Using SAS® Proc Mixed for the Analysis of Clustered-Longitudinal Data

Kathy Welch
Center for Statistical Consultation and Research
The University of Michigan

Background

- Proc Mixed can be used to fit Linear Mixed Models (LMMs) for repeated measures/longitudinal or clustered data
- In this example, we demonstrate the use of Proc Mixed for the analysis of a clustered-longitudinal data set
- The data we will use is derived from the Longitudinal Study of American Youth (LSAY, ICPSR 30263).
  - A group of 3116 students in 52 schools were followed from 1987-1994, when they were in grades 7 through 12.

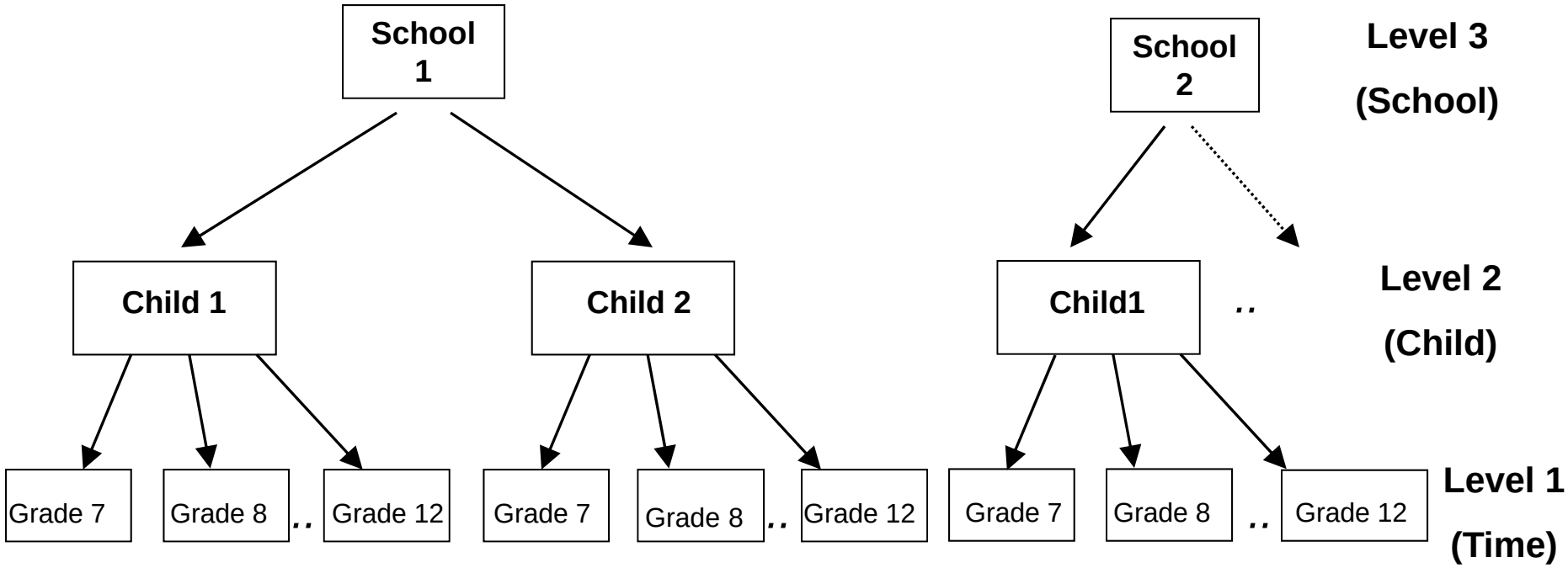
Purpose of the LSAY Study

- To examine student attitudes toward and achievement in mathematics and science
- To examine student interest in and plans for a career in science, mathematics, or engineering, during middle school, high school, and the first four years post-high school
- To estimate the relative influence of parents, home, teachers, school, peers, media, and selected informal learning

The LSAY Data Set

- The current example includes 796 students (a 25% stratified random sample) from the 52 schools in the LSAY
- We will be examining the relationship of some parent variables, as well as student variables with math achievement over time.
- Some covariates were time-varying (e.g., math achievement, student enjoyment of working on tough problems, parent academic push)
- Some variables were constant for a student (e.g., gender, student likes math, parent education level)

LSAY Data Structure



Level 3 Variables: No School-Level covariates are included

Level 2 Variables: Gender, Likes Math, Parent Education

Level 1 Variables: Math Achievement, Parent Academic Push, Tough Problems

Analysis Plan

- We will begin by checking descriptive statistics and graphs to illustrate the relationships between predictors and the outcome (math achievement)
- Cautionary Example of how to graph and interpret the effects of time-varying covariates
- We will consider LMMs with a random intercept for each school, plus a random intercept and random slope per student
- We will explore the fixed effects of parent variables and student variables on baseline math achievement and on the slope of math achievement over time

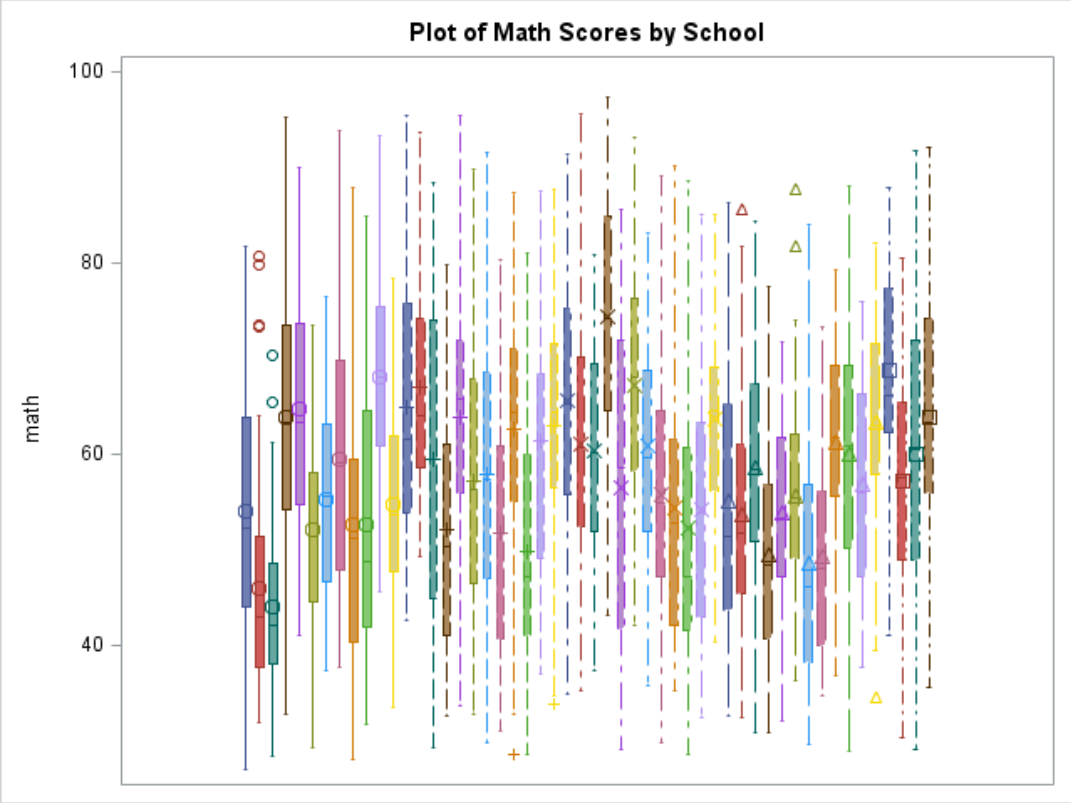
Stratified Random Sample

- Data set must be sorted by schoolid so that schoolid can be a stratifying variable for the sample
- This stratified sample will select a 25% random sample from each school
- This code can be run the first time with no seed. Then copy and include the seed for future runs.

```
proc sort data=two;
  by schoolid;
run;
proc surveyselect data=two method=srs outall
                  samprate=0.25 out=sampdat2
                  seed=507002001;
  strata schoolid;
run;
```

Variability of Math Scores Across Schools

```
title "Plot of Math Scores by School";
proc sgplot data=mathlong noautolegend;
  vbox math/ group=schoolid;
run;
```

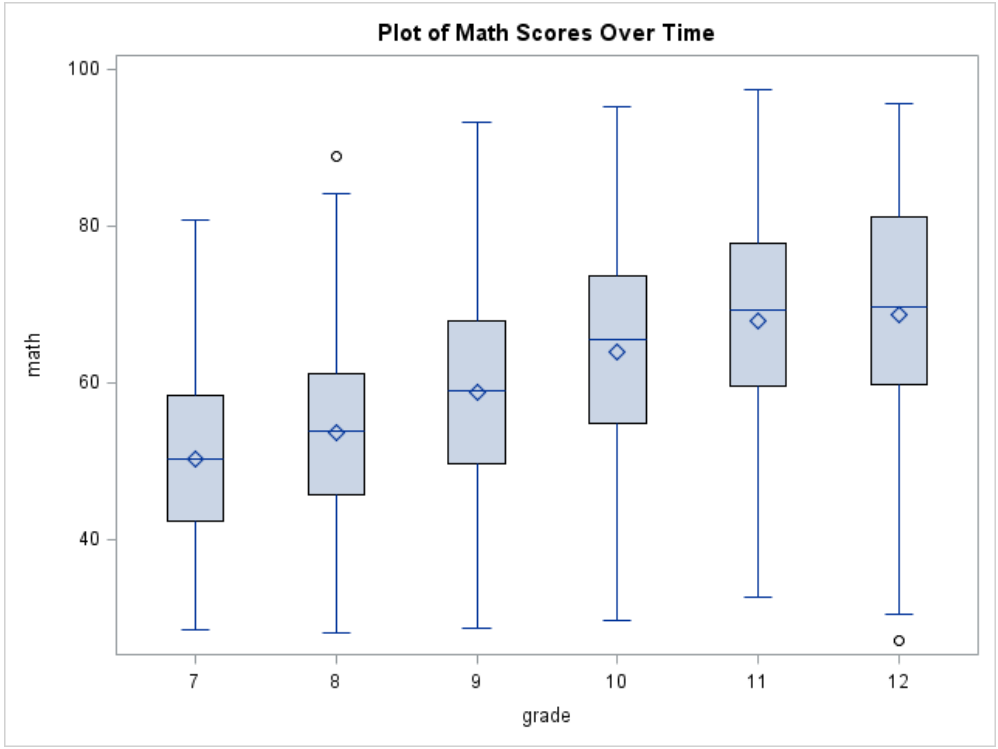


Variability of Math Scores by Schools

- There appears to be a fair amount of variability between schools in terms of their math scores
- We will want to include a random intercept for each school to capture this variability and to allow the scores for students within the same school to be correlated
- We do not have school-level variables to include in this model, but we may want to explore further whether any school-level variables (e.g., Rural/Urban) can explain these differences between schools

Math Achievement Over Time

```
proc sgplot data=math;
   vbox math / category=grade;
run;
```

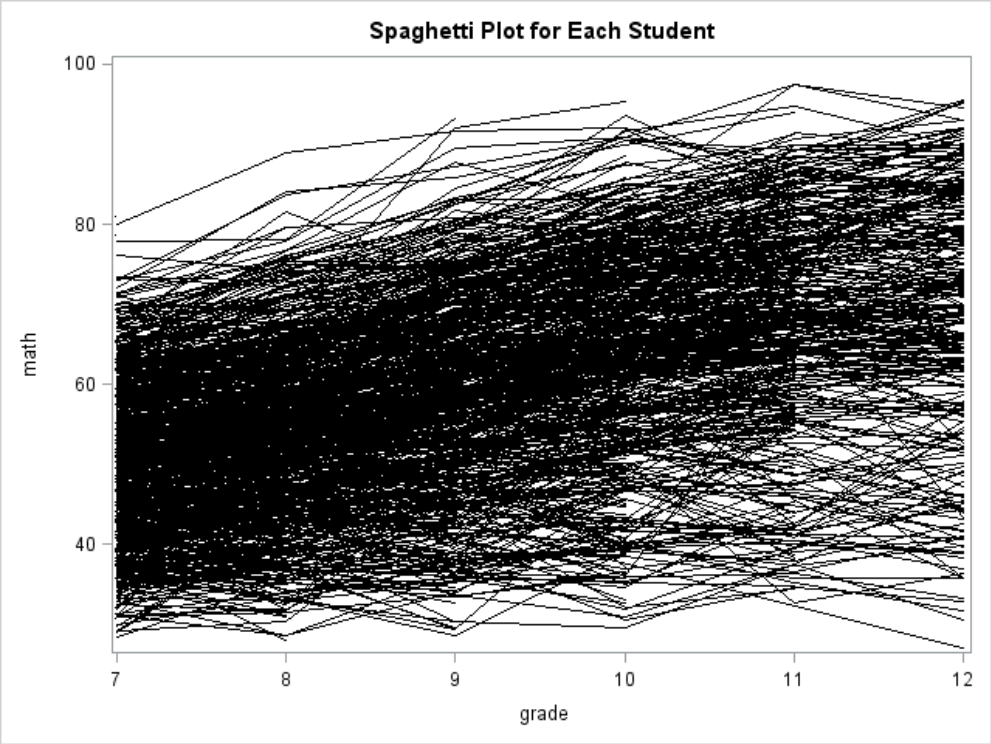


Math Achievement Over Time

- Math achievement increases over time
- This is expected
- We want to explore whether the rate of increase over time differs by child and whether child covariates or parent covariates can help to explain differences in the rate of increase

Spaghetti Plots

```
title "Spaghetti Plot for Each Student";
proc sgplot data=math noautolegend;
   series y=math x=grade / group=caseid
      lineattrs = (thickness=.5 pattern=1 color=black);
run;
```

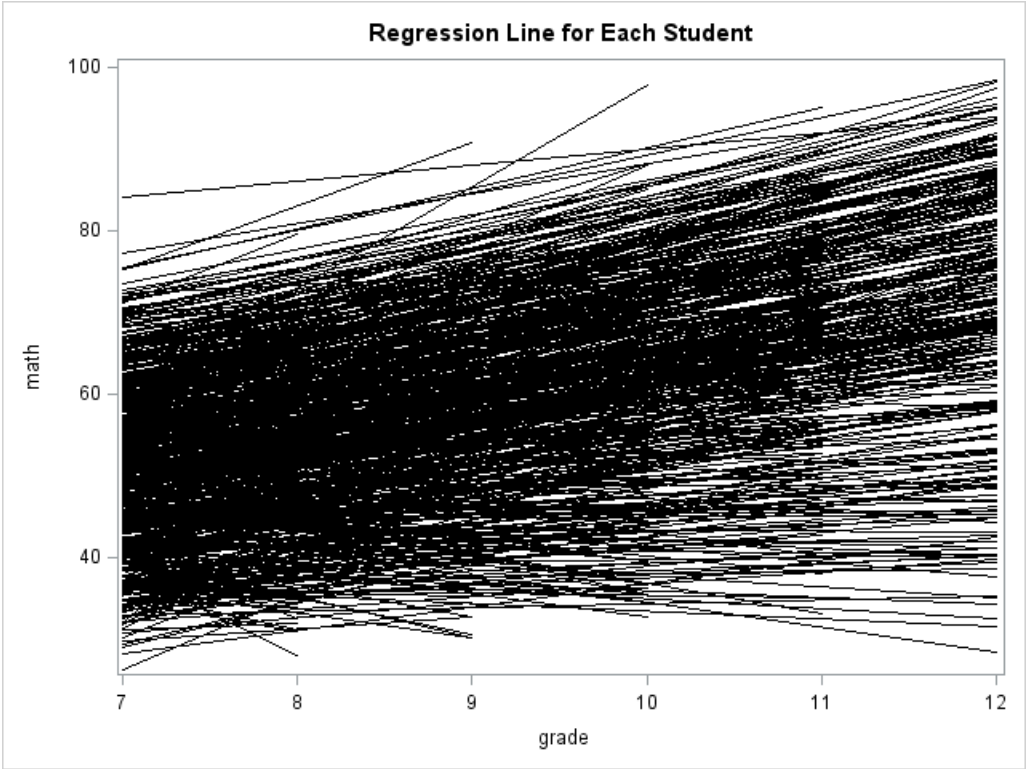


Spaghetti Plots for Each Child

- There is a general increase in math achievement over time across students
- Some students have a steeper slope over time and others have a lower slope
- The spaghetti plots emphasize the erratic nature of the math scores (they are not all smoothly increasing for each child)
- A random sample of children would more clearly show individual patterns (we don't show this)

Regression Line for Each Student

```
title "Regression Line for Each Student";
proc sgplot data=math noautolegend;
  reg y=math x=grade / group=caseid nomarkers
     lineattrs = (thickness=.5 pattern=1 color=black);
run;
```

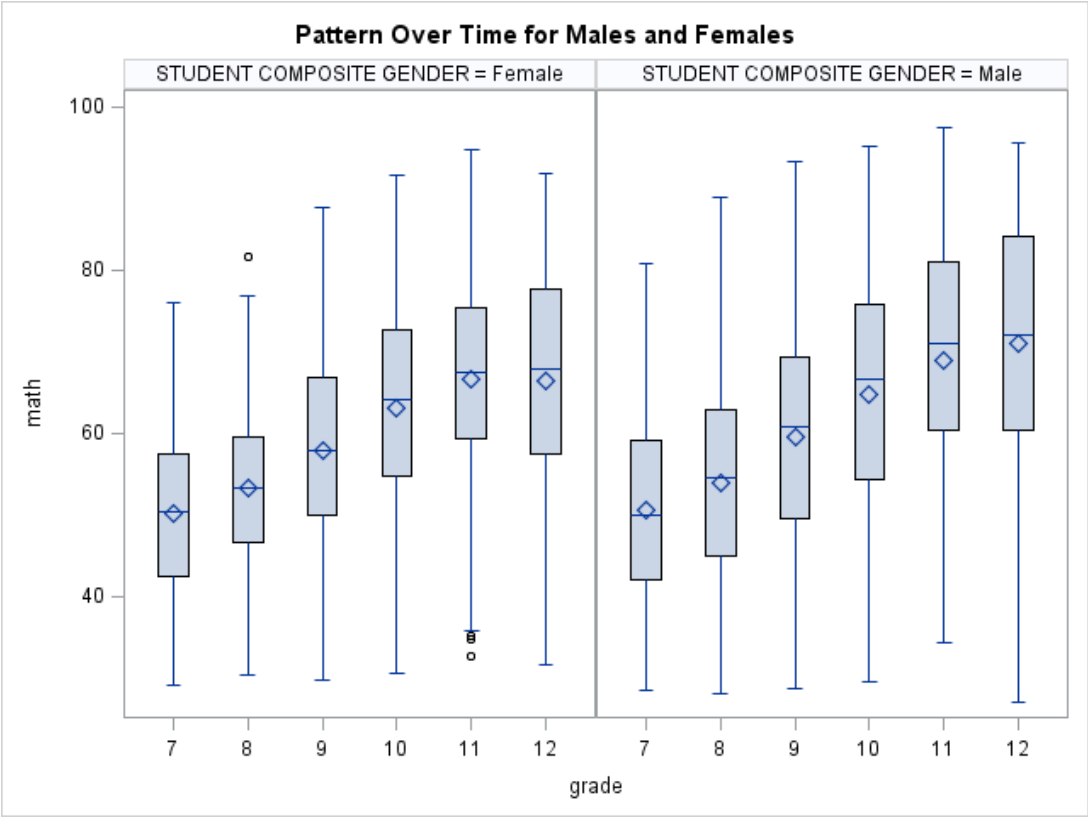


Regression Plots

- Regression plots for each child over time smooth out the relationship between grade and math achievement for each child
- Different children have higher or lower intercepts and steeper or more shallow slopes over time
- We may want to consider including a random intercept for each child and a random slope for time in our LMM specification

Gender Boxplots

```
title "Pattern Over Time for Males and Females";
proc sgpanel data=math;
  panelby gender;
  vbox math / category=grade;
  format gender genderfmt.;
run;
```

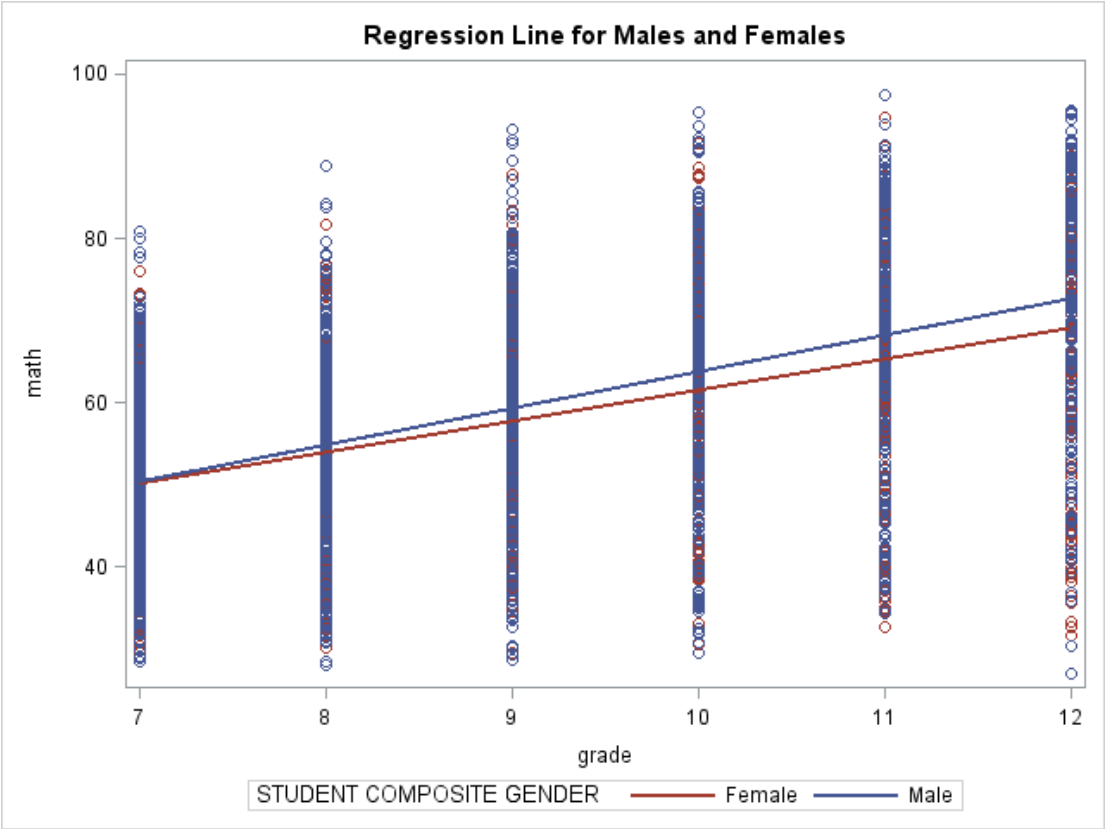


Box Plots for Gender Over Time

- The paneled boxplot shows that math scores increase generally for boys and for girls
- The variability at each grade does not appear to differ, nor does it appear to be different for boys vs. girls
- We next look at a regression line over time for males and for females

Regressions for Males and Females

```
title "Regression Line for Males and Females";
proc sgplot data=math;
   reg y=math x=grade / group=gender ;
   format gender genderfmt.;
run;
```

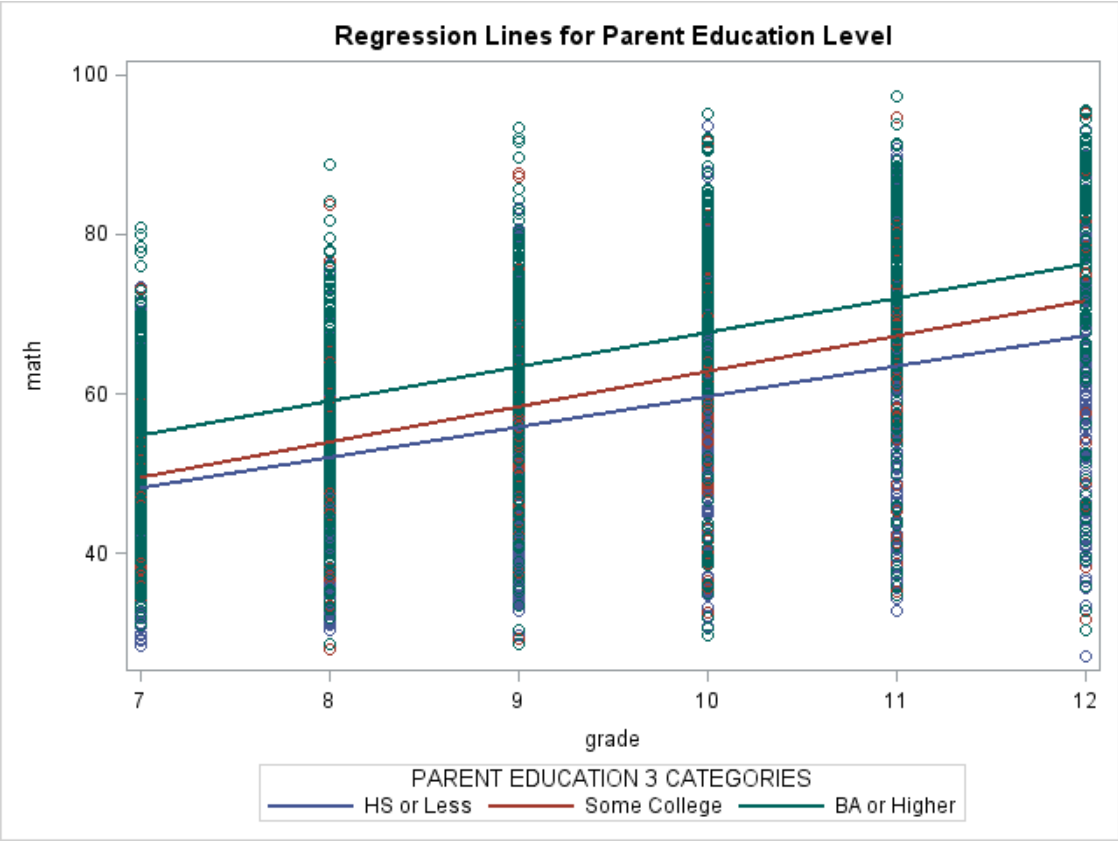


Regression Lines for Gender

- Based on these regression plots, it appears that gender does not explain much of the variability in math achievement scores
- Boys and girls appear to start out at a very similar level of math achievement
- The slope over time appears to be a bit steeper for boys than for girls, so we may expect to see a difference between boys and girls by 12^{th} grade (or sooner)
- We may want to include an interaction between gender and grade to see if this is true

Regressions for Parent Education

```
title "Regression Lines for Parent Education Level";
proc sgplot data=math;
  where peduc3 not=.;
  reg y=math x=grade / group=peduc3 ;
  format peduc3 peduc3fmt.;
run;
```

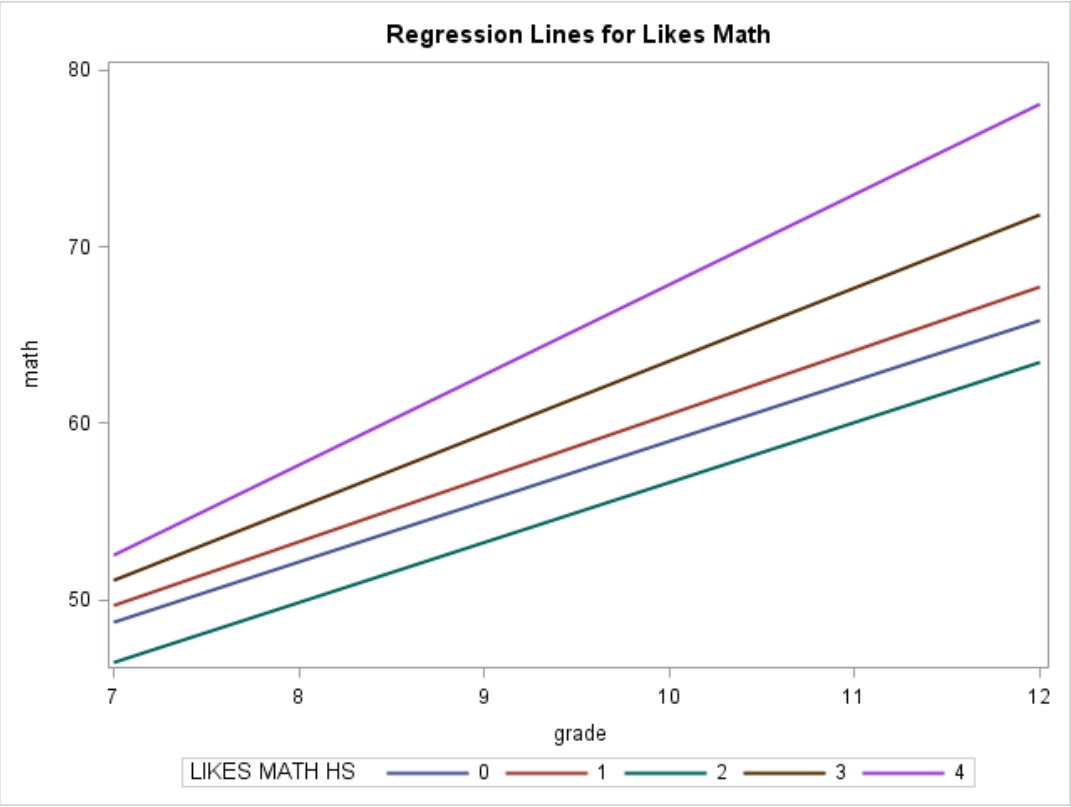


Regression Lines for Parent Education

- Based on these regression plots, it appears that children of parents with a BA or greater education level tend to do better in math achievement at 7th grade than children in the other two categories.
- As time goes on, the difference between BA or greater and some college appears to diminish while the difference between HS or less and some college appears to get wider
- We may want to include an interaction between time and parent education to see if this is true

Regressions for Likes Math Levels

```
title "Regression Lines for Likes Math";
proc sgplot data=math;
  where likemth not=.;
  reg y=math x=grade / group=likemth nomarkers;
run;
```

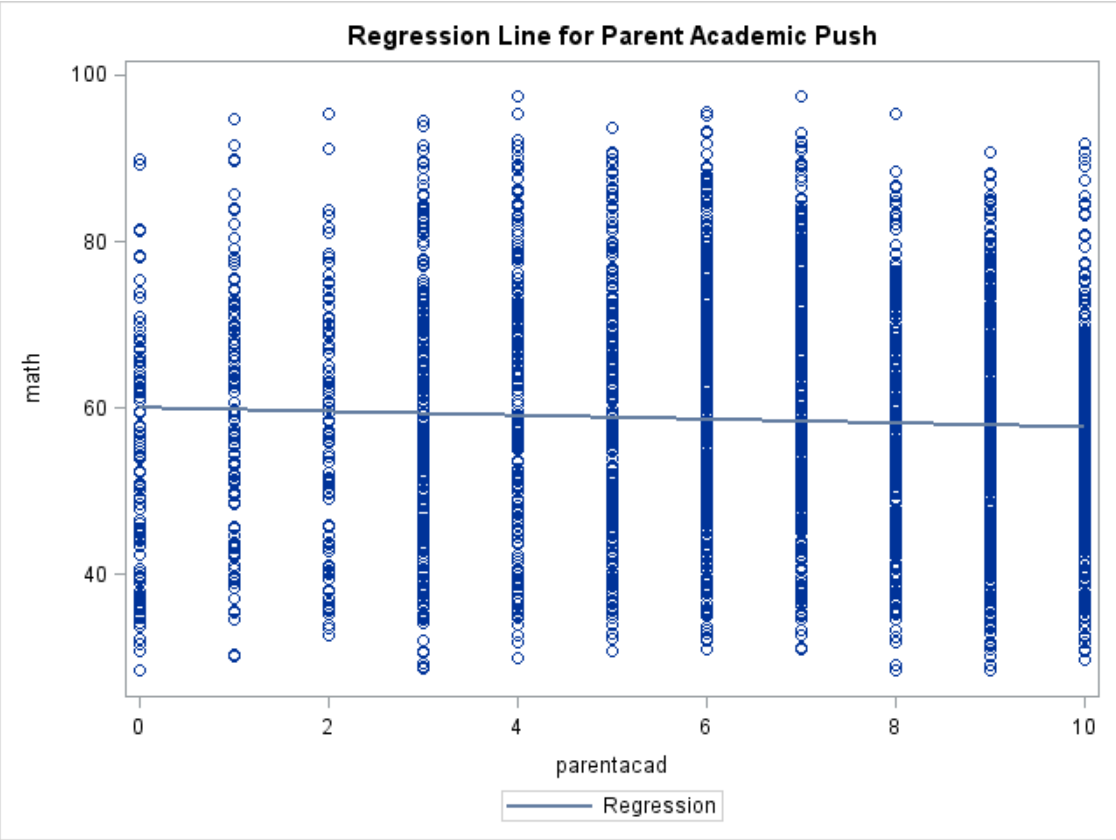


Regression Lines for Likes Math

- Students who like math more tend to do better in math achievement scores in 7th grade and then continue to do better across the grades.
- As time goes on, the difference between the kids who really love math and the other groups tends to get larger
- This indicates that including an interaction between grade and likes math may be a good idea

Regression Line for Parent Academic Push

```
title "Regression Line for Parent Academic Push";
proc sgplot data=math;
  reg y=math x=parentacad ;
run;
```

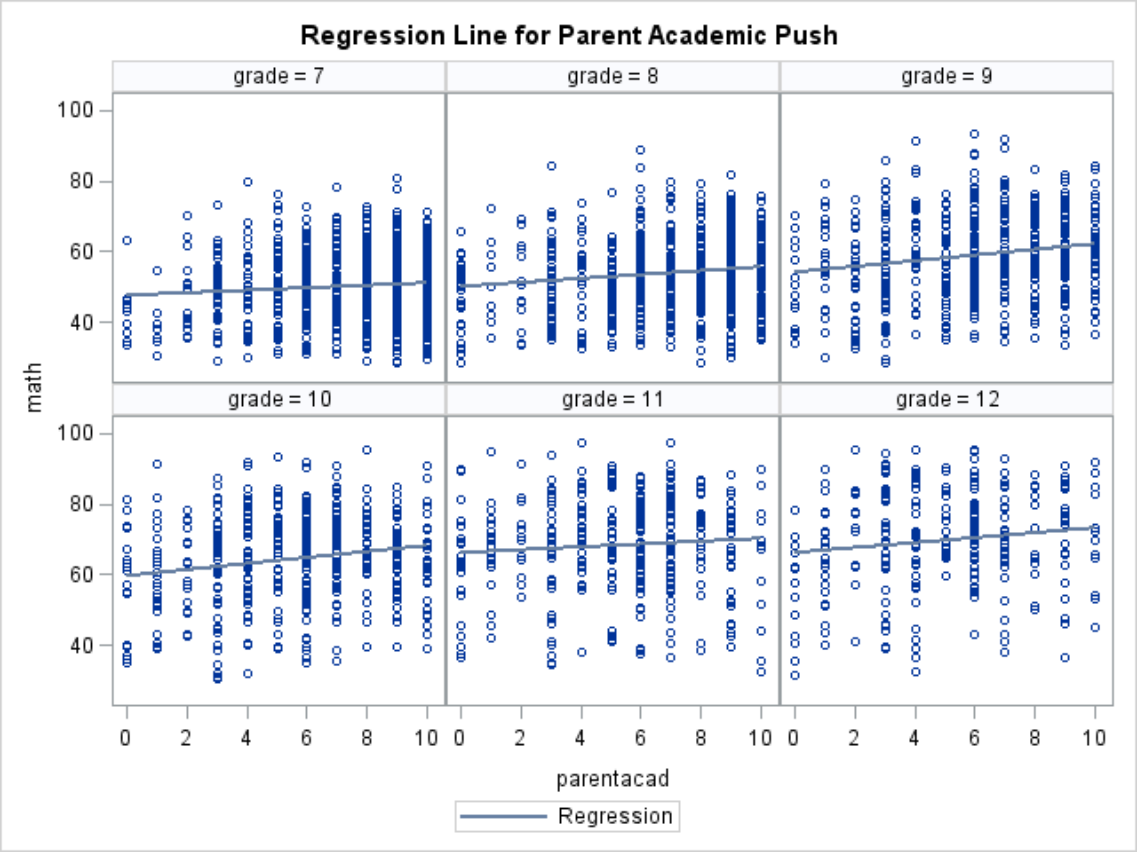


Is Parent Academic Push Negatively Related to Math Achievement?

- We expect that children of parents who push their children more academically will do better in math, but we see an apparent negative relationship between parent academic push and math achievement
- Is this real?
- Recall that parent academic push was measured at each time point
- We need to see the effects of parent academic push within each grade

Regression Line for Parent Academic Push within Each Grade

```
proc sgpanel data=math;
  panelby grade;
  reg y=math x=parentacad;
run;
```



Is Parent Academic Push Actually Negatively Related to Math Achievement?

- Within each grade, the relationship between parent academic push and math achievement is positive, yet the overall relationship is negative
- How can this be?
- Let's look at descriptive statistics for parent academic push and for math achievement for each grade

Descriptive Statistics by Grade

grade	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
7	796	math	786	50.3053562	10.1589309	28.4600000	80.8300000
		parentacad	793	7.3959647	2.3825086	0	10.0000000
8	796	math	660	53.6551515	10.9019506	28.0000000	88.8400000
		parentacad	697	6.6499283	2.8285751	0	10.0000000
9	796	math	581	58.6595353	12.5994706	28.6500000	93.2900000
		parentacad	600	6.0000000	2.7151956	0	10.0000000
10	796	math	505	63.9117822	13.5109215	29.6100000	95.2600000
		parentacad	538	5.4330855	2.7678963	0	10.0000000
11	796	math	409	67.8204156	13.9069185	32.6700000	97.3900000
		parentacad	458	5.2576419	2.7466304	0	10.0000000
12	796	math	299	68.7263545	15.3685745	27.0100000	95.5700000
		parentacad	440	4.8386364	2.7794983	0	10.0000000

What is Going On?

- Notice that math achievement is increasing across grades (as we saw in the previous graphs)
- Parent academic push is decreasing over time (parents are apparently pushing their kids less academically as they grow older)
- The result is that overall it looks like kids of parents who push more do worse, but actually, within the same grade, kids of parents who push more have higher math scores for every grade

What is Going On?

- Notice that math achievement is increasing across grades (as we saw in the previous graphs)
- Parent academic push is decreasing over time (parents are apparently pushing their kids less academically as they grow older)
- The result is that overall it looks like kids of parents who push more do worse, but actually, within the same grade, kids of parents who push more have higher math scores for every grade

Data Management

- We recode GRADE to center it at Grade 7, by subtracting 7 from each value, to make the intercept more interpretable
- We center LIKEMTH by subtracting the mean from each value (to give us an overall mean of zero).

```
data math2;
  set math;
  centgrade=grade-7;
  if peduc3=1 then r_peduc=0;
  if peduc3=3 then r_peduc=1;
  if peduc3=4 then r_peduc=2;

  centlikemth = likemth-2.42;
run;
```

Initial (Null) Model

- We fit an initial model with no fixed predictors
- We include a random intercept for each school and a random intercept for each student nested within school

```
title "Initial Null Model";
proc mixed data=math2 covtest noclprint order=internal;
   class schoolid caseid r_peduc gender;
   model math = / solution;
   random schoolid;
   random int / subject=caseid(schoolid) ;
run;
```

Covariance Parameter Estimates for Initial Model

- The estimate for SCHOOLID is the variance between schools,
- The intercept for caseid(Schoolid) is the variance of students nested within schools
- Residual is the estimated variance within a student

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
SCHOOLID		28.3660	7.2728	3.90	<.0001
Intercept	CASEID(SCHOOLID)	100.38	6.2671	16.02	<.0001
Residual		67.2217	1.9236	34.95	<.0001

Fixed Effects Estimates for Initial Model

- The Intercept represents the predicted mean of math achievement across all grades, for an average school and an average student

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	56.9062	0.8415	51	67.63	<.0001

New Model with Parent Academic Push as the Only Fixed Predictor

- We now illustrate (the incorrect) LMM including parent academic push as the only fixed predictor

```
title "Model with Parent Academic Push Only";
proc mixed data=math2 covtest noclprint order=internal;
    class schoolid caseid r_peduc gender;
    model math = parentacad / solution;
    random schoolid;
    random int / subject=caseid(schoolid) ;
run;
```

Results from Model with Parent Academic Push as the Only Fixed Predictor

- The estimate for Parent Academic Push (parentacad) is negative (-1.0576) and significant (p<0.0001)

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	63.3137	0.9732	51	65.06	<.0001
parentacad	-1.0576	0.07047	2219	-15.01	<.0001

Model with Grade and Parent Academic Push

- We now include the fixed effect of CentGrade and a random slope for centgrade for each student

```
title "Model with Grade and Parent Academic Push";
proc mixed data=math2 covtest noclprint order=internal;
    class schoolid caseid r_peduc gender;
    model math = centgrade parentacad / solution;
    random schoolid;
    random int centgrade/ subject=caseid(schoolid) ;
run;
```

Model with Grade and Parent Academic Push

- The effect of parent academic push is now comparing two students who have different levels of academic push, but are in the same grade
- Notice that we now get a positive effect of parent academic push
- The effect of centgrade is positive, as expected

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	49.5916	0.8281	51	59.89	<.0001
centgrade	3.7301	0.1019	707	36.59	<.0001
parentacad	0.1203	0.05309	1511	2.27	0.0236

Final Model

- Our final model includes some (but not all) of the fixed effects that we considered earlier.
- Model selection steps are not shown

```
ods graphics on;
proc mixed data=math2 covtest noclprint order=internal;
   class schoolid caseid r_peduc gender;
   model math =  centgrade gender centgrade*gender
                 centlikemth r_peduc r_peduc*centgrade
                  / residual solution;
   random schoolid;
   random int centgrade /solution subject=caseid(schoolid) type=un;
   ods output solutionR = eblupsdat;
   ods exclude solutionR;
```

Final Model Covariance Parameter Estimates

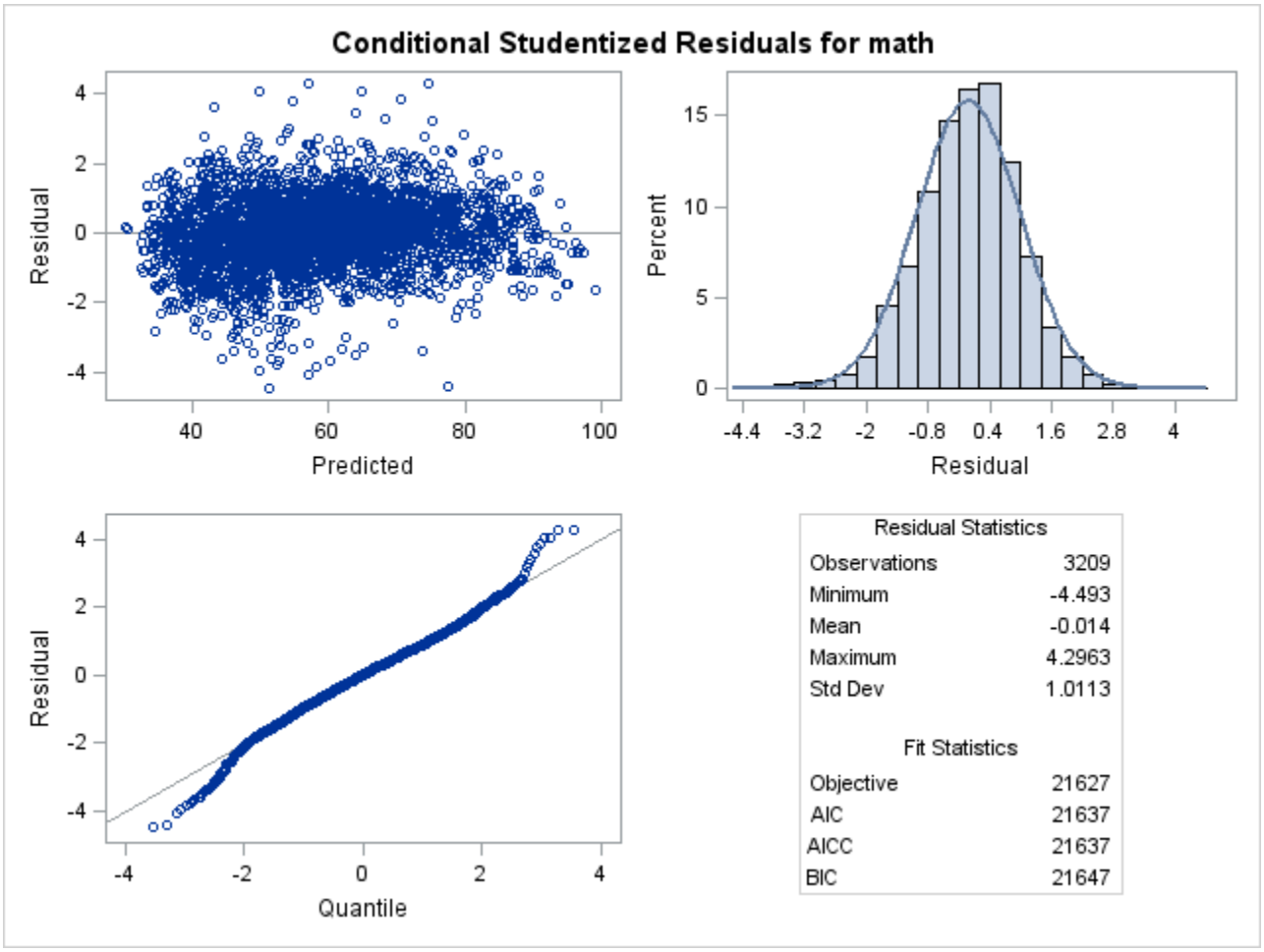
- SchoolID is the variance estimate between schools
- Un(1,1) is the variance of the random child intercepts
- Un(2,1) is the covariance between the child intercepts and slopes
- Un(2,2) is the variance of the random child slopes
- Residual is the variance of the residuals around each child's regression line

Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
SCHOOLID		13.9340	3.9398	3.54	0.0002
UN(1,1)	CASEID(SCHOOLID)	65.0081	4.1772	15.56	<.0001
UN(2,1)	CASEID(SCHOOLID)	4.7149	0.8575	5.50	<.0001
UN(2,2)	CASEID(SCHOOLID)	3.0705	0.3208	9.57	<.0001
Residual		21.4741	0.7245	29.64	<.0001

Final Model Fixed Effects Estimates

Solution for Fixed Effects							
Effect	r_peduc	STUDENT COMPOSITE GENDER	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			53.5553	0.8487	51	63.11	<.0001
centgrade			4.1117	0.1799	722	22.86	<.0001
GENDER		Female	-0.05989	0.6525	1702	-0.09	0.9269
GENDER		Male	0	.	.	.	.
centgrade*GENDER		Female	-0.4052	0.1827	1702	-2.22	0.0267
centgrade*GENDER		Male	0	.	.	.	.
centlikemth			0.9950	0.2700	1702	3.69	0.0002
r_peduc	0		-4.6057	0.7778	1702	-5.92	<.0001
r_peduc	1		-3.6799	1.0503	1702	-3.50	0.0005
r_peduc	2		0	.	.	.	.
centgrade*r_peduc	0		-0.6895	0.2031	1702	-3.40	0.0007
centgrade*r_peduc	1		-0.3980	0.2918	1702	-1.36	0.1728
centgrade*r_peduc	2		0	.	.	.	.

Final Model Residual Diagnostics



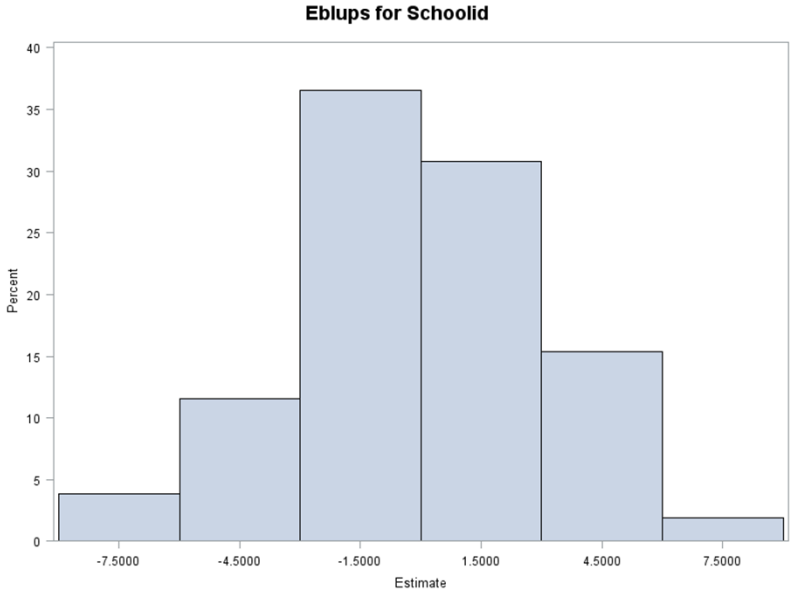
Check the Distribution of Eblups for Random Effects

- We use the output data set (Eblupsdat) generated from the Proc Mixed run to get the Eblups for each school and check the distribution

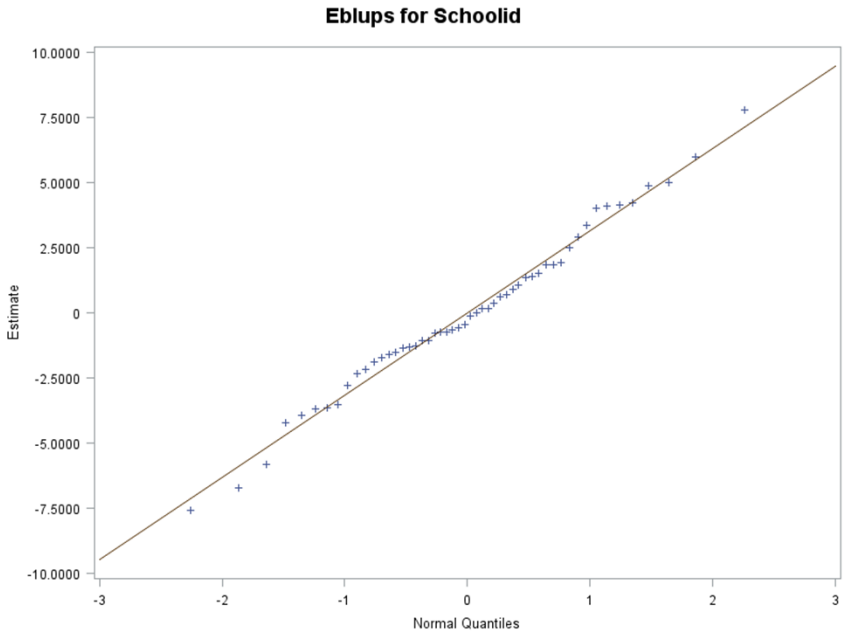
```
title "Eblups for Schoolid";
proc univariate data=eblupsdat;
  where effect="SCHOOLID";
  var estimate;
  histogram;
  qqplot/normal(mu=est sigma=est);
run;
```

Eblups for SchoolID

Histogram



Normal Q-Q Plot



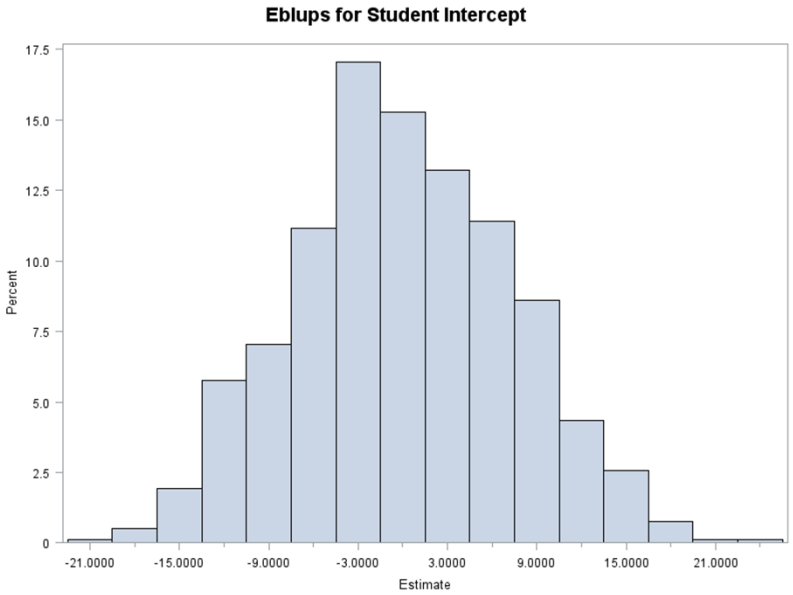
Check the Distribution of Eblups(Cont)

- We use the output data set (Eblupsdat) generated from the Proc Mixed run to get the Eblups for each intercept and slope for a child and check the distribution

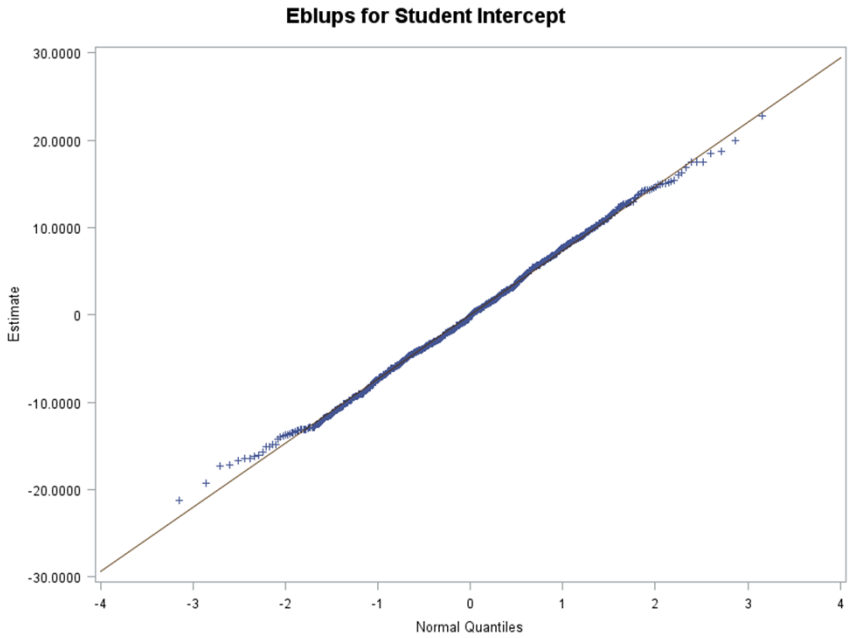
```
title "Eblups for Student Intercept";
proc univariate data=eblupsdat;
  where effect="Intercept";
  var estimate;
  histogram;
  qqplot/normal(mu=est sigma=est);
run;
title "Eblups for Student Slope";
proc univariate data=eblupsdat;
  where effect="centgrade";
  var estimate;
  histogram;
  qqplot/normal(mu=est sigma=est);
run;
```

Eblups for Random Student Intercept

Histogram

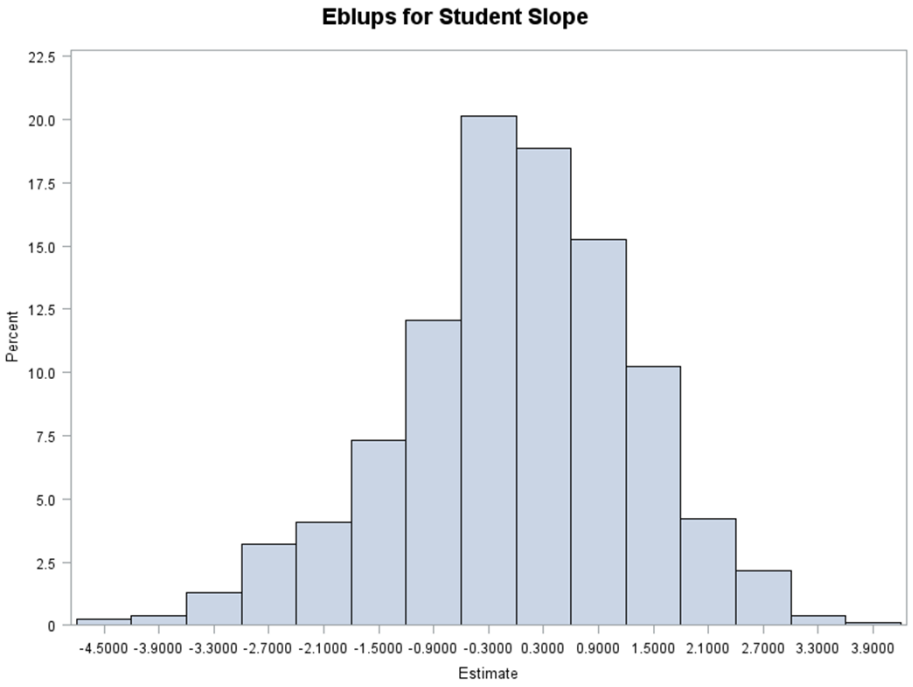


Normal Q-Q Plot

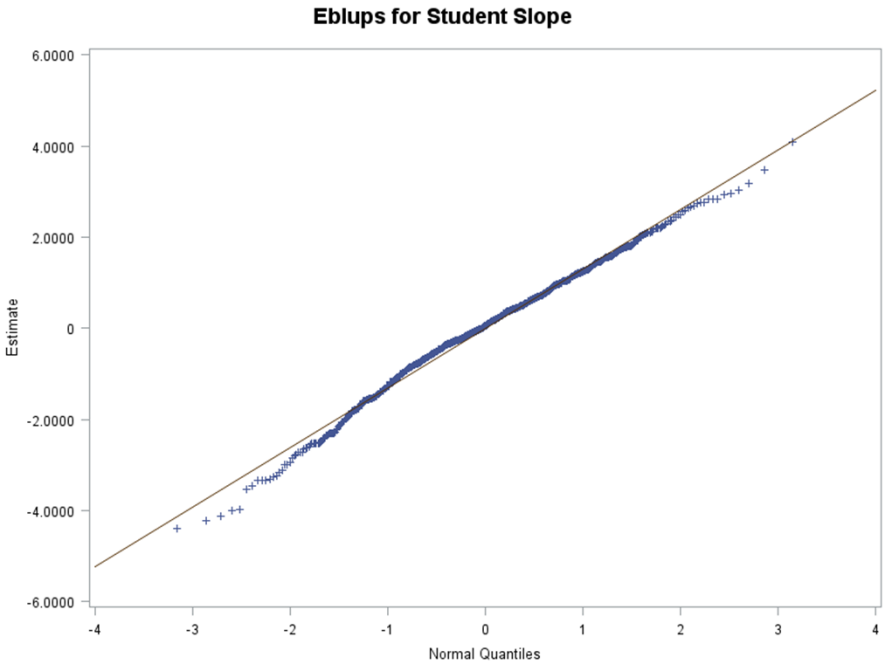


Eblups for Random Student Slope

Histogram



Normal Q-Q Plot



Summary

- Proc Mixed is a flexible tool for fitting models for clustered-longitudinal data
- Care must be taken when including time-varying predictors in the model to be sure that the interpretation of their effects is correct
- Graphics can help to understand the data before the model-fitting process begins

References: Software and Data

- The output, code and data analysis for this presentation were generated using SAS/STAT software, Version 9.3 (TS1M0) of the SAS System for Windows. Copyright ©2002-2010 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.
- The data used for these examples were derived from the Longitudinal Study of American Youth (ICPSR 30263)
- Kathy Welch is responsible for any errors or omissions in this analysis

References

- Verbeke, G., and Molenberghs, G. Linear Mixed Models for Longitudinal Data, Springer, New York, 2000.
- West, Brady T., Welch, Kathleen B., Galecki, Andrzej T., Linear Mixed Models: A Practical Guide, Chapman & Hall, 2007.