

EDUC 231D: Homework 4

Shuhan (Alice) Ai

Scenario

A national foundation focused on improving the diversity of the Science, Technology, Engineering, and Math (STEM) workforce wants to learn more about differences in math learning over the course of a student's secondary education (from grade 7 through 12). The foundation's director would like you to use the Longitudinal Survey of American Youth (LSAY) to answer some questions they have. They are particularly interested in disparities in math learning between students from ethnic/racial groups historically under-represented in STEM (defined for this task as Black, Indigenous, and Hispanic/Latinx students) and students from over-represented groups. *As a short-hand, let's refer to these students as UR students and non-UR students.*

To conduct this investigation, the director provided you with a sample of the LSAY data and a description of the file. The LSAY followed a representative sample of students in the U.S. from 7th grade through 12th grade. A math test was administered to students each year. The LSAY file was uploaded to our BruinLearn site and is named `hw4_lsay_data.RDS`.

Use this data file to answer the following questions for the director. Submit your responses as a PDF file by **12PM on February 25**. The file name for the PDF you submit should use the following naming convention:

HW4__[LastName]__[FirstName].pdf

Set Up

To get started, you need to load some R packages and the data file.

Note

If you have not already installed these packages, you will first have to install them before loading the libraries.

```
# clear the R environment just in case there are things loaded that we don't want
# (start with a clean slate)
rm(list=ls())

# load packages
library("tidyverse") # optional package useful for data processing
library("skimr") # optional package useful for summarizing data file contents
library("flextable") # optional package useful for creating tables
library("table1") # another optional package for creating tables
```

```
library("lme4") # the primary package we'll use for estimating multilevel models
library("lmerTest") # package to view p-values for estimates

# load data file: make sure the file path matches where you have the file saved
hw <- readRDS("/Users/aishuhan/Desktop/EDUC 231D Multilevel Analysis/Assignments/HW4/hw4_lsay_data")

# set a working directory for where you can save files
setwd("/Users/aishuhan/Desktop/EDUC 231D Multilevel Analysis/Assignments/HW4")
```

Description of the Data File

The data file includes data on 1,000 students, each with math test scores covering 6 grades (7 - 12). The following variables are included in the file:

- *CASENUM* = unique ID for each student
- *URSTD* = indicator for whether the student is from an under-represented group (1) or not (0)
- *ALGIN8* = indicator for whether the student took an Algebra course (or higher) in 8th grade (1) or not (0)
- *YEAR* = year of the math test: 0 = 7th grade, 1 = 8th grade, 2 = 9th grade, 3 = 10th grade, 4 = 11th grade, and 5 = 12th grade
- *MTHSCORE* = student's score on the math test

```
skim(hw) # get descriptive statistics
```

Table 1: Data summary

Name	hw
Number of rows	6000
Number of columns	5
Column type frequency:	
numeric	5
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
CASENUM	0	1	4024.43	1688.80	1005.00	2621.50	3997.50	5453.50	6943.00	
URSTD	0	1	0.19	0.39	0.00	0.00	0.00	0.00	1.00	
ALGIN8	0	1	0.20	0.40	0.00	0.00	0.00	0.00	1.00	
YEAR	0	1	2.50	1.71	0.00	1.00	2.50	4.00	5.00	
MTHSCORE	0	1	61.51	13.76	29.14	51.84	61.86	71.14	100.16	

```
# A look at the mean math scores by year and URSTD
tmp <- hw %>% filter(URSTD == 1)
table1(~ MTHSCORE | YEAR, data = tmp)
```

	0	1	2	3	4	5	Overall
	(N=190)	(N=190)	(N=190)	(N=190)	(N=190)	(N=190)	(N=1140)
MTHSCORE							
Mean (SD)	46.7 (8.44)	50.0 (9.73)	54.1 (11.3)	58.5 (12.4)	61.2 (12.8)	61.6 (13.5)	55.3 (12.8)
Median [Min, Max]	46.2 [29.1, 72.4]	49.4 [30.3, 80.3]	52.7 [30.3, 82.6]	58.7 [29.6, 90.6]	62.1 [33.4, 89.4]	63.4 [30.4, 94.3]	54.8 [29.1, 94.3]

```
tmp <- hw %>% filter(URSTD == 0)
table1(~ MTHSCORE | YEAR, data = tmp)
```

	0	1	2	3	4	5	Overall
	(N=810)	(N=810)	(N=810)	(N=810)	(N=810)	(N=810)	(N=4860)
MTHSCORE							
Mean (SD)	53.3 (9.81)	57.0 (10.2)	61.7 (11.9)	66.3 (12.7)	69.1 (13.3)	70.2 (14.0)	63.0 (13.6)
Median [Min, Max]	53.9 [29.8, 79.9]	57.5 [29.3, 84.0]	63.0 [30.0, 94.3]	67.8 [30.5, 95.3]	70.1 [32.7, 98.6]	71.4 [29.7, 100]	63.3 [29.3, 100]


Question 1

Start by fitting the following model to describe math performance growth during secondary school (Model 1):

$$Y_{ti} = \pi_{0i} + \pi_{1i}YEAR_{ti} + e_{ti}, e_{ti} \sim N(0, \sigma^2)$$

$$\pi_{0i} = \beta_{00} + r_{0i}, r_{0i} \sim N(0, \tau_{00})$$

$$\pi_{1i} = \beta_{10} + r_{1i}, r_{1i} \sim N(0, \tau_{11})$$

 Show your work

As part of your response to this question, include any R code and/or output you used to help answer the question.

1.A. What does the parameter π_{0i} represent?

Answer: π_{0i} represents the expected initial math score for student i in 7th grade.

1.B. What does the parameter π_{1i} represent?

Answer: π_{1i} represents the expected rate of change in math scores per year for each student.

1.C. Based on the model results, what point estimate do you get for the grand-mean math score in 7th grade? Interpret the meaning of this result for the director.

Answer: The expected grand-mean math score in 7th grade is 52.73. On average, students start with a math score of 52.73 in 7th grade.

```
model1 <- lmer(MTHSCORE ~ 1 + YEAR + (1 + YEAR | CASENUM), data = hw)
summary(model1)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
```

```
Formula: MTHSCORE ~ 1 + YEAR + (1 + YEAR | CASENUM)
Data: hw
```

```
REML criterion at convergence: 38923
```

```
Scaled residuals:
```

```
      Min       1Q   Median       3Q      Max
-4.7392 -0.4973  0.0052  0.5317  3.7009
```

```
Random effects:
```

```
Groups   Name             Variance Std.Dev. Corr
CASENUM  (Intercept)  84.977    9.218
          YEAR         2.457    1.567    0.39
Residual              17.645    4.201
Number of obs: 6000, groups: CASENUM, 1000
```

```
Fixed effects:
```

```
              Estimate Std. Error      df t value Pr(>|t|)
(Intercept)  52.72784    0.30695 999.01590  171.78   <2e-16 ***
```

```
YEAR          3.51202    0.05887 998.98286    59.66    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Correlation of Fixed Effects:
      (Intr)
YEAR 0.174
```

1.D. To what extent does 7th grade math achievement differ across students? What's the expected math score for a 7th grader with a score 1 standard deviation *below* the mean? What's the expected math score for a 7th grader with a score 1 standard deviation *above* the mean?

Answer: The variance of 7th grade math achievement (τ_{00}) is 84.98, the expected math score for 7th graders with 1 sd below is 43.51, the expected math score for 7th grades with 1 sd above is 61.95.

```
beta_00 <- fixef(model1)["(Intercept)"]
beta_10 <- fixef(model1)["YEAR"]
vc <- as.data.frame(VarCorr(model1))
tau_00 <- vc[1, 4]
tau_00_sd <- vc[1, 5]

mathscore_1sdlow <- beta_00 - tau_00_sd
mathscore_1sdabo <- beta_00 + tau_00_sd
mathscore_1sdlow
```

```
(Intercept)
  43.50955
```

```
mathscore_1sdabo
```

```
(Intercept)
  61.94613
```

1.E. Based on the model results, what is the grand-mean rate of change for math achievement during secondary school? Interpret the meaning of this result for the director, including the direction of the average change and whether the estimate is statistically significantly different from zero.

Answer: The expected grand-mean rate of change (β_{10}) is 3.51 with a standard error of 0.06. This means that, on average, students' math scores increase by 3.51 points per year during 7th to 12th grade. This estimate is statistically significant with $p < 0.001$, meaning that the average rate of change is significantly different from zero and unlikely due to random variation.

1.F. To what extent does the rate of change for math achievement differ across students? What's the expected rate of change for a student with a rate 1 standard deviation *below* the mean rate? What's the expected rate of change for a student with a rate 1 standard deviation *above* the mean?

Answers: The variance of between students growth rate of change (τ_{11}) is 2.46. The expected rate of change for student with 1 sd below is 1.94 points per year, while the expected rate of change for students with 1 sd above the mean rate is 5.08 points per year.

```

tau_11 <- vc[2, 4]
tau_11_sd <- vc[2, 5]

growthrate_1sdlow <- beta_10 - tau_11_sd
growthrate_1sdabo <- beta_10 + tau_11_sd

growthrate_1sdlow

```

```

YEAR
1.944556

```

```

growthrate_1sdabo

```

```

YEAR
5.07948

```


1.G. Do students with relatively higher math scores in 7th grade tend to learn more, about the same, or less math over time compared to students with relatively lower math scores in 7th grade? What led you to this conclusion?

Answer: The correlation between random intercepts (7th grade math score) and random slope (math score growth rate) is 0.39, indicating a positive relationship between initial math scores and learning rates, suggesting that students with higher math scores at 7th grade tend to have faster math growth rate compared to students with lower initial scores.

Question 2

Now estimate a model that tests the director's concern about disparities in math learning between UR and non-UR students (Model 2):

$$\begin{aligned}
 Y_{ti} &= \pi_{0i} + \pi_{1i}YEAR_{ti} + e_{ti}, e_{ti} \sim N(0, \sigma^2) \\
 \pi_{0i} &= \beta_{00} + \beta_{01}(URSTD_i) + r_{0i}, r_{0i} \sim N(0, \tau_{00}) \\
 \pi_{1i} &= \beta_{10} + \beta_{11}(URSTD_i) + r_{1i}, r_{1i} \sim N(0, \tau_{11})
 \end{aligned}$$

 Show your work

As part of your response to this question, include any R code and/or output you used to help answer the question.

2.A. Based on the model results, what's the estimated grand-mean 7th grade math score for a UR student? What's the estimated grand-mean 7th grade math score for a non-UR student? Interpret the meaning of the difference in 7th grade math scores between these two student groups for the director. Make sure to indicate whether the group difference is statistically significantly different from zero.

Answer: The estimated grand-mean 7th grade math score for a UR student is 47.28 ($54.00 - 6.72 = 47.28$). The estimated grand-mean 7th grade math score for a non-UR student is 54.00. This means for UR students, on average, begin 7th grade with significantly lower math scores compared to the non-UR students. The gap of 6.72 points is statistically significant ($p < 0.001$, $\alpha = 0.05$).

```
model2 <- lmer(MTHSCORE ~ 1 + YEAR + URSTD + YEAR:URSTD + (1 + YEAR | CASENUM), data = hw)
summary(model2)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: MTHSCORE ~ 1 + YEAR + URSTD + YEAR:URSTD + (1 + YEAR | CASENUM)
Data: hw
```

REML criterion at convergence: 38846.3

Scaled residuals:

Min	1Q	Median	3Q	Max
-4.7344	-0.4966	0.0018	0.5315	3.6784

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
CASENUM	(Intercept)	78.11	8.838	
	YEAR	2.44	1.562	0.38
Residual		17.65	4.201	

Number of obs: 6000, groups: CASENUM, 1000

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	54.00436	0.32840	998.00467	164.448	<2e-16 ***
YEAR	3.58019	0.06525	998.00975	54.867	<2e-16 ***
URSTD	-6.71852	0.75339	998.00466	-8.918	<2e-16 ***
YEAR:URSTD	-0.35878	0.14970	998.00976	-2.397	0.0167 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

	(Intr)	YEAR	URSTD
YEAR	0.160		
URSTD	-0.436	-0.070	
YEAR:URSTD	-0.070	-0.436	0.160

2.B. Based on the model results, what's the estimated grand-mean 12th grade math score for a UR student? What's the estimated grand-mean 12th grade math score for a non-UR student?

Answer: The estimated grand-mean 12th grade math score for a UR student is 63.39, the expected grand-mean 12th grade score for a non-UR student is 71.91.

```
beta_00 <- fixef(model2)["(Intercept)"]
beta_10 <- fixef(model2)["YEAR"]
beta_01 <- fixef(model2)["URSTD"]
beta_11 <- fixef(model2)["YEAR:URSTD"]

mathscore_12th_nonUR <- beta_00 + beta_10*5
mathscore_12th_UR <- beta_00 + beta_01 + (beta_10 + beta_11)*5
mathscore_12th_nonUR
```


(Intercept)
71.90528

mathscore_12th_UR

(Intercept)
63.39288

2.C. Interpret the meaning of the rate of change in math scores from 7th to 12th grade for the UR students compared to the non-UR students. Make sure to indicate whether the group difference in the rate of change is statistically significantly different from zero.

Answer: On average, non-UR students' math score increases by 3.58 points per year. Compared with non-UR students, the UR students experience a slower growth rate, increasing by only 3.22 ($3.58 - 0.36$) points per year. The interaction terms between URSTD and YEAR is negative and statistically significant ($\beta_{11} = -0.36$, $p < 0.05$, $\alpha = 0.05$). This indicates that the achievement gap between UR and non-UR students widens over time due to the slower math growth rate among UR students.

2.D. What do the model results imply about UR student math preparation at the secondary grade level?

Answer: The results indicate that UR students start the secondary school with significantly lower math scores than non-UR students ($\beta_{01} = -6.72$, $p < 0.001$) and UR students also experience a lower rate of math score growth over time, the annual growth rate gap is 0.36 points less per year for UR students ($\beta_{11} = -0.36$, $p < 0.05$). By 12th grade the gap widens to 8.52 points ($6.72 + 0.36 * 5$) due to the slower learning rate.


Question 3

The foundation director heard that UR students are less likely to take Algebra 1 in 8th grade than non-UR students, and wonders if access to Algebra 1 could explain differences in math growth between UR and non-UR students. Fit the following model to test this hypothesis (Model 3):

$$Y_{ti} = \pi_{0i} + \pi_{1i}YEAR_{ti} + e_{ti}, e_{ti} \sim N(0, \sigma^2)$$

$$\pi_{0i} = \beta_{00} + \beta_{01}(URSTD_i) + \beta_{02}(ALGIN8_i) + r_{0i}, r_{0i} \sim N(0, \tau_{00})$$

$$\pi_{1i} = \beta_{10} + \beta_{11}(URSTD_i) + \beta_{12}(ALGIN8_i) + r_{1i}, r_{1i} \sim N(0, \tau_{11})$$

 Show your work

As part of your response to this question, include any R code and/or output you used to help answer the question.

3.A. Based on the model results, report the estimated grand-mean math score in 7th grade and 12th grade for the following types of students: - UR students who do not take Algebra in 8th grade - non-UR students who do not take Algebra in 8th grade - UR students who take Algebra in 8th grade - non-UR students take Algebra in 8th grade

Answer:

- Estimated math score for the 7th grade:

(1) UR students who do not take Algebra in 8th grade: 46.13

- (2) non-UR students who do not take Algebra in 8th grade: 51.29
- (3) UR students who take Algebra in 8th grade: 58.33
- (4) non-UR students take Algebra in 8th grade: 63.49

- Estimated math score for the 12th grade:

- (1) UR students who do not take Algebra in 8th grade: 61.77
- (2) non-UR students who do not take Algebra in 8th grade: 68.1
- (3) UR students who take Algebra in 8th grade: 78.91
- (4) non-UR students take Algebra in 8th grade: 85.24

```
model3 <- lmer(MTHSCORE ~ 1 + YEAR + URSTD + ALGIN8 + YEAR:URSTD + YEAR:ALGIN8 + (1 + YEAR | CASENUM)
summary(model3)
```

Linear mixed model fit by REML. t-tests use Satterthwaite's method [lmerModLmerTest]

Formula: MTHSCORE ~ 1 + YEAR + URSTD + ALGIN8 + YEAR:URSTD + YEAR:ALGIN8 + (1 + YEAR | CASENUM)

Data: hw

REML criterion at convergence: 38514.6

Scaled residuals:

Min	1Q	Median	3Q	Max
-4.7334	-0.4959	0.0029	0.5331	3.6676

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
CASENUM	(Intercept)	54.89	7.409	
	YEAR	2.29	1.513	0.30
Residual		17.65	4.201	

Number of obs: 6000, groups: CASENUM, 1000

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	51.29442	0.31533	996.99958	162.668	< 2e-16 ***
YEAR	3.36014	0.07152	996.99983	46.985	< 2e-16 ***
URSTD	-5.16387	0.65066	996.99958	-7.936	5.58e-15 ***
ALGIN8	12.19471	0.64055	996.99958	19.038	< 2e-16 ***
YEAR:URSTD	-0.23254	0.14757	996.99983	-1.576	0.115
YEAR:ALGIN8	0.99018	0.14527	996.99983	6.816	1.62e-11 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

	(Intr)	YEAR	URSTD	ALGIN8	YEAR:U
YEAR	0.061				
URSTD	-0.443	-0.027			
ALGIN8	-0.451	-0.028	0.126		
YEAR:URSTD	-0.027	-0.443	0.061	0.008	
YEAR:ALGIN8	-0.028	-0.451	0.008	0.061	0.126

```

beta_00 <- fixef(model3)["(Intercept)"]
beta_10 <- fixef(model3)["YEAR"]
beta_01 <- fixef(model3)["URSTD"]
beta_02 <- fixef(model3)["ALGIN8"]
beta_11 <- fixef(model3)["YEAR:URSTD"]
beta_12 <- fixef(model3)["YEAR:ALGIN8"]

mathscore_7th_nonUR_nonAlg <- beta_00
mathscore_7th_UR_nonAlg <- beta_00 + beta_01
mathscore_7th_nonUR_Algebra <- beta_00 + beta_02
mathscore_7th_UR_Algebra <- beta_00 + beta_01 + beta_02

mathscore_12th_nonUR_nonAlg <- beta_00 + beta_10*5
mathscore_12th_UR_nonAlg <- beta_00 + beta_01 + (beta_10 + beta_11)*5
mathscore_12th_nonUR_Algebra <- beta_00 + beta_02 + (beta_10 + beta_12)*5
mathscore_12th_UR_Algebra <- beta_00 + beta_01 + beta_02 + (beta_10 + beta_11 + beta_12)*5

#7th grade
#Non-UR, No Algebra in 8th Grade
round(mathscore_7th_nonUR_nonAlg, 2)

```

```

(Intercept)
    51.29

```

```

#UR, No Algebra in 8th Grade
round(mathscore_7th_UR_nonAlg, 2)

```

```

(Intercept)
    46.13

```

```

#Non-UR, Took Algebra in 8th Grade
round(mathscore_7th_nonUR_Algebra, 2)

```

```

(Intercept)
    63.49

```

```

#UR, Took Algebra in 8th Grade
round(mathscore_7th_UR_Algebra, 2)

```

```

(Intercept)
    58.33

```

```

#12th grade
#Non-UR, No Algebra in 8th Grade
round(mathscore_12th_nonUR_nonAlg, 2)

```

```
(Intercept)
      68.1
```

```
#UR, No Algebra in 8th Grade
round(mathscore_12th_UR_nonAlg, 2)
```

```
(Intercept)
      61.77
```

```
#Non-UR, Took Algebra in 8th Grade
round(mathscore_12th_nonUR_Alge, 2)
```

```
(Intercept)
      85.24
```

```
#UR, Took Algebra in 8th Grade
round(mathscore_12th_UR_Alge, 2)
```

```
(Intercept)
      78.91
```

3.B. Is taking Algebra in 8th grade related with a student's math growth rate through secondary school? Describe this relationship for the director, including whether the rate for Algebra students is faster, about the same, or slower than students who don't take Algebra in 8th grade, as well as whether the difference is statistically different from zero.

Answer: According to the model results, students who take Algebra in 8th grade experience a significantly faster math score growth rate through secondary school. They experienced a 0.99 points higher in the math score growth rate than students who do not take Algebra in 8th grade. This means that students who take Algebra, on average, gain 4.35 (3.36+0.99) points per year in math score. P-value for the interaction effect is less than 0.001 with $\alpha = 0.05$, indicating that the positive effect of Algebra on students' math score growth rate is not due to random chance.

3.C. Does controlling for Algebra enrollment alter the results concerning the disparity in math learning during secondary schooling between UR and non-UR students? If so, in what way(s)? If not, what led you to that conclusion?

Answer: Controlling for the Algebra enrollment alters the observed disparity between UR and non-UR students. There are several points to justify that: (1) based on the model comparison results, model3 significantly improved the model fit compared to model2, with the likelihood ratio test is highly significant ($p < 0.001$), meaning that adding the Algebra enrollment predictor improve the explanatory power meaningfully. (2) We can also see that the UR students' initial math score gap shrinks from -6.72 in model2 to -5.16 in model3, after controlling for the Algebra enrollment. And the UR students' slower growth rate (YEAR:URSTD) is no longer significant in model3, meaning that Algebra enrollment explains part of the slower growth observed in the UR students. (3) Lastly, we can see that the interaction term between YEAR and ALGIN is significant in model3, meaning that students who take Algebra in 8th grade has a 0.99 points faster growth per year than those who do not ($\beta_{12} = 0.99$, $p < 0.001$). This suggest that lower algebra enrollment among UR students maybe a potential factor to explain the slower growth seen in UR students in model2.

```
anova(model2, model3, test = "LRT")
```

Data: hw

Models:

model2: MTHSCORE ~ 1 + YEAR + URSTD + YEAR:URSTD + (1 + YEAR | CASENUM)

model3: MTHSCORE ~ 1 + YEAR + URSTD + ALGIN8 + YEAR:URSTD + YEAR:ALGIN8 + (1 + YEAR | CASENUM)

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
model2	8	38857	38911	-19421	38841			
model3	10	38528	38595	-19254	38508	333.38	2	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

3.D. Are there any other analyses you might want to conduct before drawing conclusions concerning disparities in math growth rates between UR and non-UR students? What do you recommend to help the director get a better understanding of the magnitude of the disparity and structural factors within the secondary education system that might affect the disparity?

Answer: There maybe several confounders we didn't include in the model, such as teach quality, school resouces, and family support that are likely to influence students' math growth rate. A three level analysis (time-student-school) may better capture these stractural factors. Then, I am also curious to know "dose taking Algebra in 8th grade provide the same benefit for UR students as it dose for non-UR students?", we could further include a three-way interaction (YEAR:URSTD:ALGINB) to test that. If UR students benefit less, it suggests additional barriers beyond just Algebra enrollment. Lastly, since its a longitudinal analysis, I was wondering dose the math growth gap change (flatten out or sharpened) as the year gose, may be a seperate time peride coding strategy could use or maybe to consider fitting the data with a nonlinear growth model (probably as we taught in class, quadratic term for year) to see how the grow rate change over time.

Note

There's no right or wrong answer to 3.D. I'm just looking for a brief reflection on the meaning of the analysis you just conducted and its limitations in the context of all the material we covered about conducting a longitudinal analysis.