Multilevel Regression Models
Linear Mixed Models
Hierarchical Linear Models
Random Coefficient Models

November 18, 2003

Steve Gregorich

Standard Linear Regression

- Standard linear regression

  All observations are independent

  Units of analysis represent a single level of abstraction, e.g., patients

  Fixed effects only—parameters are unit-generic

$$y_i = \beta 0 + \beta 1 \times x1_i + \beta 2 \times x2_i + \varepsilon_i$$
$$health_i = intercept + \beta 1 \times education_i + \beta 2 \times income_i + residual_i$$
$$grade_i = intercept + \beta 1 \times gender_i + \beta 2 \times essay_i + residual_i$$

subscript *i* represents individual respondents

Example Data Set for Standard Linear Regression

Student ID	Grade	Gender	Essay Score
1	A	0	78
2	C	1	70
3	B	0	85
...	...	...	..
1205	A	1	93

Here there is one unit of analysis level, individual students

All students are considered to be independent

There is one record of data per unit (i.e., student)

Multilevel Linear Regression

Data can be represented by a set of nested levels

Each level represents a unit of analysis

Clustered sampling

Repeated measures

Fixed and random parameters

Fixed parameters are unit-generic

Random parameters are unit-specific (more later)

Big concern: Not all observations are independent

Examples of Clustered Data

Clustered Data

- A three-level data structure

  Schools, classrooms with schools, students within classrooms

  "Level-3" schools
  "Level-2" classrooms within schools
  "Level-1" students within classrooms

- Two-level data structures

  married couples, individuals within couples
  primary sampling units (e.g., area codes), households within PSUs

- Notes.

Covariates can be measured at any level
Outcome data is measured at level-1
Observations nested within higher-level units not assumed independent

Examples of longitudinal data

Longitudinal Data

- A two-level data structure
  Repeated measures "clustered" within individuals

  "Level-2" - Individuals
  "Level-1" - Repeated measures within individuals

- Note
  Repeated measures on the same individual not assumed independent

- Combinations of Clustered and Longitudinal Data

Schools, students within schools, repeated measures within students
"Level-3" - schools
"Level-2" - students within schools
"Level-1" - repeated measures nested within students

Example Data Set for Multilevel Regression

2-level data structure

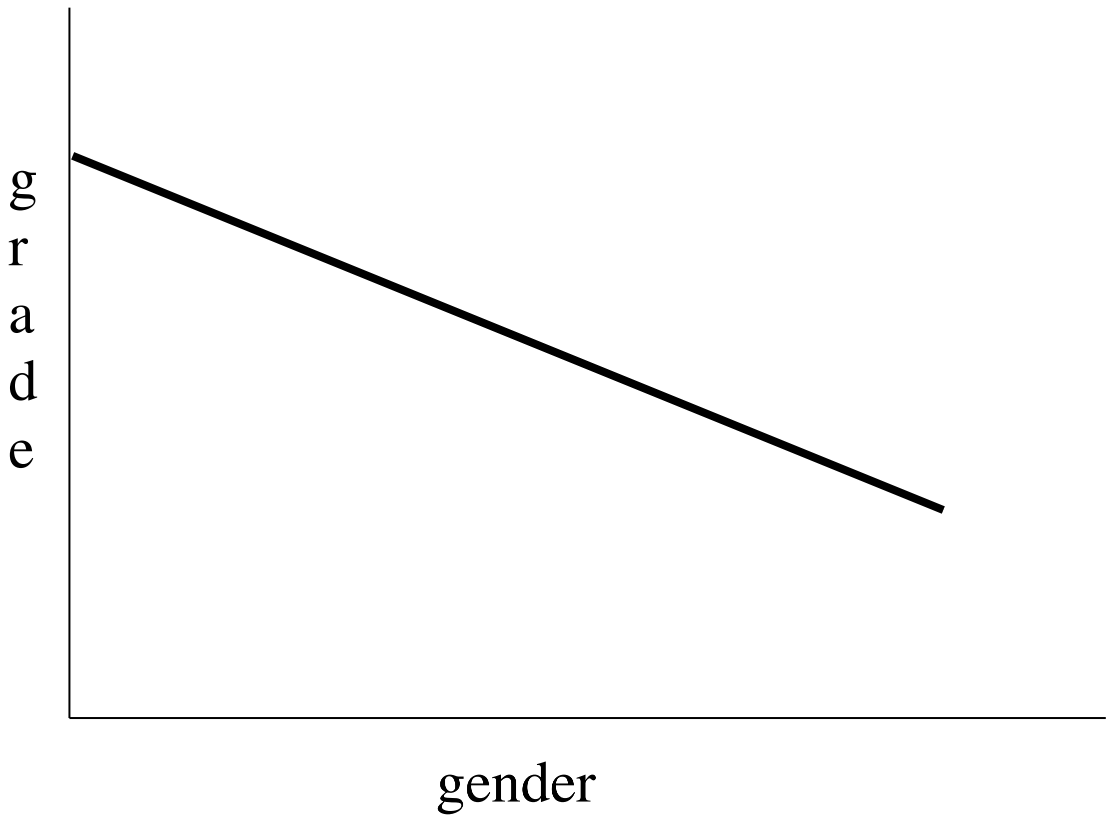
Level-2 = schools

Level-1 = students within schools

School D	Student ID	Grade	Gender	Essay Score
1	1	4	1	73
1	2	2	1	85
1	3	4	0	95
2	4	2	1	75
2	5	3	0	80
3	6	4	0	83
...	...	...	...	...
100	205	4	1	90
100	206	3	0	78

Schools are independent. Students w/in schools are not

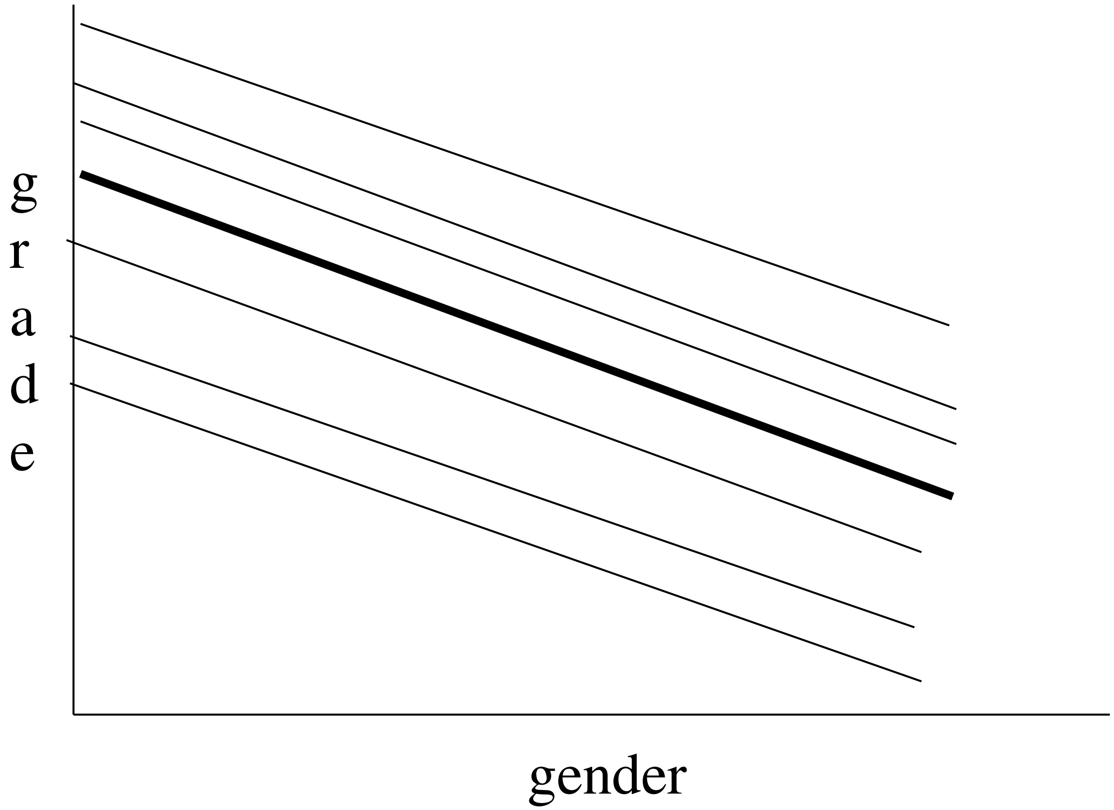
Graphical Depiction of Standard Linear Regression



$$grade_i = \beta 0 + \beta 1 \times gender_i + \varepsilon_i$$

subscript *i* represents individual students

Graphical Depiction of Multilevel Linear Regression



$$grade_{ij} = \beta 0_j + \beta 1 \times gender_{ij} + \varepsilon_{ij}$$

subscripts *i* and *j* represent students and schools, respectively

Multilevel Linear Regression

$$grade_{ij} = \beta 0_j + \beta 1 \times gender_{ij} + \varepsilon_{ij}$$

$\beta 0_j$ the intercept for school *j*, which varies by school, a random effect

$\beta 1$ the effect of gender on grades, which is constant, a fixed effect

$gender_{ij}$ the gender of student *i* in school *j*

ε_{ij} the student-level residual

Usually, the ε_{ij} are not output, but their variance is estimated, $\hat{\sigma}_{\varepsilon}^2$

this is known as the within-schools *variance component*

Multilevel Linear Regression

The same model can be re-expressed

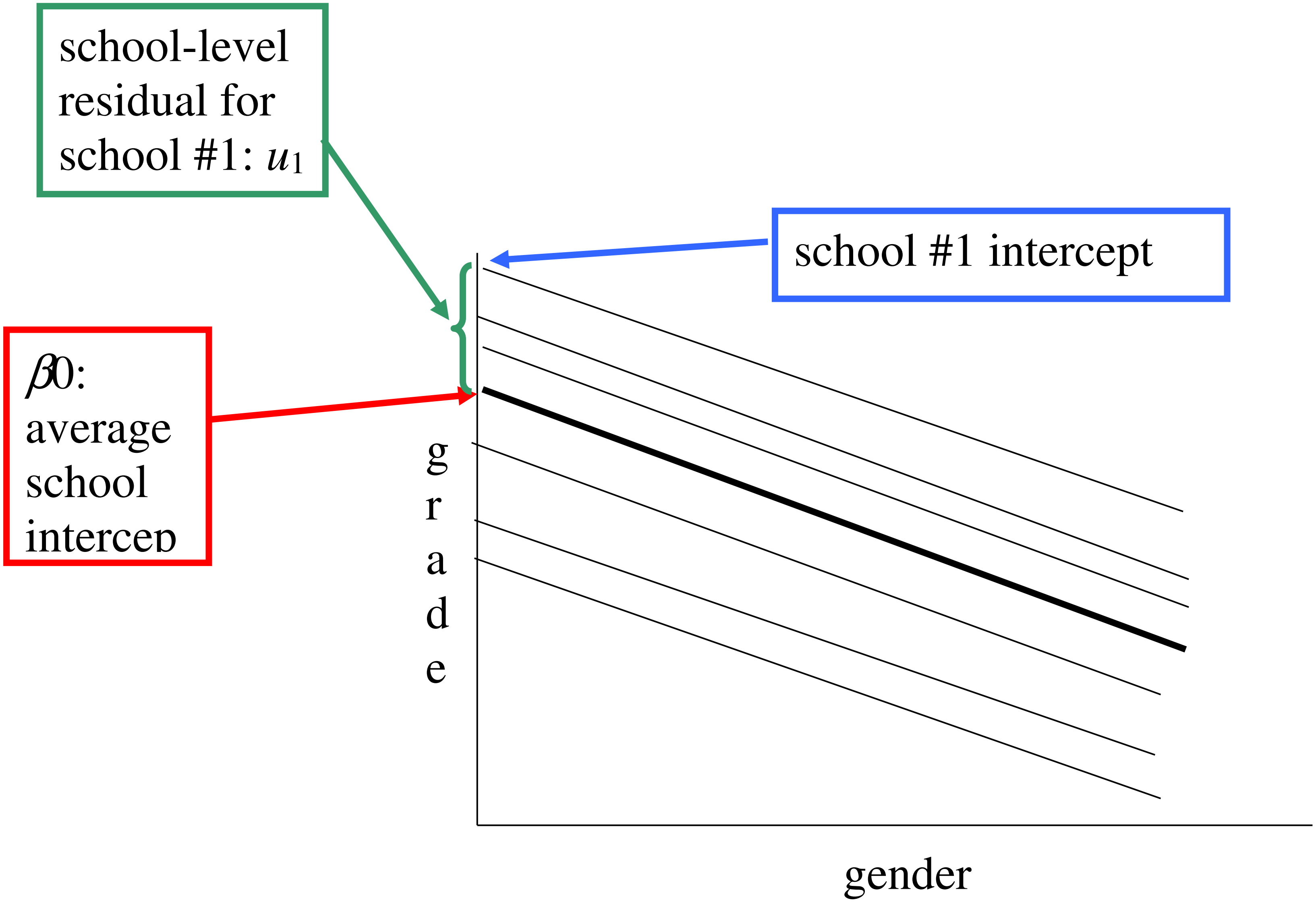
$$
\begin{aligned}
grade_{ij} &= \quad \beta 0_j \qquad + \beta 1 \times gender_{ij} + \varepsilon_{ij} \\
&= \left(\beta 0 + u_j \right) + \beta 1 \times gender_{ij} + \varepsilon_{ij}
\end{aligned}
$$

$\beta 0$ average of all school intercepts

u_j school-level residual

Usually, the u_j are not output, but their variance is estimated, $\hat{\sigma}_u^2$

this is known as the between-schools *variance component*



school-level residual for school #1: u_1
school #1 intercept
β0: average school intercep
g r a d e
gender

Benefits of Multilevel Models

- Does not assume that all observations are independent
- Correct standard errors
- Estimate and explain variation in random parameters
- Simultaneously model effects of different units of analysis

Example Data

- 1905 students within 73 schools
  From 2 to 104 students per school

- ID Variables
  School ID
  Student ID

- Outcome
  Student score on coursework, 'grade' (mean 79.03, range 10 - 108)

- Explanatory variables: Student-level
  Student score on essay (mean-centered)
  Student gender (0=girl, 1=boy)

- Explanatory variables: School-level
  Average essay score for each school (mean-centered)

Unconditional Variance Components Model

Research questions

- How much variation in coursework scores is attributable to schools?
- How much variation is attributable to students within schools?
- What is the intra-school correlation of coursework scores?

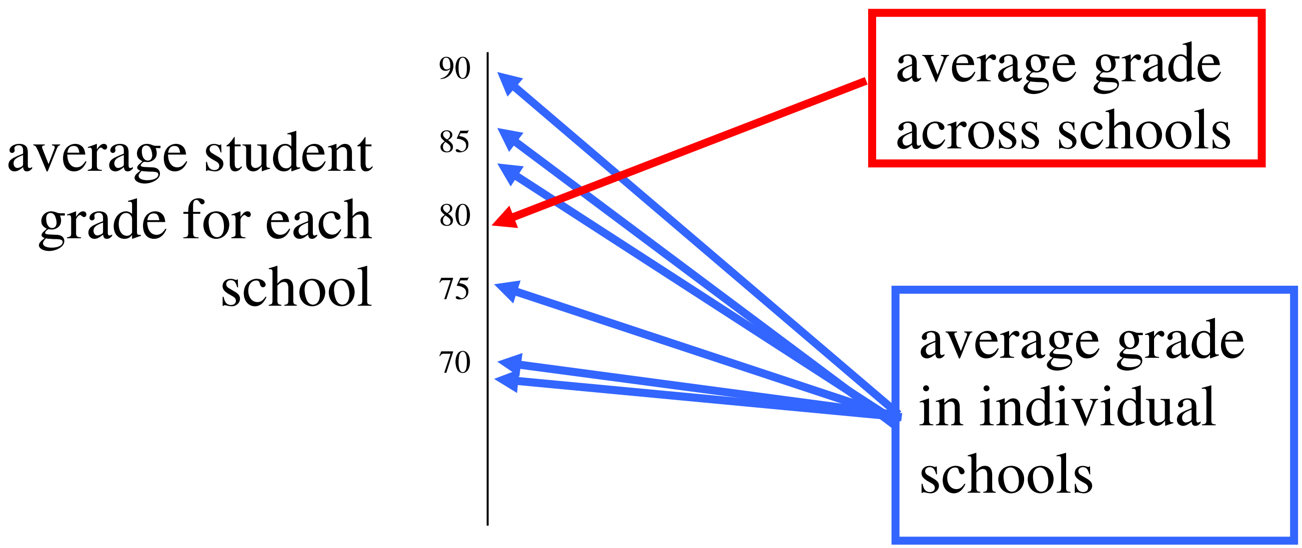
The Unconditional Variance Components Model

grade_{ij} = grand_mean + school residual + student residual

$$= \beta 0 \quad + u_j \quad + \varepsilon_{ij}$$

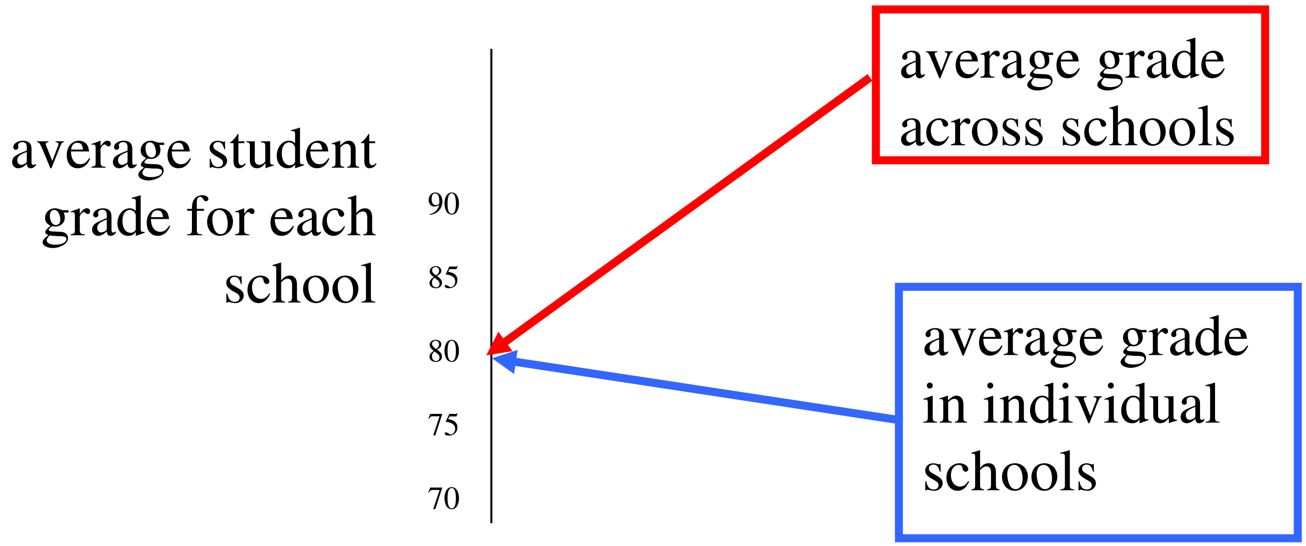
- Fit a model with no explanatory variables, only a random intercept
- Implicitly—not explicitly—an intercept is estimated for each school
- The school-level variance component, $\hat{\sigma}_u^2$, represents the variance of the average school grades around the grand mean (between school variation).
- The residual variance component, $\hat{\sigma}_\varepsilon^2$, represents the variance of student grades around their school mean (within-school variation)

- The school-level variance component, $\hat{\sigma}_u^2$, represents the variance of the average school grade around the grand mean (between school variation).



What if the school-level variance component equaled zero?

What if the school-level variance component equaled zero?



Then...

Knowing which school a student was from would not predict their grade

Students w/ a school would be no more alike than students across schools

That is, all students would be independent

No need for multilevel model, just use standard regression

What if the student-level variance component equaled zero?

That is, if the student-level residuals all equaled zero

Then...

All students within a school would have the same final grade

There would be no need to model student-level outcomes

PROC MIXED Syntax and Results

```
proc mixed;
  class school;
  model grade = / solution;
  random intercept / subject=school;
```

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Intercept	school	90.4250	17.7282	5.10	<.0001
Residual		226.64	7.4892	30.26	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	79.3581	1.2055	72	65.83	<.0001

-2 Res Log Likelihood 15893.7

Results Summary

Fixed Effect

- Grand mean for final grades = 79.36

Variance Components

- Between-school variation in average grades = 90.43
- Within-school variation in final grades = 226.64

Intra-school correlation

$$\rho = \text{between school variation} \div \text{total variation}$$
$$= 90.43 \div (90.43 + 226.64)$$
$$= 0.285$$

Adding a Fixed Student-Level Explanatory Variable: Gender

$$\text{grade}_{ij} = \beta 0 + \beta 1 \times \text{gender} + u_j + \varepsilon_{ij}$$

PROC MIXED Syntax and Results

```
proc mixed;
  class school;
  model grade = gender / solution;
  random intercept / subject=school;
```

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Intercept	school	91.5083	17.7969	5.14	<.0001
Residual		213.76	7.0656	30.25	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	82.4609	1.2427	72	66.36	<.0001
gender	-7.4189	0.7033	1831	-10.55	<.0001

-2 Res Log Likelihood 15784.5

Model Comparison

Fixed Effects	intercept only	+ gender
intercept	79.36	82.46
gender (student)	.	-7.42
Random Effects		
$\hat{\sigma}_u^2$ (school)	90.43	91.51
$\hat{\sigma}_\varepsilon^2$ (student)	226.64	213.76

all estimates, $p < .001$

Adding a Fixed School-Level Explanatory Variable: Average School Essay Score

$$\text{grade}_{ij} = \beta 0 + \beta 1 \times \text{gender} + \beta 2 \times \text{mean essay} + u_j + \varepsilon_{ij}$$

PROC MIXED Syntax and Results

```
proc mixed;
  class school;
  model grade = gender mean_essay/solution;
  random intercept / subject=school;
```

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Intercept	school	73.9983	14.8052	5.00	<.0001
Residual		213.68	7.0609	30.26	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	81.8434	1.1509	71	71.11	<.0001
gender	-7.5042	0.7030	1831	-10.67	<.0001
mean_essay	0.3529	0.08844	1831	3.99	<.0001

-2 Res Log Likelihood 15772.8

Model Comparison

Fixed Effects	intercept only	+ gender	+ mean_essay
intercept	79.36	82.461	81.84
gender (student)	.	-7.419	-7.50
essay (school)	.	.	0.35

Random Effects			
$\hat{\sigma}_u^2$ (school)	90.43	91.51	74.00
$\hat{\sigma}_\varepsilon^2$ (student)	226.64	213.76	213.68

all estimates, $p < .001$

Adding a Student-Level Fixed Explanatory Variable: Student Essay Score

$\text{grade}_{ij} = \beta 0 + \beta 1 \times \text{gender} + \beta 2 \times \text{mean essay} + \beta 3 \times \text{essay} + u_j + \varepsilon_{ij}$

PROC MIXED Syntax and Results

```
proc mixed;
 class school;
 model grade = gender mean_essay essay/solution;
 random intercept / subject=school;
```

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Intercept	school	77.1996	14.8609	5.19	<.0001
Residual		161.67	5.3445	30.25	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	82.4152	1.1433	72	72.08	<.0001
gender	-9.0170	0.6153	1829	-14.65	<.0001
mean_essay	-0.0446	0.08938	1829	-0.50	0.6173
paper	0.4049	0.01666	1829	24.30	<.0001

-2 Res Log Likelihood 15267.1

Model Comparison

Fixed Effects	Intercept only	gender only	gender + mean_essay	current model
intercept	79.36	82.46	81.84	82.42
gender (student)	.	-7.42	-7.50	-9.02
mean essay (school)	.	.	0.35	-0.04
essay (student)	.	.	.	0.41

Random Effects				
τ_{00}	90.43	91.51	74.00	77.20
σ^2	226.64	213.76	213.68	161.67

all estimates, $p < .001$, except mean essay, n.s.