Psych 610: MLM, Fall 2014

**Pre-cursors to Hierarchical Linear Modeling (HLM) in SPSS**

FOR THE HOMEWORK: Be sure to interpret your results in meaningful terms! For example, don’t just say “the intercept,” write about everything in terms of what the variables actually refer to. You’re working with a real dataset. Use the analyses to try to understand the relationships between the variables, and then explain those relationships to your reader.

# **Disaggregation vs. Aggregation**

|  |  |
| --- | --- |
| **Disaggregation** | **Aggregation** |
| Ignores group-level data  (discards between-group variability) | Ignores individual-level data  (discards within-group variability) |

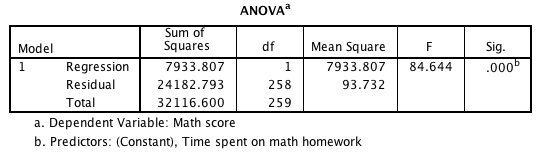
Open *NELS88.sav*

Variables of interest: (1) schoolid, (2) mathscore, and (3) timeonmath

***Disaggregation****:* Estimate a disaggregated model predicting math achievement from time spent on math homework each week. Since disaggregation ignores group-level data, we will not use the *schoolid* variable in our regression model.

REGRESSION

/MISSING LISTWISE

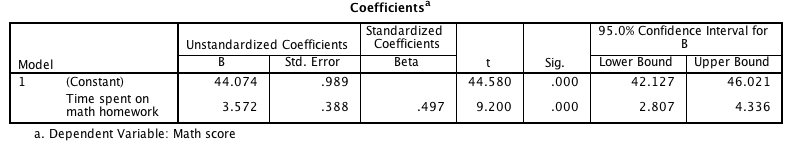
/STATISTICS COEFF OUTS R ANOVA CI(95)

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT mathscore

/METHOD=ENTER timeonmath.

**

*Question: What do the results suggest?*

*Question: What is the problem with conducting the analysis in this manner?*

***Aggregation****:* Estimate an aggregated model predicting mean math achievement from mean hours spent on homework each week. Since aggregation ignores individual-level data, we need to compute the mean math score and the mean time spent on math homework for each of the 10 schools:

DATASET DECLARE aggregated.

*Point-and-click:*

Data 🡪 Aggregate

“Schoolid” 🡪 “Break Variable(s)”

“timeonmath” & “mathsocre” 🡪 “Aggregated Variables”

Select “Create a new dataset…”

AGGREGATE

/OUTFILE='aggregated'

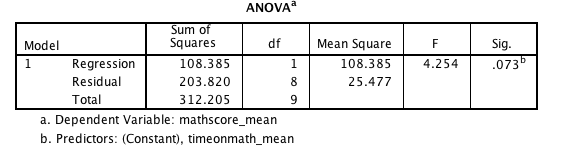
/BREAK=Schoolid

/timeonmath\_mean=MEAN(timeonmath)

/mathscore\_mean=MEAN(mathscore).

This opens a new dataset with mean time spent on math and mean math score, with one row for each school. In this new dataset, run a regression predicting mean math score from mean time spent on math.

DATASET ACTIVATE aggregated.

REGRESSION

/MISSING LISTWISE

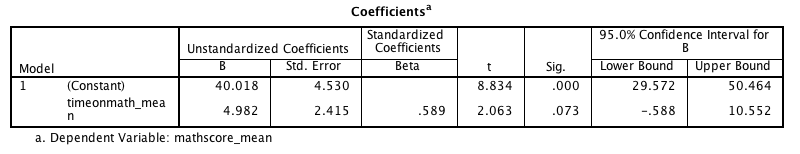
/STATISTICS COEFF OUTS R ANOVA CI(95)

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT mathscore\_mean

/METHOD=ENTER timeonmath\_mean.

**

*Question: What do the results suggest?*

*Question: What is the problem with conducting the analysis in this manner?*

***Within and Between Groups Analysis (WABA)***

Unlike running a disaggregated model (which ignores grouping) or an aggregated model (which ignores within-group variation), the WABA model lets you include both within- and between-group effects in the same model.

The *within-group* effect is captured by a variable that represents the difference between an individual’s score and their group mean:

**Add the group means to the disaggregated dataset:**

AGGREGATE

/OUTFILE=\* MODE=ADDVARIABLES OVERWRITEVARS=YES

/BREAK=schoolid

/time\_mean=MEAN(timeonmath).

**Compute the within-group effect:**

COMPUTE time\_wi = timeonmath - time\_mean.

EXECUTE.

The *between-group* effect is captured by a variable that represents the difference between a group’s mean and the grand mean:

**Find the grand mean:**

MEANS timeonmath.

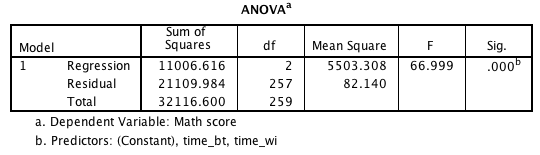
**Compute the between-group effect:**

COMPUTE time\_bt = time\_mean - 2.023077.

EXECUTE.

**Run the WABA model:**

REGRESSION

 /DESCRIPTIVES MEAN STDDEV CORR SIG N

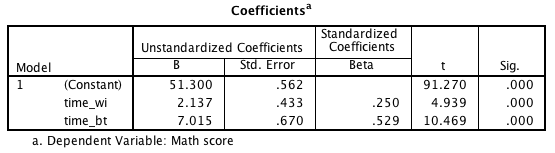
/MISSING LISTWISE

/STATISTICS COEFF OUTS R ANOVA

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT mathscore

/METHOD=ENTER time\_wi time\_bt.

*Within:* average relationship between time on math and math score when the between-group effect is zero

*Between*: average relationship between time on math and math score when the within-group effect is zero***Calculating the ICC for Math Achievement:***

*Question: How can we get the between-schools variability, within-schools variability, and the total variability in math achievement?*

Note that both the GLM and One-way ANOVA functions will give you the same result, though the GLM results are somewhat less straight-forward (the SS for the grouping variable is the between-groups SS; “error” = within groups SS; “corrected total” = total SS)

GLM mathscore BY schoolid

/PRINT = DESCRIPTIVE PARAMETER HOMOGENEITY

/DESIGN = schoolid .

ONEWAY mathscore BY Schoolid

/MISSING ANALYSIS.

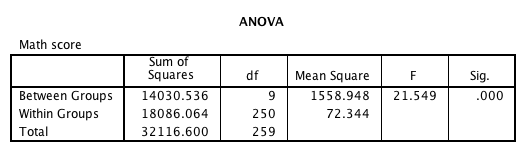
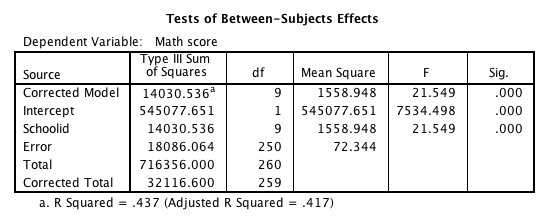
*What is the between-schools variability (SSb)? \_\_\_*14030.54*\_\_\_\_\_\_\_\_\_*

*What is the within-schools variability (SSw)?* \_\_\_\_18086.06\_\_\_\_\_\_\_\_

*What is the total variability? \_* 32116.6*\_*

*What is the ICC? \_\_*SSb / SSt = 14030.54 / 32116.6 = **.44**\_\_\_\_\_\_\_\_\_\_

*What does this mean?*

***Examining Variability in Slopes***

Because we have substantial between-group variability, it’s useful to visualize our data to see if they vary in slopes, intercepts, or both. You can do this by generating a separate scatterplot for each group or by plotting all of the groups together.

**To generate separate plots:**

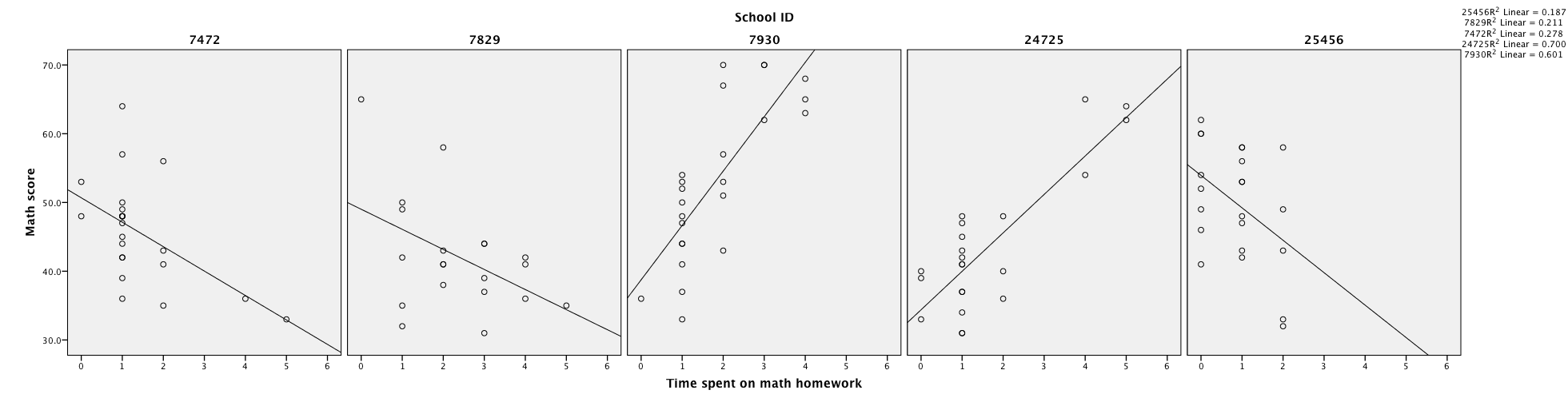
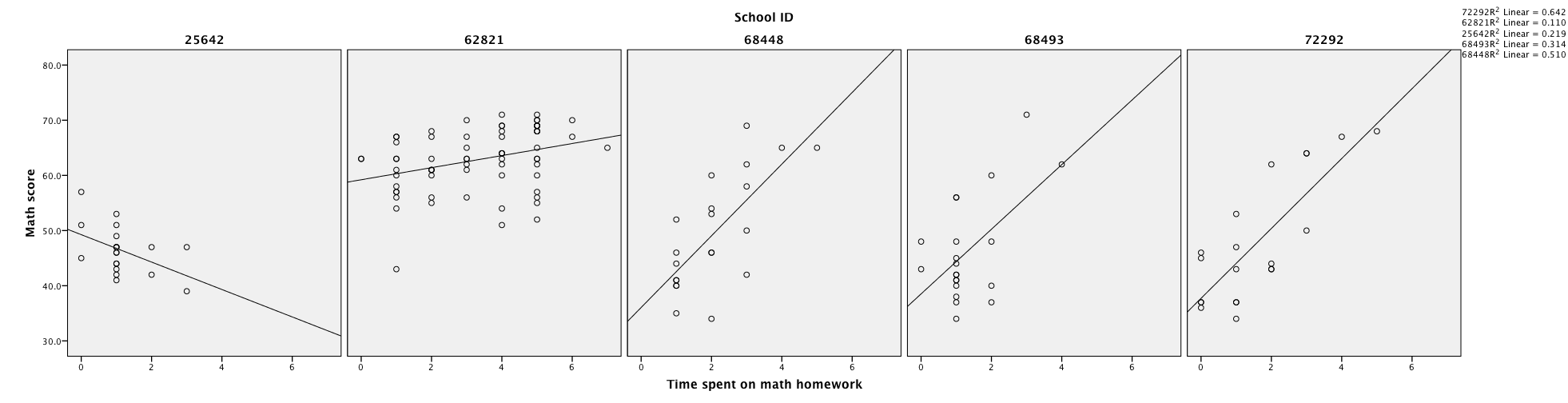
SORT CASES BY schoolid.

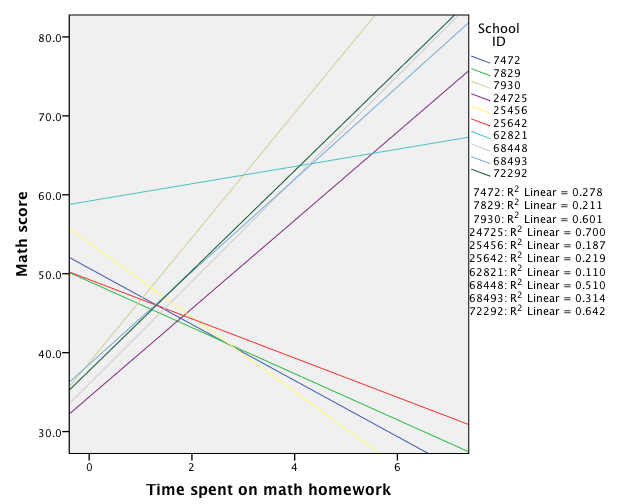
SPLIT FILE LAYERED BY schoolid.

GRAPH

/SCATTERPLOT(BIVAR)=timeonmath WITH mathscore

/MISSING=LISTWISE.





**To generate a single plot (be sure that split file is off!):**

GRAPH

/SCATTERPLOT(BIVAR)=timeonmath WITH mathscore BY Schoolid

/MISSING=LISTWISE.

\*then, in the chart editor, select "add fit line at subgroups"

\*then, double click on the lines to get the line properties, and un-select "attach label to line"

\*then, in properties window, delete the "element type: marker" element.

The plots suggest that the regression lines predicting math achievement from hours spent on homework are pretty different from school to school. If that variability is real, we usually want to try to figure out what characteristics of a school predict its regression line (i.e. its slope and intercept) so we can say something more substantial than just “the way time spent on homework predicts math scores is different school to school.”

***Slopes and Intercepts as Outcomes, Part 1: Run regressions at L1, within each school.***

***Step 1a:***Since we will now include a categorical variable (to mark the groups) in our regression model, we need to create dummy variables. With 10 schools, we will have 9 dummy variables.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **D1** | **D2** | **D3** | **D4** | **D5** | **D6** | **D7** | **D8** | **D9** |
| **7472** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **7829** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **7930** | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **24725** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **25456** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **25642** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **62821** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **68448** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **68493** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **72292** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

COMPUTE D1=0.

IF (Schoolid=7472) D1=1.

exe.

*Question: Which school is the comparison school? \_\_\_\_\_\_\_\_\_\_*

***Step 1b:*** While we’re at it, we should make the interaction terms that will represent the difference in the slope (i.e., the effect of time spent on math homework) between schools.

COMPUTE D1HW = D1\*timeonmath.

exe.

***Step 2:*** Check to see if the grouping variable (i.e., now represented as D1 to D9) significantly contributes to our model.

Model 1: 

Model 2: 

REGRESSION

/MISSING LISTWISE

/STATISTICS COEFF OUTS R ANOVA CHANGE

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT mathscore

/METHOD=ENTER timeonmath

/METHOD=ENTER d1 d2 d3 d4 d5 d6 d7 d8 d9.

*Question:* Looking at the table below, do we have a main effect of our grouping variable (i.e., school)? In other words, is the change in R squared significant?

***Step 3:*** Check to see if the interaction terms significantly contribute to our model.

REGRESSION

/MISSING LISTWISE

*Question:* If the interaction terms did NOT significantly contribute to our model, what type of analysis would we have?

/STATISTICS COEFF OUTS R ANOVA CHANGE

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT mathscore

/METHOD=ENTER timeonmath

/METHOD=ENTER d1 d2 d3 d4 d5 d6 d7 d8 d9

/METHOD=ENTER d1hw d2hw d3hw d4hw d5hw d6hw d7hw d8hw d9hw.

Model 1:



Model 2:

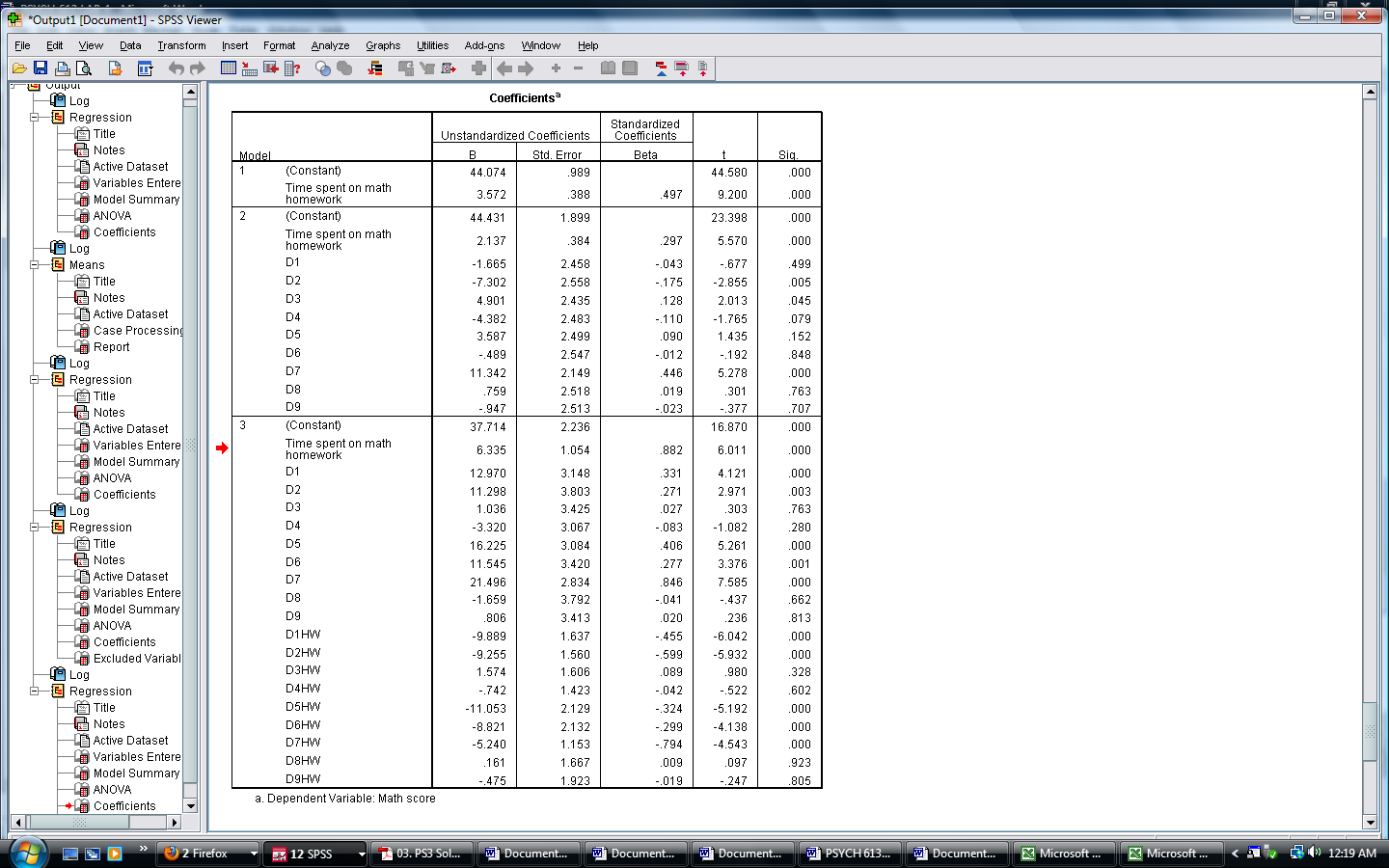


Model 3:



Now we know that there is significant between-school variance in both the intercepts and the slopes. Now we need to find the **unique intercept and slope** for each of the 10 schools.

***Option 1:***



We can calculate the slope and intercept for each school by hand from this output. The equation for each school contains the (1) common intercept, (2) group intercept, (3) common slope, and the (4) group slope.

*Question:* What is the unique intercept for school 7930 (group 3)?

*Answer:*

*Question:*  What is the unique slope for school 7930 (group 3)?

*Answer:*

***Option 2:***

*Point-and-Click*:

Data 🡪 Split file 🡪 Compare groups

Put *Schoolid* into the “groups based on” box

SORT CASES BY schoolid.

SPLIT FILE LAYERED BY schoolid.

REGRESSION

/MISSING LISTWISE

/STATISTICS COEFF OUTS R ANOVA CHANGE

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT mathscore

/METHOD=ENTER timeonmath.

After you run the syntax, you’ll get output for ten regressions, one for each school. The values you get for each school’s slope and intercept here are exactly the same as what you would get calculating them by hand from the one multiple regression model.

***Slopes and Intercepts as Outcomes, Part 2: Run two regressions at L2, one predicting intercept and one predicting slope.***

Let’s use a group-level variable (e.g., type of school – private or public) to predict the variability in the (1) intercepts and (2) slopes.

***Step 1:*** Create a new dataset that includes the slope and intercept of each school (open *NELS88\_L2*), as well as whether the school is public or private. *Pro Tip: This is 100% easier in R; see additional handout.*

***Step 2:*** Run a regression where you predict the *variability in the intercepts* from *type of school*



Remember: Type of school (Private = 0, Public = 1)

*Question:* What is the regression equation predicting intercept from school type?

*Question:* When time spent on math homework is 0, what is the expected math achievement for a private school? How about for a public school?

*Question:* The slope is marginally significant. What does this mean?

***Step 3:*** Run a regression where you predict the *variability in the slopes* from *type of school*



*Question:* What is the regression equation predicting slope from school type?

*Question:* What is the relationship (slope) between time spent on homework and math achievement for a private school? How about for a public school? Is the relationship (slope) significantly different for the two types of schools?

**Hierarchical Linear Modeling (HLM) in SPSS**

***Random Coefficient Regression: The “full” MLM***

In RCR, every L1 coefficient (the betas) is allowed to vary randomly at the group level (L2). What would the RCR model be to predict math scores from hours spent on math homework, school type (public vs. private), and the interaction between time on math and school type?

L1:

What if you found that the slopes did NOT vary randomly across groups? How would that model be different?

L2:

Single Equation:

What is the meaning of each parameter?

g00= u00 =

g01=

g10= u11 =

g11= eij =

**MLM in SPSS**

Categorical predictor

Continuous predictor

DV

Print z-tests for the variances components.

MIXED mathscore WITH timeonmath BY schooltype

Print the parameter estimates.

/PRINT TESTCOV SOLUTION

Get fixed effects (*g*s) for these variables.

/FIXED INTERCEPT timeonmath schooltype timeonmath\*schooltype

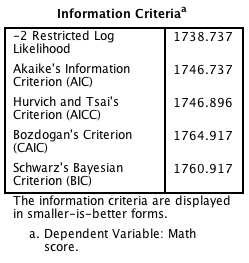
/RANDOM INTERCEPT timeonmath | SUBJECT(schoolid) COVTYPE(UN) .

This is the variable that demarcates the units we’re assuming are independent.

Estimate the variances and covariance(s).

Get random effects (*u*s) for these variables.

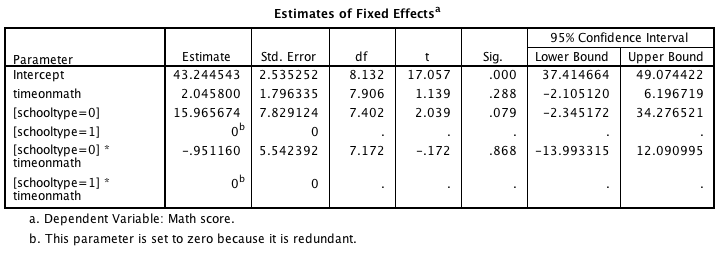
***Output:***



These stats are measures of model fit (or rather, of badness of fit – the smaller the values, the better the model).

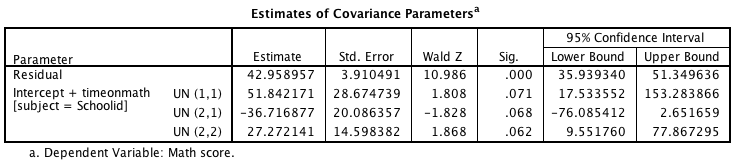
-2 Log Likelihood is also called “deviance”. Alone, these values are meaningless, but they can be compared across nested models to determine which model has better fit (i.e., in a deviance-change test).

SPSS uses some crazy magic called the Satterthwaite approximation to calculate df for multilevel designs, so you’ll see weird degrees of freedom.



σ2

(L1 res. variance)



The L2 variance components.

(1,1) = τ00, (2,1) = τ01, and (2,2) = τ11.



g00= τ00 =

g01= τ01 =

g10= τ11 =

g11= σ2 =