several assumptions about your data that you should consider prior to conducting a hierarchical linear model analysis. Bryk and Raudenbush (1992) identify five assumptions that should be met:

- The error term of each level-1 unit should have a mean of zero and the residuals should be normally distributed. For example, if the level-1 units are students and level-2 units are classrooms, then the mean of the error within each classroom should be zero, the residuals should be normally distributed, and all classrooms should have variances equal to the other classrooms in the sample.
- Level-1 predictors are independent of the level-1 error term. That is, the covariance between the level-1 predictors and the error term should equal zero.
- Level-2 error terms each have a mean of zero and adhere to a multivariate normal distribution.
- Level-2 predictors are independent of all level-2 error terms. Thus, all variables in the second level of the model are not related to any of the error terms on that level of the model, including the error term for the level-1 intercept, and the error term for any of the slopes of level-1 variables.
- The level-1 error terms are independent of level-2 error terms. That is, there is not relationship between the error term at level-1 and the error term in the level-2 equation for the level-1 intercept, or the error term in any of the equations used to estimate the slopes of level-1 variables.

The size of your sample is best determined through conducting a power analysis. As with other statistical models, statistical power is a function of three factors: the sample size, the variance in the sample, and the size of the effect being studied. Currently, HLM does not feature power analysis, however, there are some resources available for determining power. One resource is a freeware application, PINT (Power analysis IN Two-level designs) that can be downloaded from the Web.

Section 4: Two-Level Models

A two-level model consists of level-2 units in which the level-1 units are nested. For example, students, the level-1 unit of analysis, can be nested within schools, the level-2 unit. In the sample dataset created in the previous section, the units of observation in the first level are students that are grouped within the second level unit, school. In this example, we are examining the relationship between SES and math achievement for students from 160 schools. By using a level-2 model, we can determine whether students from different schools show systematic differences in the strength of the relationship between SES and math achievement.

4.1 Level-1 Model

After you have created an SSM file, you can set up your model in HLM. The first step is to specify your level-1 model. To do this, first open the previously constructed SSM file by selecting the *Old* option from the SSM submenu of the *File* menu:

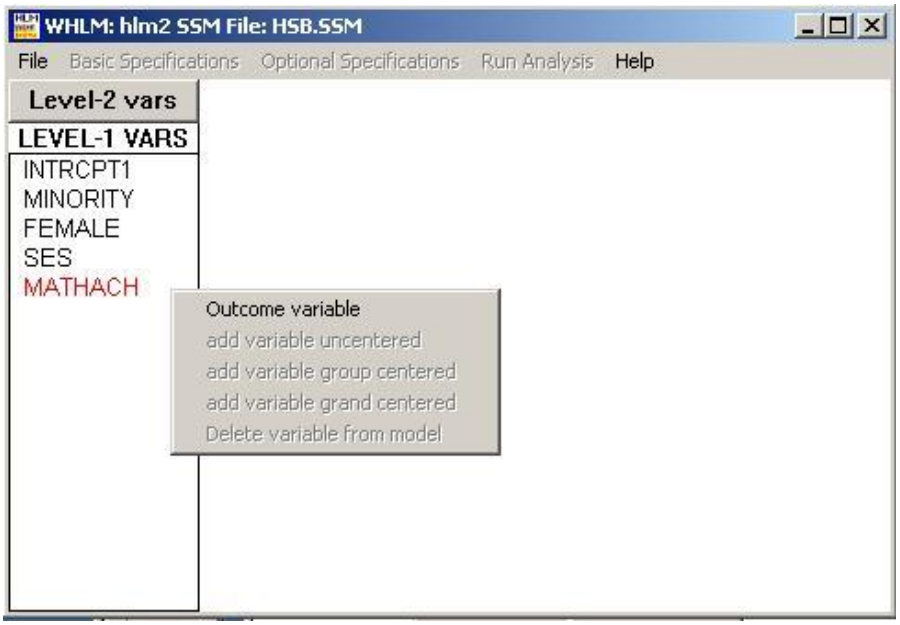
File

SSM...

Old

This will produce a dialog box that allows you to select SSM files. In the present example, the SSM file, *HSB.SSM*, was selected from the *Chapter2* directory within the *Examples* directory that resides within the *HLM 5* directory. When you have opened the file, there will be a new column on the left side of the HLM window containing the names of the variables in the level-1 dataset, as seen in the figure below.

The first step in specifying the level-1 model is to designate a dependent variable. To do this, click the name of the dependent variable in the list of level-1 variables on the left side of the HLM window. This will produce the pop-up menu as is shown in the example below.



The *outcome variable* is the level-1 dependent variable; in this case, we designate *mathach*, or math achievement scores, as the dependent variable. Once you have designated a variable as an outcome variable, you can begin adding independent variables. When you designate an independent variable, there are three choices which should be considered carefully as, although they do not affect the outcome, they do affect the interpretation of the intercept in level-1 models. Each of these choices is discussed below with regard to its effect on the level-1 intercept:

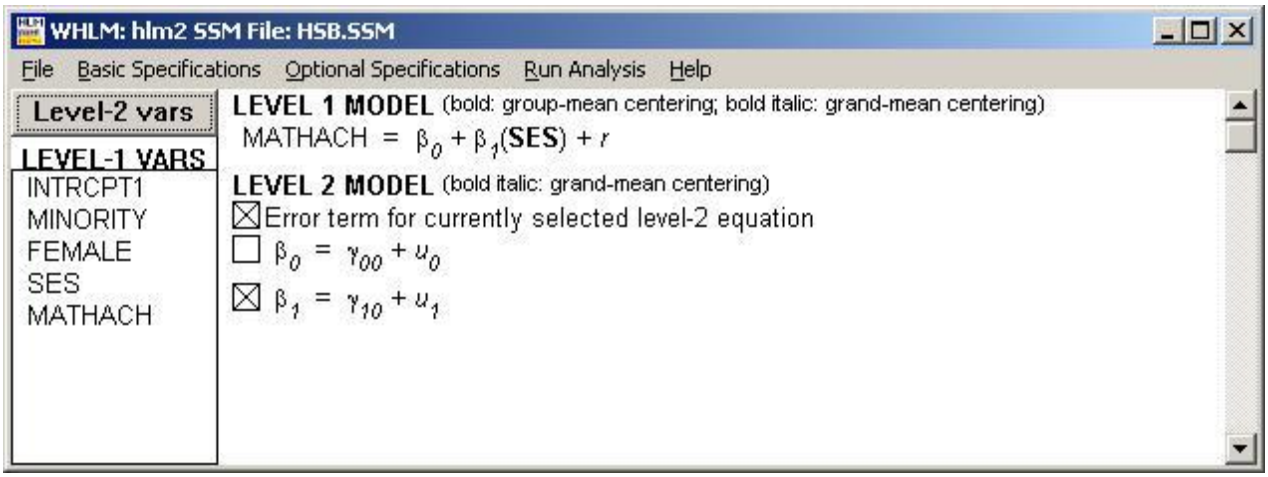
- *Variable uncentered*: This option uses predictors in their natural metric. This method is appropriate in situations in which your independent variable has a restricted range. For example, IQ scores are never 0 and are rarely even close to 0. You should always use this option when the independent variable is a dummy coded variable (it only takes a value of 0 or 1) as the mean of a dummy variable is not meaningful.

- *Variable grand-mean centered*: Independent variables are centered around a grand mean by subtracting each participant's value on the independent variable from the mean of that variable across the mean of all other participants in the study. When grand mean centering is used, the intercept is interpreted as the predicted score of an individual whose value for that independent variable is equal to the grand mean. For example, if the grand mean of a test is 70, a person with a test score of 70 would have a grand-mean centered value of 0 which is equal to the intercept. This option is useful when you want the intercept of the model to provide information about mean differences as predicted by the independent variables.

- *Variable group-mean centered*: Independent variables are centered around the mean of their level-2 group. For group mean centered variables, the intercept is the mean of the outcome at a given level-2 unit in the study's sample. For example, if you examined a school with a mean score of 72 on a test score and a second school with a mean of 68, the students within those schools would have different group centered scores for the same test score. An individual student's score of 72 in the first school would result in a group-

mean centered value of zero for the independent variable as the score is identical to the mean of the group, but were this student in the second school that had a mean of 68, this would result in a value of 4 as the score is above the average for that school.

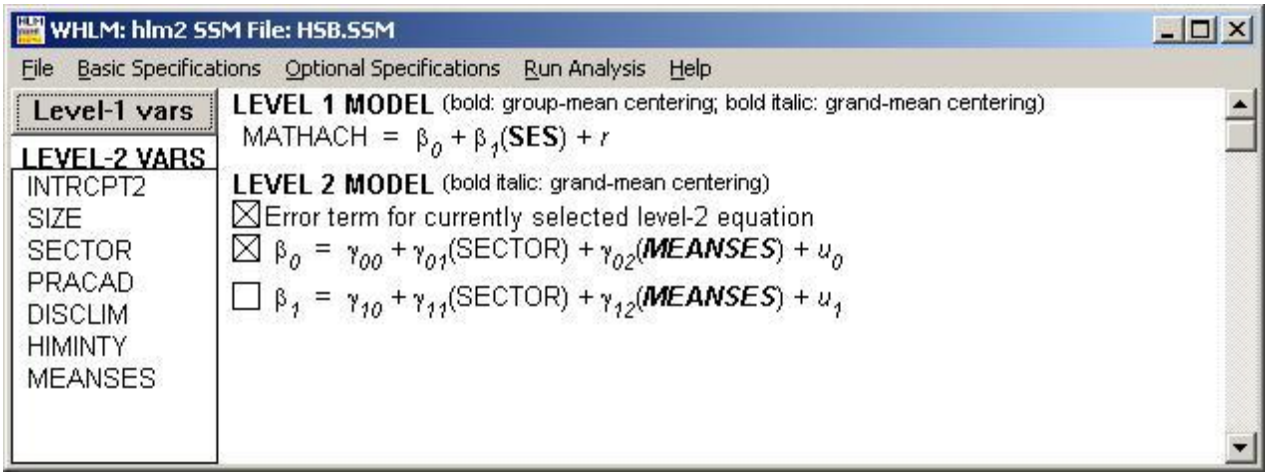
In the present example, *SES* is selected as a group-mean centered independent variable. As you can see below, the level-2 variable is bolded, indicating that it has been group-mean centered. Thus, the level-1 model specifies that students' individual math scores are a function of (1) math achievement scores at the school's average SES, and (2) the students' deviation from his or her school's average SES. The *Error term for currently selected level-2 equation* checkbox controls whether the level-1 terms are treated as fixed or random and is discussed in greater detail in the next section. The selection of the choices in the above example will result in the following model being displayed in the HLM window:



4.2 Level-2 Model

The level-2 model can be specified by first clicking on **Level-2 vars**. In the present example, the units in the level-2 data are the schools to which the students in the level-1 model belong. This will result in the level-2 variables being displayed in the variable list rather than the level-1 variables. Clicking on the name of a variable will produce a menu with two options for independent variables: *add variable uncentered* and *add variable grand centered*. These options have the same meaning as in the level-1 model.

Two variables are entered into the model in the present example for estimating both the intercept and slope of the level-1 model: *Meanses*, the average SES for each school, and *Sector*, an indicator variable that contains a value of 0 if the school is in the public sector or a value of 1 if the school is a Catholic school. *Meanses* is grand-mean centered whereas *Sector* is uncentered in both cases. These two variables need to be entered separately for the slope and the intercept. To enter variables for the intercept, B0, click the box next to this term, and then enter variables as described. Repeat this process for the slope, B1, or in the case of multiple level-1 predictor variables, for each of the slope terms. Note that it is not necessary to have the same variables predicting the intercept and slope as is shown below. The example model appears as follows after the level-2 variables have been entered:



The final step to setting up your model is to specify level-1 coefficients as random or fixed. The check box labeled, *Error term for currently selected level-2 equation* controls whether level-1 coefficients are treated as random or fixed. The default is that both the intercept and slope are treated as random coefficients. To designate either the intercept or slope as fixed, first click the box next to the intercept term, B0, and then click on the X in the *Error term for currently selected level-2 equation* box. When you do this, the error term (*u*) will disappear. This process can be repeated to remove the error term for the slope. Here, we have retained the B0 and B1, as random parameters because the addition of the level-2 parameters result in the following conceptual model: (1) a school's sector status and average SES impact the school's overall math achievement scores, and (2) a school's sector status and average SES impact the relationship between individual students' SES and math achievement scores.

After you have specified your model, you can run the analysis by clicking on the *Run Analysis* menu item at the top of the window. When you select this menu item, you receive a dialog box that gives you the option to save your analysis or to run it as is. If you wish to save your model, click **Save** and assign a name to the file. This will save the HLM command file that generates the analysis that you have just specified and can be used to rerun your model at a later time. If you click **Run the model shown**, the model will run immediately. One frequent problem that is frequently encountered at this point is a failure of the model to converge.

4.3 Output

The output for the most recently executed model can be viewed by selecting *View Output* from the *File* menu:

File

View Output

Doing so for the example in the previous section produces the output shown below. This output is described in this section with the intent of covering the output in its entirety so that you understand what each element of the output represents. However, this section is not intended to provide interpretation of output, as interpretation of selected portions of the output are discussed and interpreted in subsequent sections. Thus, this section contains little interpretation of the

output; instead it focuses on providing you with an overview of all of the output, much of which is not interpreted until subsequent sections. The output begins with the *Summary of the model specified* section:

```
Summary of the model specified (in equation format)    -----------------------
----------------------------  Level-1 Model    Y = B0 + B1*(SES) + R   Level-
2 Model    B0 = G00 + G01*(SECTOR) + G02*(MEANSES) + U0    B1 = G10 +
G11*(SECTOR) + G12*(MEANSES) + U1   Level-1 OLS regressions     -------------
----------    Level-2 Unit     INTRCPT1         SES slope    -----------------
-------------------------------------------------------------
1224      9.71545         2.50858           1288     13.51080         3.25545
          1296      7.63596         1.07596           1308
16.25550         0.12602           1317     13.17769         1.27391
1358     11.20623         5.06801           1374      9.72846         3.85432
          1433     19.71914         1.85429           1436
18.11161         1.60056           1461     16.84264         6.26650      The
average OLS level-1 coefficient for INTRCPT1 =     12.62075   The average OLS
level-1 coefficient for      SES =      2.20164
```

The above output reports the ordinary least-squares (OLS) coefficients for the first ten level-2 units (schools) in the sample. The equation for this model is shown above: Y = B0 + B1*(SES) + R. This is the standard regression equation that is computed in statistical packages when you conduct a linear regression. Here, the independent variable, *SES*, has been centered around the school mean. Thus, the intercept is the mean for each school and the *SES slope* is the change in math achievement scores that is predicted by each unit of change in socioeconomic status.

```
Least Squares Estimates    ----------------------    sigma_squared =
39.03409    The outcome variable is  MATHACH     Least-squares estimates of
fixed effects    ----------------------------------------------------------
---------------                                              Standard       Fixed
Effect         Coefficient   Error      T-ratio   d.f.     P-value    --------
-----------------------------------------------------------------------
   For       INTRCPT1, B0       INTRCPT2, G00          12.083837   0.106889
113.050      7179    0.000         SECTOR, G01           1.280341
0.157845     8.111      7179    0.000        MEANSES, G02           5.163791
0.190834    27.059      7179    0.000    For      SES slope, B1
INTRCPT2, G10           2.935664   0.155268    18.907      7179    0.000
        SECTOR, G11          -1.642102   0.240178    -6.837      7179
0.000        MEANSES, G12           1.044120   0.299885     3.482      7179
0.001    ------------------------------------------------------------------
-------      The outcome variable is  MATHACH    Least-squares estimates of
fixed effects    (with robust standard errors)    ---------------------------
-----------------------------------------------
                                          Standard      Fixed Effect
Coefficient   Error      T-ratio   d.f.     P-value    ----------------------
-------------------------------------------------    For       INTRCPT1,
```

```
B0       INTRCPT2, G00          12.083837   0.169507    71.288      7179
0.000         SECTOR, G01           1.280341   0.299077     4.281      7179
0.000        MEANSES, G02           5.163791   0.334078    15.457      7179
0.000    For      SES slope, B1       INTRCPT2, G10           2.935664
0.147576    19.893      7179    0.000         SECTOR, G11          -1.642102
0.237223    -6.922      7179    0.000        MEANSES, G12           1.044120
0.332897     3.136      7179    0.002    ------------------------------------
----------------------------------------         The least-squares likelihood
value = -23362.111325    Deviance =  46724.22265    Number of estimated
parameters =    1
```

The tables shown above use OLS to estimate coefficients. The statistics in these tables are identical to those that would be obtained from a standard OLS regression and do not take into account the random variation among the level-2 units. There are two tables in the above output with identical values in the *Coefficient* columns. The difference between the two tables is the method used to calculate the standard error. Looking only at the first table, you can see that there are two lines that begin with *For* and these two lines represent the sections for the statistics used to compute the level-1 intercept and slope. You can also see that there are identical terms used for these equations: *Intrcpt2*, *Sector*, and *Meanses*. These terms correspond to the level-2 model constructed in the previous section: in the model for the level-1 intercept, B0, and for the level-1 slope, B1, there was an intercept term and *sector* and *meanses* were entered as independent variables. To examine the relationship between an independent variable and the parameters, you look at the statistics for that variable under both the intercept and slope sections. For example, you can see that *sector* had a significant effect on both the intercept of each school's overall math achievement, $t = 8.11$, $p < .001$, and the slope on the relationship between math achievement and SES, $t = -6.84$, $p < .001$. Thus, a general conclusion that can be reached here is that there are differences between public and Catholic schools in their slopes and intercepts. In addition, schools with different SES's also vary in terms of their effect on the level-1 slope, $t = 27.06$, $p < .001$, and level-1 intercept, $t = 3.48$, $p < .001$. Specific interpretations of these results are discussed in the following section.

```
STARTING VALUES    ---------------   sigma(0)_squared =     36.72025    Tau(0)
   INTRCPT1,B0      2.56964       0.28026         SES,B1      0.28026      -
0.01614       New Tau(0)    INTRCPT1,B0      2.56964       0.28026
SES,B1      0.28026       0.43223     The outcome variable is  MATHACH
   Estimation of fixed effects   (Based on starting values of covariance
components)    ------------------------------------------------------------
-------------                                              Standard
Approx.       Fixed Effect         Coefficient   Error      T-ratio   d.f.
P-value    ---------------------------------------------------------------
---------    For       INTRCPT1, B0       INTRCPT2, G00          12.094864
0.204326    59.194       157    0.000         SECTOR, G01           1.226266
0.315204     3.890       157    0.000        MEANSES, G02           5.335184
0.379879    14.044       157    0.000    For      SES slope, B1
```

```
INTRCPT2, G10           2.935219   0.168674    17.402       157    0.000
        SECTOR, G11          -1.634083   0.260672    -6.269       157
0.000        MEANSES, G12           1.015061   0.323523     3.138       157
0.002    ----------------------------------------------------------------
-------      The value of the likelihood function at iteration 1 = -
2.325199E+004   The value of the likelihood function at iteration 2 = -
2.325182E+004   The value of the likelihood function at iteration 3 = -
2.325174E+004   The value of the likelihood function at iteration 4 = -
2.325169E+004   The value of the likelihood function at iteration 5 = -
2.325154E+004                                                        .
                                                                     .
                                                                     .   The value of the
likelihood function at iteration 57 = -2.325094E+004   The value of the
likelihood function at iteration 58 = -2.325094E+004   The value of the
likelihood function at iteration 59 = -2.325094E+004   The value of the
likelihood function at iteration 60 = -2.325094E+004   Iterations stopped due
to small change in likelihood function   ******* ITERATION 61 *******
```

The above output describes the computation of the fixed effects and their associated coefficients. While this output is important for understanding how the effects were estimated, it is not typically used to interpret hypotheses tested by your model. Thus, as the focus of this document is on testing models and interpreting results, the above output is not discussed in detail.

```
Sigma_squared =     36.70313    Tau    INTRCPT1,B0      2.37996       0.19058
        SES,B1      0.19058       0.14892     Tau (as correlations)
   INTRCPT1,B0  1.000  0.320         SES,B1  0.320  1.000    -----------------
-----------------------------------     Random level-1 coefficient
Reliability estimate    ------------------------------------------------------
    INTRCPT1, B0                         0.733          SES,
B1                        0.073    ------------------------------------------
---------   The value of the likelihood function at iteration 61 = -
2.325094E+004       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          12.095006
0.198717    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           2.937787   0.157119    18.698       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    ----------------------------------------------------------------------
-------    The outcome variable is  MATHACH    Final estimation of fixed
effects    (with robust standard errors)    ---------------------------------
----------------------------------------

                                          Standard             Approx.
Fixed Effect         Coefficient   Error      T-ratio   d.f.     P-value    --
----------------------------------------------------------------------------

   For       INTRCPT1, B0       INTRCPT2, G00          12.095006   0.173688
69.637       157    0.000         SECTOR, G01           1.226384
0.308484     3.976       157    0.000        MEANSES, G02           5.333056
0.334600    15.939       157    0.000    For      SES slope, B1
INTRCPT2, G10           2.937787   0.147615    19.902       157    0.000
        SECTOR, G11          -1.640954   0.237401    -6.912       157
0.000        MEANSES, G12           1.034427   0.332785     3.108       157
0.002    ----------------------------------------------------------------------
-------
```

The preceding output is identical in format to the OLS estimates of fixed effects shown and described above. The key difference between the two tables is that the second table used *robust standard errors* as opposed to model-based estimates of the standard errors reported in the first table. Generally, it is preferable to use the robust standard errors and the *t* ratios and *p* values associated with these standard errors. When it is inappropriate to use the robust standard errors, HLM will indicate this in a note following the output. Notice that the values in the *Coefficient* columns are identical in both tables, as the effects being estimated are identical.

```
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    ---------------------------------
----------------------------------------
```

Thus far, the values of the level-1 and level-2 intercept and slope coefficients have been discussed, but the error terms have not. The above table contains these values and the hypothesis tests that these values are equal to zero. You can see that there is an error term for each regression equation at each level of the model containing a random effect. In the table above, U0 is the error term for the level-2 intercept, U1 is the error term for the level-2 slope, and R is the error term for the level-1 equation. These tests are used to evaluate the amount of unexplained variance and can be useful for determining whether an error term is necessary. For example, the chi square value of 605.30, $p < .001$, indicates that the error term associated with estimating the value of the intercept is significantly different than zero; therefore there is significant variability among schools in their average math achievement scores. In contrast, the error term used to estimate the slope for the SES parameter has a smaller, nonsignificant chi square of 162.31, $p =$

.369 indicating that the error term for this parameter does not differ significantly from zero and could therefore be dropped from the model. That is, the relationship between SES and math achievement does not appear to differ between schools. There is not a significance test associated with the level-1 variance, but it can be seen that it has the largest standard deviation, indicating that there is a large amount of variance that is not explained by the current model.

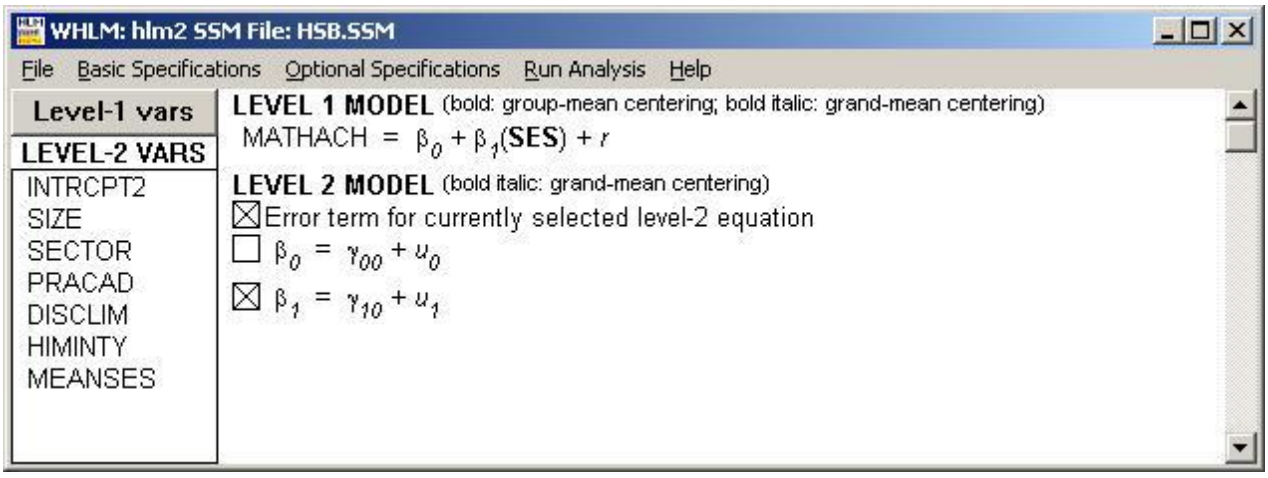
```
Statistics for current covariance components model    ----------------------
-------------------------    Deviance =  46501.87563    Number of estimated
parameters =    4
```

The deviance statistic is used in comparing models. For example, you may run one model including the effects of gender and one model with no gender effects. The deviance statistic outputted from each model can help you compare the relative utility of the two models. However, the deviance statistic is not typically interpreted on its own and is therefore not discussed here.

Section 5: Examples of Two-Level Models

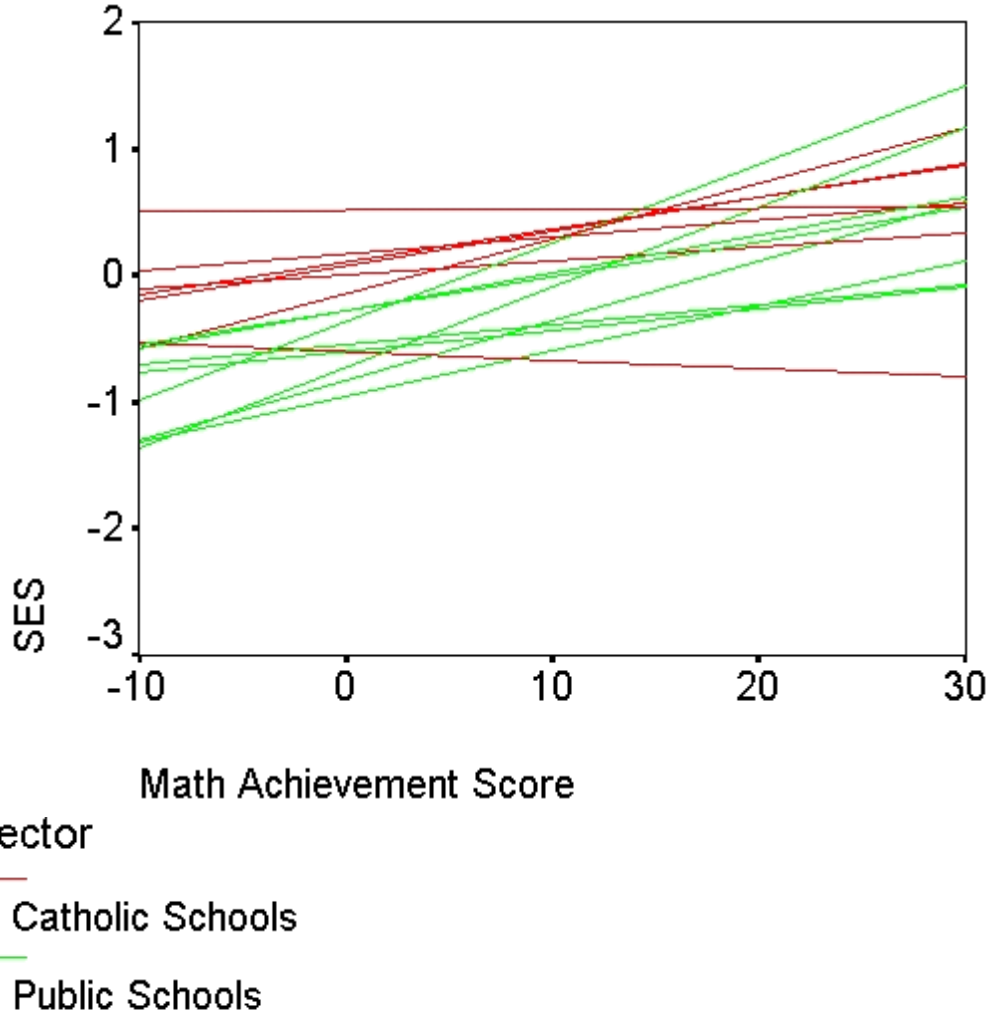
5.1 Random Coefficients

The random coefficients model derives its name for the fact that it contains random terms in the level-2 equations used to predict the level-1 coefficients. This model is used in situations where there are not any level-2 variables, but there is variation between the level-2 units and therefore, a single regression equation is not appropriate. An example of a random coefficients model can be seen in the dialog box below where there is an error term for both the intercept and slope. This error term represents the unique effect of each individual school on the slope and intercept of the level-1 model.



As each of the schools have unique intercepts and slopes there is essentially a regression equation for each of these level-2 units, as the error terms in the level-2 equations are unique to each school are used to calculate the level-1 intercept and slope coefficients. In the case of the current example, there are 160 schools, each of which has a separate regression equation on level-1. This can be illustrated by considering the regression equation for a single school. To obtain this regression equation, you would calculate the level-1 intercept, B0, by adding the

average of math achievement scores across the population of schools to the error term U0 which is the unique increment in the intercept associated with an individual school. The slope, B1, is calculated in the same manner: the average regression slope for SES and math achievement scores is added to the unique change in the slope associated with an individual school. The random coefficients model reflects the fact that each school has a separate slope and intercept and therefore it is not necessarily the case that a single regression equation is appropriate across all schools in the model. This is illustrated in the graph below that plots the regression lines for six Catholic schools and eight public schools.



The output below contains coefficients for all the parameters in the model and their associated significance tests. Only the *Final Estimation of Fixed Effects* table for robust standard errors is shown below, as it is the most appropriate for this example.

```
The outcome variable is  MATHACH    Final estimation of fixed effects    (with
robust standard errors)    ----------------------------------------------------
------------------------
Standard             Approx.       Fixed Effect         Coefficient
```

```
Error      T-ratio   d.f.     P-value    ------------------------------------
----------------------------------------    For       INTRCPT1, B0
INTRCPT2, G00          12.636197   0.243738    51.843       159    0.000
   For      SES slope, B1       INTRCPT2, G10           2.193157   0.127846
17.155       159    0.000    ------------------------------------------------
--------------------------
```

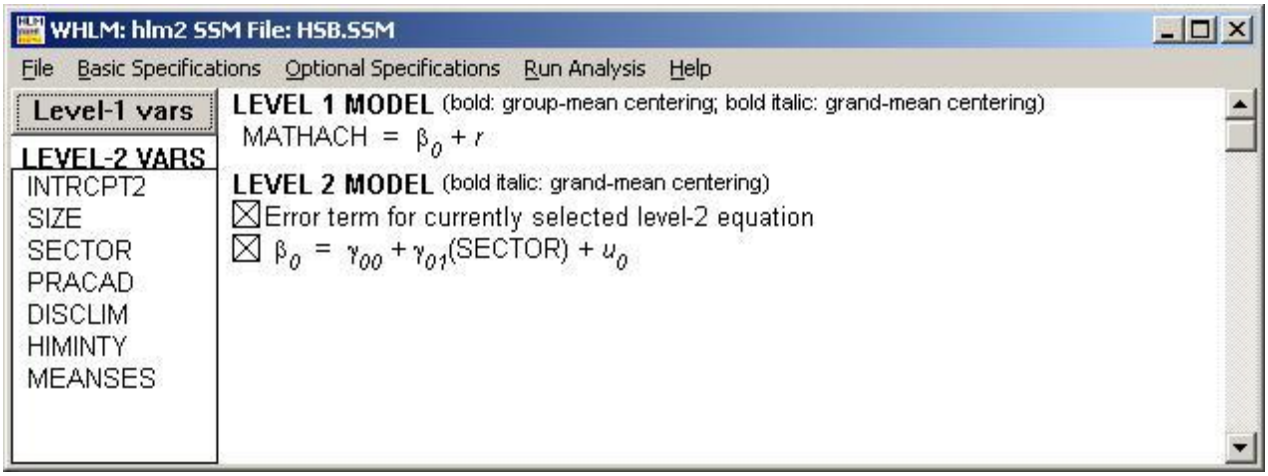
The table above is interpreted in a manner similar to a standard regression equation. The first term, B0, is the level-1 intercept and is computed with a regression equation that contains only a single term, the level-2 intercept, G00. The *t* ratio associated with this term tests the possibility that the intercept equals zero (which is typically not an interesting comparison, as it amounts to testing the hypothesis that students averaged a score of zero on their math achievement tests). The next term that is estimated, B1, is the level-1 slope; it is estimated by the term G10, which is the average slope and, similar to the intercept, the associated *t* ratio tests the possibility that this value is zero. This test is a theoretically interesting test because it examines the relationship between SES and the dependent variable: if this value were indeed zero, this would indicate that there were no relationship between SES and students' math achievement scores. In this case, the table above indicates that the slope is not zero, as the *t* ratio is 17.16, $p < .001$. Examining the value in the *Coefficient* column, it can be seen that the increase in math achievement scores associated with each unit of SES is 2.19, indicating that as SES increases, so do math achievement scores.

```
Final estimation of variance components:   ---------------------------------
----------------------------------------   Random Effect
Standard      Variance     df    Chi-square  P-value
                           Deviation     Component   ------------------------
-------------------------------------------------   INTRCPT1,
U0        2.94633       8.68087   159    1770.85120    0.000         SES
slope, U1         0.82485       0.68038   159     213.43769    0.003     level-
1,       R          6.05835      36.70356   ---------------------------------
----------------------------------------
```

The error terms are examined to determine whether it is necessary to have different regression equations for each of the schools in the model, or whether it might be sufficient to average all the error terms. That is, if there is not significant variability in math achievement across schools, or in the relationship between SES and math achievement scores across schools, then there is not any variability in the slope, then there is not a need to include an error term in the level-2 model. In such a situation, a hierarchical model is likely to not be necessary as there is not any significant variation across higher-level units that could potentially affect the interpretation of the level-1 variables. Remember that these error terms are what makes each school's regression equation unique; if there were no intercept error term, each school would have an identical intercept, and if there were no slope error term, each school would have an identical slope. In the table above, you can see that there is significant variability in both the intercept and the slope, indicating that there is variability among the schools in the model, and, therefore, these terms are important for predicting level-1 coefficients.

5.2 Intercept-as-Outcome Models

Differences in intercepts represent mean differences in the dependent variable that can be predicted from independent variables. In fact, models in which only the intercept is predicted from level-2 variables are also known as mean-as-outcome models because a difference in the intercept represents a difference in means in the dependent variable that can be predicted from the independent variables. Returning to our previous example, it is possible that it is only of interest as to whether there is a mean difference in *mathach*, the variable representing math achievement scores between the Catholic and public schools in our dataset. In this case, you would construct a model with *mathach* as the dependent variable and *sector*, the variable representing whether a school is public or Catholic, as the only level-2 predictor. The HLM model appears as shown below:



Running this model produces the following output:

```
The outcome variable is  MATHACH    Final estimation of fixed effects    (with
robust standard errors)    ----------------------------------------------------
-------------------------
Standard              Approx.       Fixed Effect         Coefficient
Error      T-ratio    d.f.      P-value    ------------------------------------
---------------------------------------    For       INTRCPT1, B0
INTRCPT2, G00          11.393044   0.292258    38.983       158    0.000
        SECTOR, G01           2.804889   0.435823     6.436       158
0.000    ------------------------------------------------------------------------
-------
```

The *t* ratio represents the test that the intercepts (the average math achievement score) are equal across two types of schools. Examining the output, you can see that there is a significant *p* value for *sector*, indicating that intercepts differ across the two types of schools. This represents the test that the intercepts, or means, are equal. As the coefficients in this model are estimated using the restricted maximum likelihood (REML) estimation algorithm, the coefficients do not represent the mean difference exactly as does the output that uses ordinary least squares to estimate coefficients. However, the coefficients are quite close. To illustrate this, consider that the mean score for *mathach* for public schools is 11.36, as opposed to Catholic schools whose mean *mathach* score is 14.17; the mean difference between these two values is 2.81 which

closely approximates the estimated difference between the groups as seen in the coefficient for sector. As public schools are coded as 0 in the example dataset, their predicted value is the intercept of the level-2 model, which is 11.39. The Catholic schools, which are coded as 1, have a predicted value of the intercept plus the change in one unit of the independent variable, *sector*, and thus have a predicted value of 14.20. Comparing the predicted values and the observed means, you can see that the above model does a good job of estimating the mean differences in *mathach*.

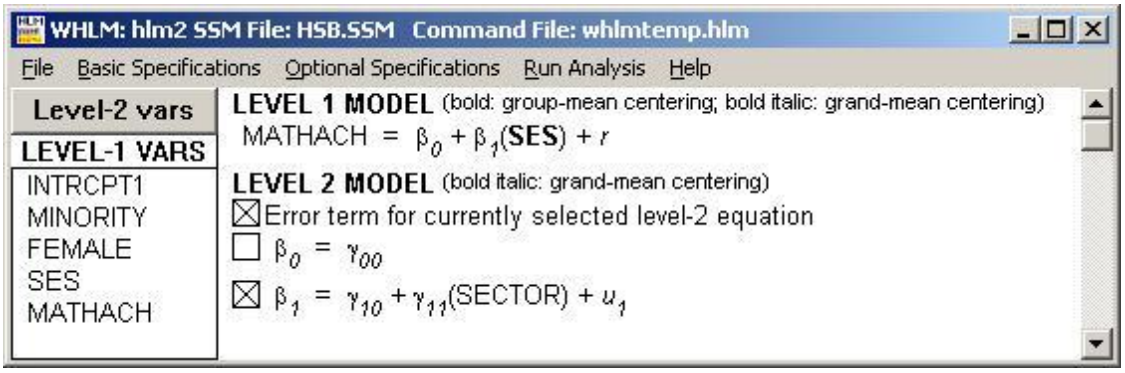
```
Final estimation of variance components:   ---------------------------------
-----------------------------------------   Random Effect
Standard      Variance     df    Chi-square  P-value
                           Deviation     Component   ------------------------
-------------------------------------------------------   INTRCPT1,
U0        2.58413       6.67771   158    1296.76563    0.000     level-
1,       R         6.25710      39.15135   ---------------------------------
-----------------------------------------
```

The table above summarizes the variance components for the intercepts-as-outcomes model. The chi square value of 1296.76 is quite large and the *p* value is less than .001, which indicates that error term for the level-2 equation used to predict the slope is significantly different from zero. This tells you that there is significant variation among the level-2 units (schools). Thus, a traditional regression that ignored the effects of individual schools would be misspecified, and therefore incorrect. Indeed, a comparison between the HLM output and an OLS regression equation shows that the standard error and *t* values from the OLS substantially overestimate the effect of sector when the school level variation is ignored: the OLS results produce a standard error of .159 and a *t* value of 17.66, in contrast to .436 for the standard error and a *t* value of 6.436. Thus, it is essential to include the error term in the model in order to avoid overestimating the effects of level-2 independent variables.

5.3 Slopes-as-Outcomes Models

In addition to examining questions about differences in the intercept that can be predicted by level-2 variables, hierarchical models are also used to examine differences in slopes for a dependent variable (which represent the change in the dependent variable that can be accounted for by the independent variable). Returning to the above example, one feasible research question is whether the relationship between SES and math achievements scores is the same across public and Catholic schools. In other words, it is already known that Catholic schools have a higher math scores than do the public schools, but it is possible that these scores are a function of socioeconomic status.

This model would have a level-1 regression equation in which the group-centered value of the variable *ses* was used to predict *mathach*. The level-2 equation is somewhat more complex as there is now an equation for modeling both the slope and the intercept. In this model, the only effect of interest is the effect of *sector* on the intercept of the level-1 equation, so *sector* is added into the equation used to predict the slope for the level-1 model. The HLM model takes the following form:



Running the model shown above produces the following output:

```
The outcome variable is  MATHACH    Final estimation of fixed effects    (with
robust standard errors)    ----------------------------------------------------
-------------------------
Standard             Approx.       Fixed Effect         Coefficient
Error      T-ratio   d.f.     P-value    ------------------------------------
---------------------------------------    For       INTRCPT1, B0
INTRCPT2, G00          12.637854   0.243595    51.881       159    0.000
   For      SES slope, B1       INTRCPT2, G10           2.972029   0.154085
19.288       158    0.000         SECTOR, G11          -1.727501   0.224486
-7.695       158    0.000    ----------------------------------------------------
--------------------------
```

In contrast to previous examples interpreted in this section, there are two level-1 coefficients that are being estimated: the level-1 intercept, B0, and the level-1 slope, B1. There are no independent variables or error term used to estimate the level-1 intercept in this example, therefore this term is set as the grand average of math achievement scores across both sectors. The first term, G00, indicates that on average, schools have math achievement scores of 12.63. This term does not provide any statistics of theoretical interest to the current example; the reported *t* ratio is simply a test of the null hypothesis that the average math achievement score is zero, however, the intercept must be included as it is important for interpreting the remaining coefficients in the model. The term G10 indicates that for public schools, and for each unit of increase in SES, there is a resulting increase in *mathach* of 2.97. The term G11 indicates that there is a decrease in the strength of the relationship between SES and *mathach* for students that are attending private Catholic schools relative to public schools.

The coefficients for the level-1 slope, G10 and G11, show that there is a statistically significant effect for both the intercept and the slope used to predict the level-1 slope. Both of these provide theoretically interesting information about the relationship between SES and and math achievement scores. First, consider the regression equation for computing the coefficients: B1 = G10 + G11(*sector*) + error. If you replace sector with the values of that variable, there are two possible outcomes to that equation as sector is a 0 in the case of public schools and 1 in the case of Catholic schools. Thus, for public schools, B1 = 2.972 + (-1.728) (0), resulting in a value of 2.972, and for Catholic schools, B1 = 2.972 + (-1.728) (1), resulting in a value of 1.244. This equation results in a positive slope for both sectors, indicating that as SES increases, so do math

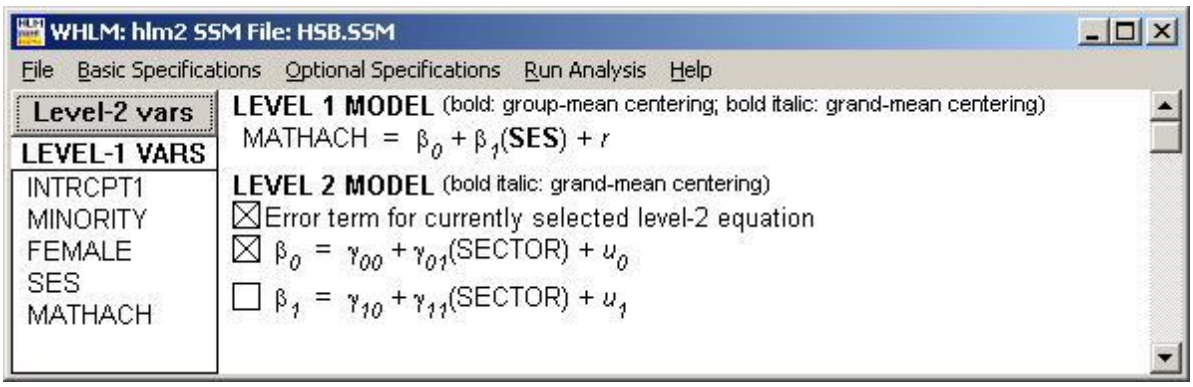
achievement scores. However, completing the regression equation also reveals that the public schools have a steeper slope: for each unit increase of SES, there is a 2.972 increase in math achievement scores for public school, whereas for the Catholic schools, there is a 1.244 increase in math achievement scores for each unit increase in SES. Furthermore, G11, which is the change in the level-1 slope attributable to the level-2 slope for *sector*, reveals that *sector* has a statistically significant effect on B1 as indicated by the large *t* ratio, -7.695 ($p < .001$). In terms of the theoretically interesting questions that are implicit in this model, this tells us that there is a statistically significant difference in the relationship between SES and math achievement scores across sectors. The fact that students who attended public schools have a stronger relationship between SES and math achievement supports the conclusion that SES is better predictor of math achievement score in the public schools than in Catholic schools. This reflects the fact that there is a good deal of disparity in math achievement scores between students with lower SES and students with higher SES in the public schools whereas students in the Catholic school, who generally tend to have a higher SES, do not show as strong a relationship between these variables. SES has a stronger relationship with math achievement among public schools than among Catholic schools.

```
Final estimation of variance components:   ---------------------------------
-----------------------------------------   Random Effect
Standard      Variance     df    Chi-square  P-value
                           Deviation     Component   ------------------------
-----------------------------------------------------   INTRCPT1,
U0        2.94579       8.67765   159    1771.42510    0.000         SES
slope, U1        0.57724       0.33321   158     178.33005    0.128     level-
1,       R          6.05724      36.69017   ---------------------------------
-----------------------------------------
```

You can see that the error term in the equation used to predict the level-1 slope is not significantly different from zero as indicated from the small chi square value, 178.33, relative to its critical chi square value of 157 and the large *p* value, .128. Thus, it could be concluded that there is not a significant difference between schools in their relationship between *sector* and their level-1 slope. Given the lack of variation across schools, it is conceivable that the error term, U1 could be dropped from the analysis.

5.4 Random Slopes and Intercepts

The random slopes and intercepts model combines the two previous models so that both mean differences in *mathach* and the differences in slope can be evaluated across sectors. In this model, the level-l coefficients for both the intercept and slope are predicted based on information about the sector to which a school belongs. To construct a model that addresses hypotheses about differences between sectors in intercepts and slopes, you need level-2 equations that include both intercepts and slopes for both the level-1 slope and intercept. The following dialog box provides an example of such a model:



Running the above model produces the following output:

```
Final estimation of fixed effects    The outcome variable is  MATHACH    (with
robust standard errors)    ----------------------------------------------------
-------------------------
Standard             Approx.       Fixed Effect         Coefficient
Error      T-ratio   d.f.     P-value    ------------------------------------
---------------------------------------    For       INTRCPT1, B0
INTRCPT2, G00          11.393837   0.292348    38.974       158    0.000
        SECTOR, G01           2.807465   0.435633     6.445       158
0.000    For      SES slope, B1       INTRCPT2, G10           2.802449
0.157937    17.744       158    0.000         SECTOR, G11          -1.340634
0.230324    -5.821       158    0.000    ------------------------------------
---------------------------------------
```

Examining the above output, the information about the intercept, B0, is nearly identical to the intercepts-as-outcomes model discussed in the previous section: it shows us that there is a significant amount of variance attributable to differences between schools. The information about the slope, B1, shows similar results as were seen in the intercepts-as-outcomes models. As in the previous examples, the parameters for level-1 coefficients can be calculated by substituting the level-2 values into the level-2 regression equation. Thus, the public sector's slope can be estimated by adding the intercept to the slope multiplied by the sector variable to obtain the public school's value for B1: 2.80 + (-1.34) (0) = 2.80. The change in math achievement scores that can be attributed to each unit of SES among public schools is 2.80. The slope for the Catholic schools can be calculated in the same manner: the intercept is added to the slope multiplied by the sector variable to obtain the Catholic school's value for B1: 2.80 + (-1.34) (1) = 1.44. You can see that the coefficient for slopes differ across public and private schools and that public schools have a greater increase in math achievement per unit of SES than do Catholic schools. Thus, students in Catholic schools have higher math achievement scores. Also, where difference in SES exist, they have a stronger impact on kids in public schools than kids in Catholic schools.

In addition to the coefficients, there is important information in the *T-ratio* columns that helps us evaluate whether the differences in slopes and intercepts across sectors are statistically significant. For both B0 and B1, there are intercept terms, G00 and G10. The level-2 intercept

term, G00, for the level-1 intercept, B0, is typically not of interest as it tests whether math achievement scores are significantly different from 0. The *t* ratio for the level-2 slope for the level-1 intercept is more interesting as it tests the hypothesis that the increment in the slope that can be attributed to the variable *sector* is zero. In fact, this hypothesis is rejected; the *t* ratio, -5.821 ($p < .001$), is significant, indicating that the difference in math achievement scores and SES relationship between the two types of schools is not likely to occur by chance.

```
Final estimation of variance components:   ---------------------------------
-----------------------------------------   Random Effect
Standard      Variance     df    Chi-square  P-value
                           Deviation     Component   ------------------------
-----------------------------------------------------   INTRCPT1,
U0         2.59609       6.73966   158    1383.78481    0.000         SES
slope, U1         0.55141       0.30405   158     175.31196    0.164     level-
1,       R          6.05722      36.68995   ---------------------------------
-----------------------------------------
```

The table above again shows that there is a significant difference in the variance between schools for the equation estimating their level-1 intercepts, indicating that there are mean differences between schools. In contrast, the error term for the slope, U1, reveals that there is not a significant difference across schools in the level-1 slopes. This means that there is not a significant difference in the relationship between SES and math achievement scores across schools. Thus, you could potentially drop the error term, U1, from the model as the variation between slopes across schools does not explain a significant portion of the variance and therefore does not differ from that which we would expect purely due to chance.

Section 6: The Hierarchical Generalized Linear Model

The HGLM component of HLM 5 is designed to accommodate models that have a categorical outcome variable (such as a yes or no response) or a count variable (one that represents the number of responses falling into a particular category). One of the principal assumptions of hierarchical linear models is that the distribution of level-1 residuals adheres to a normal distribution. This assumption is not reasonable when a dependent variable is not continuously distributed. For example, if the outcome were a yes or no response, or a response that asked participants to classify themselves into one of several possible categories, then it is not reasonable to assume that the residuals would be normally distributed. Fortunately, HLM has adopted the *generalized linear model*, an approach that allows you to model data with categorical outcomes in a manner similar to traditional linear models.

6.1 Theoretical Background

Generalized linear models are quite similar to the linear models that have been discussed thus far. The differences between generalized linear models and standard linear models can essentially be summarized by the sampling and structural models.

The *sampling model* refers to the model of the distribution of the dependent variable. While linear models assume that the dependent variables are normally distributed, categorical outcomes

may adhere to one of several possible distributions, including the Bernoulli, binomial, and Poisson distributions. In order for variables from one of these distributions to be used in a linear model, they need to be transformed using a *link function*. Link functions serve to transform the values of the dependent variables so that they adhere to linear model assumptions and the outcome is still a possible value of the dependent variable (e.g., a model with a binary outcome can only have a predicted value of 0 or 1).

The *structural model* refers to the form of the equation describing the model. The structural model results in a regression equation where the link function is substituted for the dependent variable (for categorical outcomes), or more than one probability is calculated (for models where there are more than two possible values to the dependent variable).

While understanding the essential features of the generalized linear models is important for understanding how your data are being treated by HGLM, it is only necessary to know what type of data you have . Once you have identified your sampling model, HGLM will automatically implement the proper link function for the distribution of your data and will also use the correct structural model. The following list provides an overview of the types of categorical data that can be analyzed in HGLM:

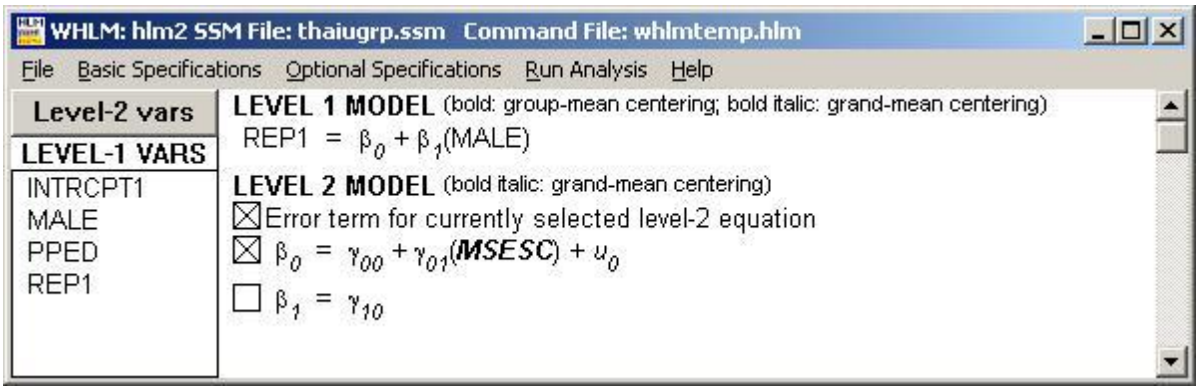
- *Binary*: a variable with two possible outcomes. The parameters of a binomial distribution are the number of trials and the probability of success. An example is the number of hits a baseball player has out of the total number of attempts at hitting the ball.
- *Bernoulli*: a binary model where the outcome is only measured once. The predicted values equal the probability of success. An example is whether a person voted yes or no on a referendum, as there is only one outcome per person.
- *Multinomial*: a multi-category response variable. For example, a questionnaire that asked you to select your favorite movie from four possible choices would be a multinomial dependent variable. Multinomial responses have a similar distribution to Bernoulli models with the key difference being that there is a probability associated with each of the possible outcomes.
- *Count*: Count data are measures of the number of times an event occurs during a period of time. Count data adhere to the Poisson distribution that has parameters for the time interval and the number of occurrences of an event. An example of count data is the number of times a person visited a doctor during a period of three years.
- *Ordinal*: Ordinal data model the probability that a response falls into an ordered set of categories. It extends the count model by treating the probabilities as cumulative. An example of an ordinal outcome variable is the ranking on a skills exam where participants were rated as novice, intermediate, and advanced.

6.2 A Hierarchical Generalized Linear Model

The process for implementing an analysis with a categorical outcome in HLM is very similar to an analysis with a continuous outcome variable. The SSM dataset, *THAIGRP.SSM* is created from the level-1 dataset, *UTHAI.SAV* and the level-2 dataset, *THAI2.SAV*, using the method as described in Section 3. The dataset contains data on 7516 sixth graders nested within 356

schools. The level-1 unit of analysis is the student while the level-2 unit of analysis is the school.

Setting up the model is done the same way as was described in Section 4. The following model will be used in this example:



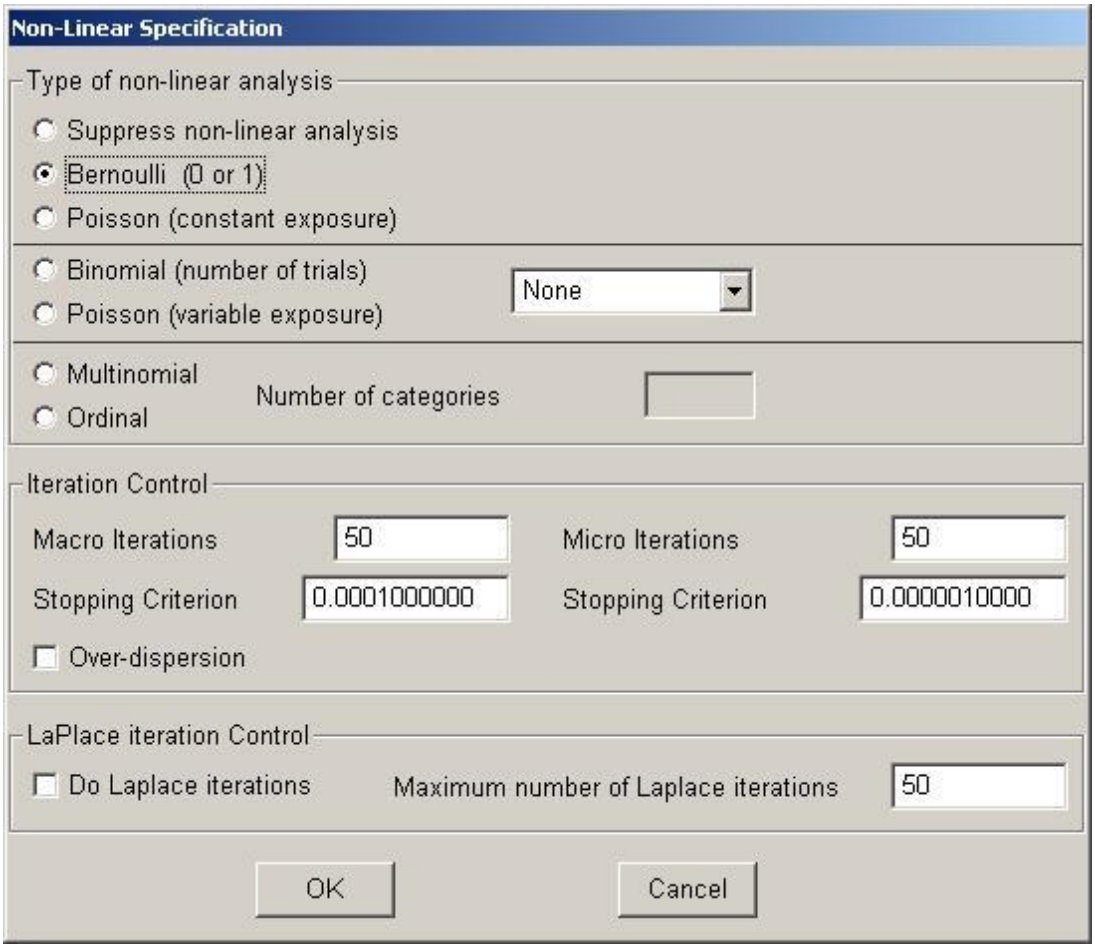
In the above model, the level-1 model has one predictor variable, *Male*, that is used to predict the dependent variable *Rep1*. The dependent variable, *Rep1*, is a Bernoulli distributed variable that is coded as 0 for students who did not repeat a year during their primary school years and 1 otherwise. The variable *Male* is a dummy variable that has a value of 1 if the student is male and 0 if the student is female. The level-2 model has one independent variable for estimating the level-1 intercept, *MSESC*, which is the mean socioeconomic status of the schools included in the analyses and has no independent variables for the level-1 slope, B1. *MSESC* is grand-mean centered in the second level of the model. Notice that the error terms have been removed from the slope coefficient making that term non-random, or constant across schools, whereas the error term remains for the intercept, indicating that the intercept is randomly varying across the level-2 unit, school.

After you have set up the model that you desire to test, the next step is to specify the type of dependent variable in your model. To do this, select the *Setup Non-linear Model* from the *Optional Specifications* menu:

Optional Specifications

Setup Non-linear Model

This will produce the following dialog box:



The critical specification for this dialog box is to select the appropriate type of dependent variable in the *Type of non-linear analysis* section of the dialog box. In the present example, the dependent variable is Bernoulli distributed, so the *Bernoulli (0 or 1)* option has been selected by clicking on the button to its left. If you have a multinomial or ordinal dependent variable, you will also need to specify the number of possible outcomes for that variable. For example, if you have a multinomial dependent variable with the possible outcomes, *yes*, *no*, and *maybe*, you would type 3 in the *Number of categories* box to indicate that there were three possible outcomes for this variable. After specifying the type of variables and optionally modifying the number of iterations, click OK to exit this dialog box.

At this point, you are ready to execute the analysis. Do so by selecting the *Run Analysis* menu item. After the analysis is completed, view the output by selecting the *View Output* option from the *Edit* menu:

Edit

View Output

There are three sets of parameter estimates that will be obtained in the final output: the linear model with the identity link function, the unit-specific model with a link function, and the population average model with a link function. The *linear model with the identity link function* is

the model without a link function; it is only used to obtain starting values for the estimation of the models containing link functions, and is therefore typically not interpreted. The *unit-specific* model contains the random effect from the level-2 model and thus is a prediction of a school typical of the independent variables in the model. In this case, the coefficients would represent a model of a school typical of its SES. In contrast, the *population-average* model does not contain a random effect, thus producing output that is typical of the population average. Notice that if you do not have any error terms in your level-2 model, these two models reduce to the same model, as the distinction between the two is that the unit-specific model contains random effects where the population-average model does not.

One way to conceptualize the difference is to consider the predicted value for two male students from two different schools with the same SES. In the population-average model, the effects of individual schools are not considered and therefore, the predicted values would be the same for both of these students. However, in the unit-specific model, the unique effect of individual schools are incorporated, which would thus result in different predicted values for two students that were identical on the above variable values.

Running the analysis for the model described above will produce the following unit-specific output for the model shown above:

```
Final estimation of fixed effects: (Unit-specific model)   ------------------
----------------------------------------------------------

                                        Standard             Approx.
Fixed Effect         Coefficient   Error      T-ratio   d.f.     P-value    --
-------------------------------------------------------------------------------
   For       INTRCPT1, B0       INTRCPT2, G00          -2.317829   0.085380
-27.147       354    0.000          MSESC, G01          -0.403177
0.193792    -2.080       354    0.037    For     MALE slope, B1
INTRCPT2, G10           0.512402   0.073732     6.950      7513    0.000    --
-------------------------------------------------------------------------------
```

As this is the unit-specific output, it contains the random effect for level-2 units and should thus be used in situations where you are interested in the unique effects of individual level-2 units, such as schools. Examining this output, there are two coefficients that are of greatest interest: the effect of *MSESC* on the slope of the intercept, G01, and the intercept of the slope, G10. While there is a third coefficient, G00, which is the intercept of the level-1 intercept, it is typically not of interest as it only tells us that the grand intercept of *Rep1* is not zero. The effect of the level-2 variable, *MSESC*, on the slope of the intercept, G01, examines the relationship between the level-2 variable and the differences in the averages of *Rep1* that can be attributed to *MSESC*. In the present example this would be the relationship between the mean SES of the schools in the sample and the likelihood of repeating a grade. The output above indicates that as *MSESC* increases in value, there is a decrease in the value of the intercept. Translated into the research question, this indicates that higher levels of SES are associated with fewer instances of repeating a grade.

The second coefficient of interest is the intercept of the slope, B1. As can be seen by examining the model above, the intercept of the slope is the only variable that is used to estimate the slope

and therefore it represents the average change in likelihood of repeating a grade per unit of the level-1 independent variable. In this specific example, the intercept of the level-1 slope is estimated at .51 which becomes the slope of the level-1 equation, B1. The fact that the slope is positive indicates that as level-1 units increase in value, so does the independent variable; that is, when the value of the indicator variable, *Male*, is 1, it increases the likelihood that the dependent variable is 1, meaning that males are more likely to repeat a grade.

The second set of output that is likely of interest is the coefficients for the population-average model. These coefficients are shown below:

```
Final estimation of fixed effects: (Population-average model)    -------------
----------------------------------------------------------------

                                        Standard             Approx.
Fixed Effect         Coefficient   Error      T-ratio   d.f.      P-value    --
-------------------------------------------------------------------------------
   For       INTRCPT1, B0       INTRCPT2, G00          -1.980885   0.079551
-24.901       354    0.000          MSESC, G01          -0.381542
0.185779    -2.054       354    0.040    For     MALE slope, B1
INTRCPT2, G10           0.449826   0.066657     6.748      7513    0.000    --
-------------------------------------------------------------------------------
```

This output contains the same coefficients for the same parameters as the prior output and the same coefficients are of interest. As the interpretation is the same as above, these coefficients will not be discussed in detail. However, the general conclusion is different as this output is appropriate for population-average conclusions. The critical difference between the unit-specific and population-average models is that this output does not include the random error associated with individual schools. Thus, the level-1 coefficients are not random because they do not include the error term that is associated with individual schools.

References

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage.

Snijders, T. A. B., Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Newbury Park, CA: Sage.