Assignment #3

Data Summary:

This analysis uses data from the Study Instructional Improvement, 1 or SII was a research done by researchers at the University of Michigan. The research was done to study the math achievement scores of first and third grade students in randomly selected classrooms from a sample of U.S. elementary schools. In 107 schools, 312 classrooms, 1190 first grade students were selected. MATHGAIN is the dependent variable, it measures change in student math achievement scores from the spring of kindergarten to spring of first grade. The data is a 3-level data set, students is level 1 and it is nested within classrooms (level 2), and classrooms are nested within level 3 which is schools. The purpose of this analysis is to see the influence that the covariates in 3 levels can have on the students' math achievement gain(MATHGAIN).

Data Dictionary Table:

Variable name	Variable description	Variable type	Variable Categories	Variable Mean	Variable Normality	Variable Level
SCHOOLID	School ID number	numerical	-	-	-	3
HOUSEPOV	Percent of households in the neighborhood of school below the poverty level	numerical	-	0.17824 (17.824%)	-	3
CLASSID	Classroom ID number	numerical	-	-		2
YEARSTEA	First grade teacher's year of teaching experience	numerical	-	12.21216		2
MATHPREP	First grade teacher's mathematics preparation: number of math	numerical	-	2.61249		2

	content and method courses					
MATHKNOW	First grade teacher's mathematics content knowledge; based on a scale based composed of 30 items	numerical		0.31206		2
CHILDID	Student ID number	numerical		-	-	1
MATHGAIN	Student's gain in math achievement score from the spring of kindergarten to the spring of first grade(dependent variable)	numerical		57.56639	Fairly normal with a slight skewness of 0.50256	1
MATHKIND	Student's math score in the spring of their kindergarten year	numerical		466.65882	Fairly normal with a slight skewness of -0.30260	1
SEX	Sex of student	Indication	0=boy, 1=girl	-	-	1
MINORITY	Whether the student is a minority	Indication	0=non-minority, 1=minority	-	-	1
SES	Student socioeconomic status	numerical		-0.01298	Fairly skewed to the right with a skewness of 1.20684	1

NOTE: "-"signs in table means the variable's value for particular column has no meaning, thus are not provided. For instance, there is no meaning in this study to provide the mean or shape of ID numbers.

Dimensions	
Covariance Parameters	3
Columns in X	1
Columns in Z per Subject	10
Subjects	107
Max Obs per Subject	31

Table 1.1

Dimension table above shows that this contains 3 covariance parameters in the matrix and 107 subjects(schools) in this model. SAS provides 10 columns in the Z matrix per school. The maximum number of student per each subject(School) is 31.

mathknow	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	17802	9.50	17802	9.50

Using SAS procedure proc freq, the variable mathknow contains missing values was detected. All other variables are complete with no missing values or entry typos found.

Graphs of interest:

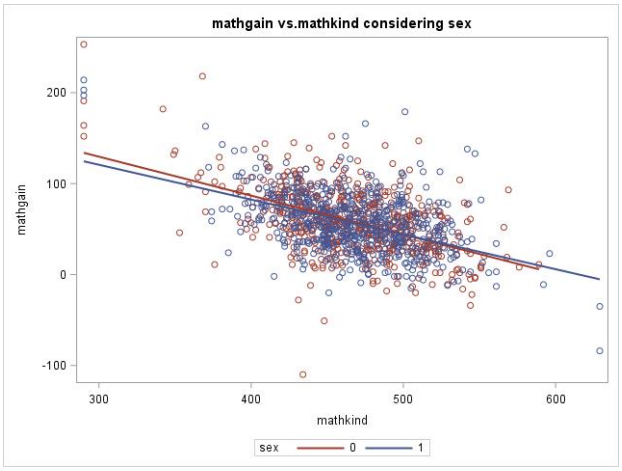

Figure 1.1

Figure 1.1 illustrates the relationship between student's gain in math achievement score from the spring of kindergarten to the spring of first grade and student's math score in the spring of their kindergarten year. Sex was also taken into consideration with the 2 score variables. Figure 1.1 suggests that regardless of sex, most students start with mid-range math scores (mathkind between 400-500) earn fairly similar increases during the1st grade. The trend seems like the students who started with lower mathkind scores have higher mathgain scores. The students who started with higher mathkind scores tend to keep the score they have in kindergarten and did not

improve much during the 1st year. Due to the cluster shape graph in the center of the graph, and the outliers on the ends, it is reasonable to doubt this trend is heavily influenced by the outliers. there is no sufficient evidence to claim such trend is the genuine relationship.

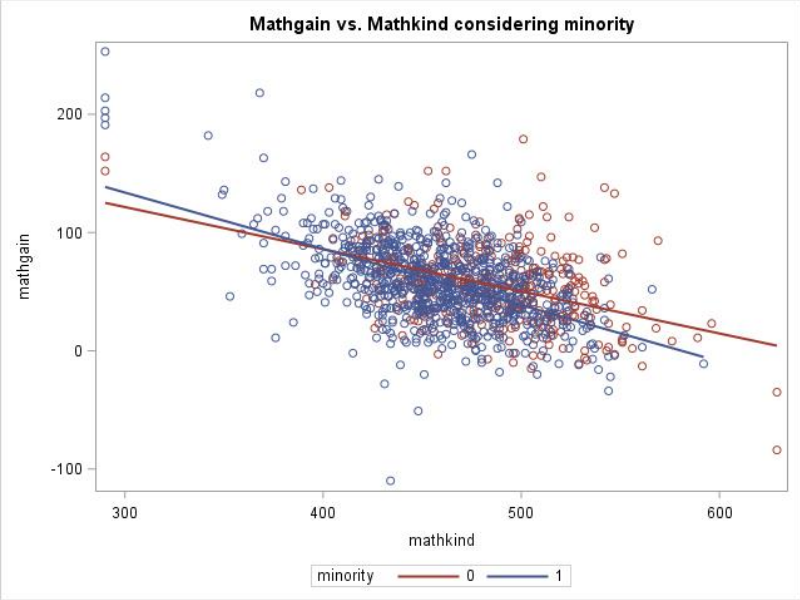

Figure 1.2

Figure 1.2 illustrates the relationship between mathgain and mathkind taking minority into consideration. The relationship between mathkind and mathgain still seems negative, but similar to figure 1.1, the trend is heavily biased by the influential outliers. The interesting part of this graph is in the center, the dots indicating non-minority students has slightly higher mathkind to begin with.

Model building and comparison[1]:

Model 0(with random effects school and class nested in school):

$$Mathgain_{ijk} = \beta_0 + u_k + u_{j|k} + \varepsilon_{ijk}$$

[1] All models formulas are written in latex lyx.

Fit Statistics	
-2 Res Log Likelihood	11768.8
AIC (Smaller is Better)	11774.8

Table 2.1

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	57.4271	1.4431	106	39.80	<.0001

Table 2.2

From table 2.2 obtain that model 0 for overall intercept with random effect classid nested with in schoolid model is $Mathgain_{ijk} = 57.4271 + u_k + u_{j|k} + \varepsilon_{ijk}$

Model 0A was built to see whether there is a difference between including random effect classid nested with in schoolid or only considering the random effect classid.

Model 0A(not nesting class in school):

$Mathgain_{ijk} = \beta_0 + u_k + \varepsilon_{ijk}$

Fit Statistics	
-2 Res Log Likelihood	11776.7
AIC (Smaller is Better)	11780.7

Table 2.3

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	57.5835	1.4413	106	39.95	<.0001

Table 2.4

From table 2.4 obtain that the overall intercept with no classid nested in schoolid random effect $Mathgain_{ijk} = 57.5835 + u_k + \varepsilon_{ijk}$

Table 2.1 and 2.3 suggests that model 0 has a slightly smaller AIC value of 11774.8 verses 11780.7 for model 0A. Therefore, model 0 will be used for further model building.

Model 1:

$Mathgain_{ijk} = \beta_0 + \beta_1 * mathkind_{ijk} + \beta_2 * sex_{ijk} + \beta_3 * minority_{ijk} + \beta_4 ses_{ijk} + u_k + u_{j|k} + \varepsilon_{ijk}$

Fit Statistics	
-2 Res Log Likelihood	11385.8
AIC (Smaller is Better)	11391.8

Solution for Fixed Effects

Fit Statistics

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	282.79	10.8533	106	26.06	<.0001
mathkind	-0.4698	0.02227	874	-21.10	<.0001
sex	-1.2511	1.6577	874	-0.75	0.4506
minority	-8.2620	2.3401	874	-3.53	0.0004
ses	5.3464	1.2411	874	4.31	<.0001

Table 2.5

Model 1 was built based on model 0, and adding level 1 covariate variables, mathkind, sex, minority and ses. The purpose of model 1 was to test whether the covariates in level 1 have significant influence on mathgain. Model 1 is in default procedure REML, and table 2.5 suggests that all variables aside from sex are significant by t-test using alpha=0.05. Test using ML procedure for model 0 and model 1 was also done to see whether the variables should be kept in the model. The ML method shows that model 1 is the better model (see details in appendix). Therefore, the covariates at level 1[2] do have significant influence over mathgain scores.

Model 2:

$$Mathgain_{ijk} = \beta_0 + \beta_1 * mathkind_{ijk} + \beta_2 * sex_{ijk} + \beta_3 * minority_{ijk} + \beta_4 ses_{ijk} + \beta_5 * yearstea_{jk} + \beta_6 * mathprep_{jk} + \beta_7 * mathknow_{jk} + u_k + u_{j|k} + \varepsilon_{ijk}$$

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	282.02	11.7016	103	24.10	<.0001
mathkind	-0.4750	0.02275	792	-20.88	<.0001
sex	-1.3395	1.7186	792	-0.78	0.4360
minority	-7.8680	2.4181	792	-3.25	0.0012
ses	5.4194	1.2760	792	4.25	<.0001
yearstea	0.03975	0.1170	792	0.34	0.7343
mathprep	1.0948	1.1482	792	0.95	0.3406
mathknow	1.9149	1.1468	792	1.67	0.0954

Table 2.6

[2] Sex will be kept at this point for further model building.

Model 2 was built by adding level 2 covariate variables, yearstea, mathprep, and mathknow to model 1. The purpose of model 2 is to test whether level 2 covariates have significant influence over mathgain. Table 2.6 suggests that all three of these variables do not have statistical significance at alpha=0.05. Therefore, there is not sufficient evidence to conclude that covariates in level 2 have significant influence over mathgain.

Model 3:

$$Mathgain_{ijk} = \beta_0 + \beta_1 * mathkind_{ijk} + \beta_2 * sex_{ijk} + \beta_3 * minority_{ijk} + \beta_4 ses_{ijk} + \beta_5 * housepov + u_k + u_{j|k} + \varepsilon_{ijk}$$

Fit Statistics	
-2 Res Log Likelihood	11378.1
AIC (Smaller is Better)	11384.1

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	285.06	11.0208	106	25.87	<.0001
mathkind	-0.4709	0.02228	873	-21.13	<.0001
sex	-1.2345	1.6574	873	-0.74	0.4566
minority	-7.7557	2.3850	873	-3.25	0.0012
ses	5.2398	1.2450	873	4.21	<.0001
housepov	-11.4398	9.9373	873	-1.15	0.2500

Table 2.7

Model 3 was built by adding level 3 covariate variables(housepov) to model 1, the purpose of model 3 was to test whether housepov have significant influence on mathgain. Table 2.7 suggests that housepov's p-value from t-test is 0.25, which is larger than significance level alpha=0.05. Therefore, there is no evidence to conclude that the covariate housepov in level 3 have significant influence over mathgain.

Model 3.1:

$$Mathgain_{ijk} = \beta_0 + \beta_1 * mathkind_{ijk} + \beta_2 * sex_{ijk} + \beta_3 * minority_{ijk} + \beta_4 * ses_{ijk} + \beta_5 * housepov + \beta_6 * (housepov * mathkind) + \beta_7 * (mathkind * ses) + \beta_8 * (mathkind * minority) + \beta_9 * (mathkind * sex) + \beta_9 * (housepov * minority) + u_k + u_{j|k} + \varepsilon_{ijk}$$

Fit Statistics

-2 Res Log Likelihood	11379.6
AIC (Smaller is Better)	11385.6

Table 2.8

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	278.99	25.1273	106	11.10	<.0001
mathkind	-0.4652	0.05156	868	-9.02	<.0001
sex	-26.3619	18.7476	868	-1.41	0.1600
minority	19.5757	23.9256	868	0.82	0.4135
ses	-10.0176	13.7822	868	-0.73	0.4675
housepov	-2.2461	90.4697	868	-0.02	0.9802
mathkind*housepov	0.02430	0.1843	868	0.13	0.8951
mathkind*ses	0.03222	0.02912	868	1.11	0.2688
mathkind*minority	-0.05046	0.04954	868	-1.02	0.3086
mathkind*sex	0.05402	0.03999	868	1.35	0.1771
minority*housepov	-24.1716	22.7314	868	-1.06	0.2879

Table 2.9

Since model 3 have not taken interaction into consideration, model 3.1 was built to see whether any of the interactions between level 3 covariates and level 1 covariates are significant. Table 2.9 indicates that none of the interactions between housepov and level 1 covariates are statistically significant at alpha=0.05. Table 2.8 suggests that the AIC of model 3.1 is 11384.1 and the AIC for model 3 is 11378.1. Which indicates that model 3.1 is not a better model than model 3. Therefore, interactions will not be considered in futher models.

At this point, model 1 is the best model for this dataset, which have only level 1 covariate variables and taken random effects of classid and nested classid within schoolid into consideration. But table 2.5 from model 1 suggests that the variable sex is not statistically significant. To make the model as efficient as possible, model EXTRA was developed to see whether the covariate sex is needed and which of the models have a lower AIC and -2REML value.

Model EXTRA:

$$Mathgain_{ijk} = \beta_0 + \beta_1 * mathkind_{ijk} + \beta_3 * minority_{ijk} + \beta_4 ses_{ijk} + u_k + u_{j|k} + \varepsilon_{ijk}$$

Fit Statistics	
-2 Res Log Likelihood	11389.2
AIC (Smaller is Better)	11395.2

Table 2.10

After the variable sex is taken out of the model, all variables are significant. Table 2.10 shows that the -2REML for model extra is 11389.2, which is larger than model 1's -2REML of 11385.3. And the AIC for EXTRA is 11395.2, while the AIC for model 1 in table 2.5 is 11391.8. Although the difference between the AIC's is not very large, the model with lower AIC(model 1) will be chosen as the best model.

Summary of Results:

The final model is model 1 :

$$Mathgain_{ijk} = \beta_0 + \beta_1 * mathkind_{ijk} + \beta_2 * sex_{ijk} + \beta_3 * minority_{ijk} + \beta_4 ses_{ijk} + u_k + u_{j|k} + \varepsilon_{ijk}$$

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	282.79	10.8533	106	26.06	<.0001
mathkind	-0.4698	0.02227	874	-21.10	<.0001
sex	-1.2511	1.6577	874	-0.75	0.4506
minority	-8.2620	2.3401	874	-3.53	0.0004
ses	5.3464	1.2411	874	4.31	<.0001

Table 3.1

The student's gain in math achievement score form the spring of kindergarten to the spring of first grade is explained by random effect class ID number, random effect class ID number nested within school ID number, fixed effects student's math score in the spring of their kindergarten year, the sex of student, whether the student is a minority, and the socioeconomic status of the student(specific numerical influences indicated in table 3.1).

This model has an AIC of 11391.8, which is one of the lowest AIC among all models compared. The statistically significant covariate variables in the final model are logically reasonable as well. Sex

being a variable that is left in the model, not statistically significant based on table 3.1 and does not suggest what kind of influence it has on mathgain. Does the gender of the child really effect the math achievement score of the child during early stages of systematic learning? If further analysis will be made on this dataset, sex will certainly be of the variables of interest.

Appendix

```
/*STAT*/
proc import out = work.class
datafile = "C:\Users\Jun2017\Desktop\6640\classroom.csv" dbms = csv replace;
getnames = YES;
datarow = 2;
guessingrows = 20;
run;

proc print data=work.class;
run;
/*math know has missing values*/
proc contents data=class;
RUN;
proc freq data=class;
      tables mathknow mathkind mathgain mathprep yearstea housepov sex
minority ses/missing;
      weight classid;
      title 'checking for missing values';
   run;
/*data preparation EDA/means/graphs*/
proc univariate data=work.class plot;
title"mathgain, dependent";
var mathgain;
run;
proc univariate data=work.class plot;
title"other variables";
var mathkind housepov yearstea mathknow mathprep ses;
run;
proc sgplot data=work.class;
title"mathgain vs.mathkind considering sex";
reg y=mathgain x=mathkind/group=sex;
run;
proc sgplot data=work.class;
title"Mathgain vs. Mathkind considering minority";
reg y=mathgain x=mathkind/group=minority;
run;
/*first model*/
proc sort data=work.class;
by schoolid classid;
run;
/*initial tests of classroom random effect*/
proc mixed data = work.class noclprint covtest;
title "Model 0";
class classid schoolid;
model mathgain =/ solution;
random intercept/ subject = schoolid v vcorr;
random intercept/ subject = classid(schoolid);
run;

proc mixed data = work.class noclprint covtest;
title "Model 0A";
class classid schoolid;
model mathgain =/ solution;
random intercept/ subject = schoolid;
run;
```

```
/* I would go with model 0, the model with random factor classroom nested in
the model,
this is determined using p-value*/


/*building model 1*/
proc mixed data = work.class noclprint covtest;
title "Model 1";
class classid schoolid;
model mathgain = mathkind sex minority ses/ solution;
random intercept/ subject = schoolid;
random intercept/ subject = classid(schoolid);
run;
proc mixed data = work.class noclprint covtest method=ML;
title "Model 0ML";
class classid schoolid;
model mathgain =/ solution;
random intercept/ subject = schoolid;
random intercept/ subject = classid(schoolid);
run;
proc mixed data = work.class noclprint covtest method=ML;
title "Model 1ML";
class classid schoolid;
model mathgain = mathkind sex minority ses/ solution;
random intercept/ subject = schoolid;
random intercept/ subject = classid(schoolid);
run;
/*the covariates does have meaning, and we will go with model 1, because it
is more significant and lower AIC*/


/*building model 2 adding level 2 covariates*/
proc mixed data = work.class noclprint covtest;
title "Model 2";
class classid schoolid;
model mathgain = mathkind sex minority ses yearstea mathprep mathknow/
solution;
random intercept/ subject = schoolid;
random intercept/ subject = classid(schoolid);
run;
/*none of the newly added covariate is significant, therefore, we still keep
model 1 as the current stage reference model*/
/*building model 3 adding level 3 covariates*/
title "Model 3";
proc mixed data = work.class noclprint covtest;
class classid schoolid;
model mathgain = mathkind sex minority ses housepov/ solution;
random intercept/ subject = schoolid;
random intercept/ subject = classid(schoolid);
run;

/*model 3.1 based on model 3 with possible/logically reasonable
interactions*/
proc mixed data = work.class noclprint covtest;
title "Model 3.1";
class classid schoolid;
model mathgain = mathkind sex minority ses
```

```
housepov housepov*mathkind mathkind*ses mathkind*minority mathkind*sex
housepov*minority/ solution;
random intercept/ subject = schoolid;
random intercept/ subject = classid(schoolid);
run;
/*the final model*/
proc mixed data = work.class noclprint covtest;
title "Model 1(final model)";
class classid schoolid;
model mathgain = mathkind sex minority ses/ solution outpred = pdatl;
random intercept/ subject = schoolid solution;
random intercept/ subject = classid(schoolid);
run;
 /*trying extra models*/
proc mixed data = work.class noclprint covtest;
title"Model EXTRA";
class classid schoolid;
model mathgain = mathkind minority ses/ solution outpred = pdatl;
random intercept/ subject = schoolid solution;
random intercept/ subject = classid(schoolid);
run;
```