Lecture 4

Lecture 4 Outline

- Focus on separating individual-level and cluster-level covariate effects in multi-level models
  - This continues our discussion from Lecture 3 where we utilized graphical displays to separate the within and between cluster effects for a level-1 covariate
  - Focus will be on estimation and interpretation of fixed effects within a random intercept model (for now)
  - Begg and Parides 2003 Statistics in Medicine paper

- We will also consider the case where the goal is to estimate the effect of a Level-2 covariate while adjusting for level-1 covariates.

- Example: Inner London School Data
  - Students (level 1) nested within schools (level 2)
  - OUTCOME: GCSE score, age 16 exam score G_{ij}
  - Level-1 Covariate: LRT score, age 11 reading test score L_{ij}

Scientific Questions

For now, focus on fixed effects; i.e. interpretation of regression coefficients within mixed model (not interpretation of variance of random effects).

1. Quantify the relationship between GCSE score and LRT score
2. Within a school, quantify the relationship between GCSE score and LRT score
3. Does the "context" of the school matter? i.e. do students from schools with higher school-average LRT scores fair better than otherwise similar students in schools with lower school-average LRT scores
   – Defined as the "contextual" effect (see "Brief conceptual…"

Exploratory Data Analysis

(continued)

```
* Create some new variables

sort school student

* Generate the number of students within each school
by school: egen totalstudents = count(student)

* Generate a counter for the number of students within each school
by school: gen withinschoolcount = _n


* EDA

* What is the distribution of number of students in each school
summ totalstudents if withinschoolcount==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
totalstude~s |         65    62.44615    29.74844          2        198
```

65 schools in the dataset:

Number of students ranges from 2 to 198, average 62

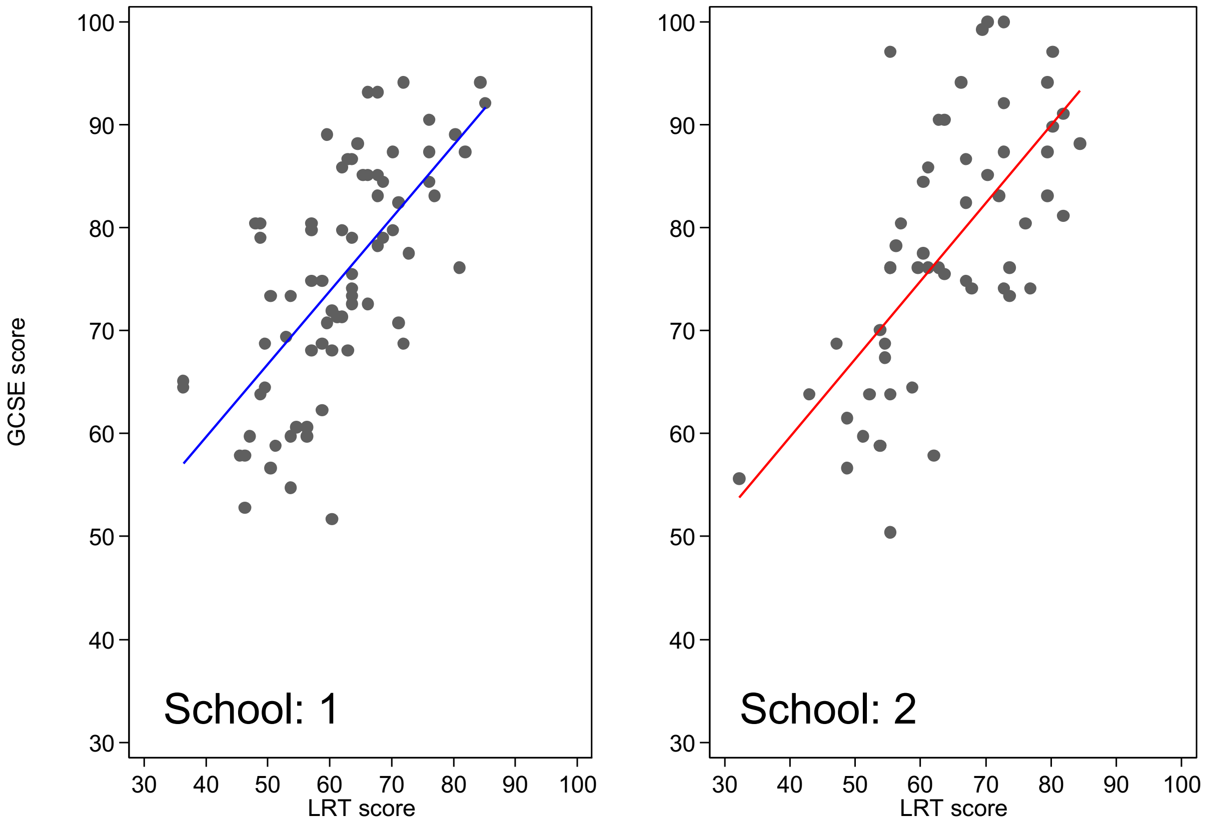
Exploratory Data Analysis

```
* What is the distribution of the gcse and lrt scores

. summ gcse

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        gcse |      4059    69.99527    9.977929     33.339        100

. summ lrt

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         lrt |      4059     60.0181     9.93223      30.65      90.16
```

NOTE: This data was extracted from your textbook. The data is provided as z-scores with standard deviation of 10 instead of 1. I added 60 to the LRT scores and 70 to the GCSE scores. So we will interpret the data in the lecture as the raw test score.

Exploratory Data Analysis

- Relationship between gcse and lrt among two schools (1 and 2)



Data by school: $(L_{ij}, G_{ij}, j = 1, \dots, n_i)$

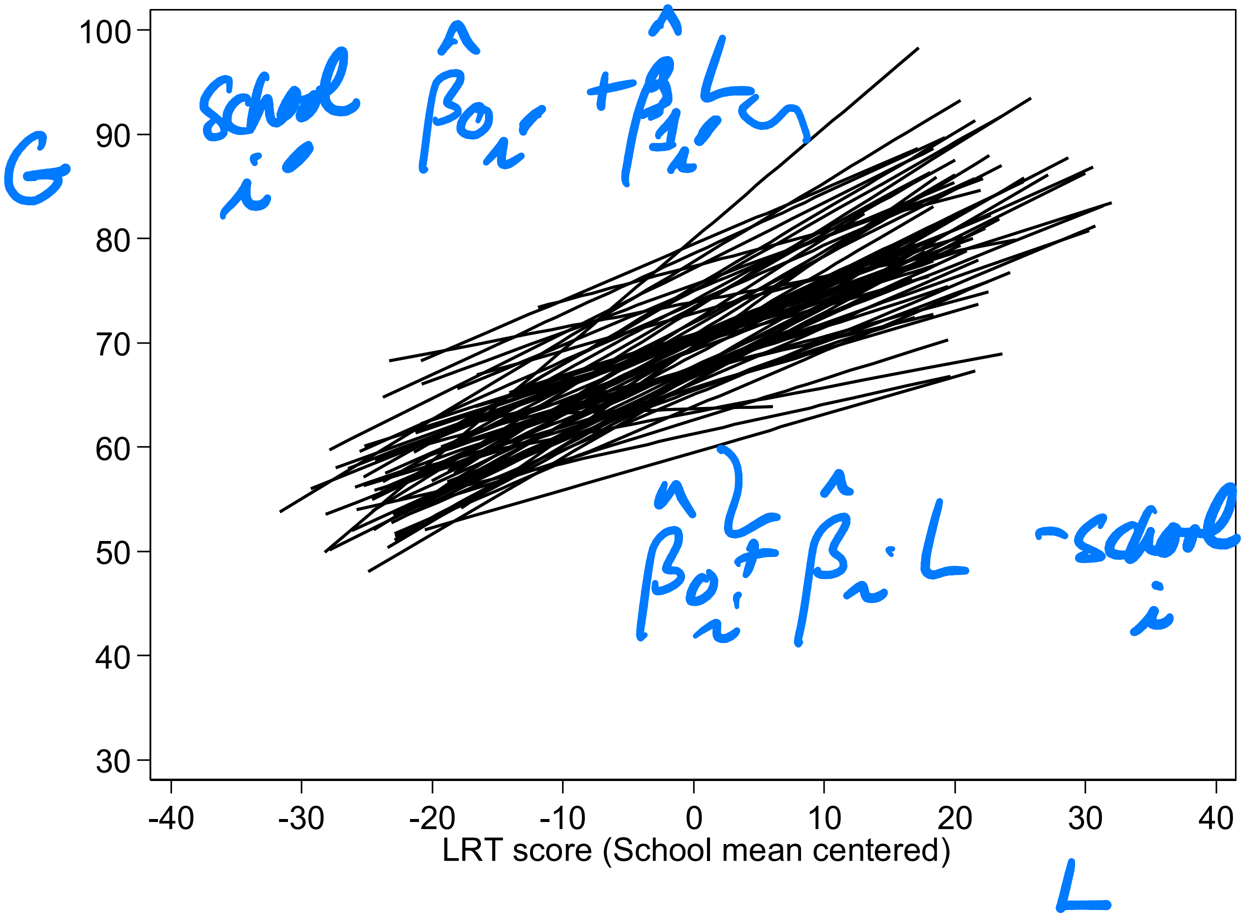
Regress G_{ij} on L_{ij} for each i

$$G_{ij} = \beta_{0i} + \beta_{1i} L_{ij} + \varepsilon_{ij}, \quad i = 1, \dots, m.$$

$$\Longrightarrow (\hat{\beta}_{0i}, \hat{\beta}_{1i}) \quad \left(\text{and } Var\begin{pmatrix}\hat{\beta}_{0i} \\ \hat{\beta}_{1i}\end{pmatrix} = V_i\right)$$

$$i = 1, \dots, m.$$

School-specific relationships among schools with at least 5 students



- One of our objectives is to estimate the within cluster association.

- There appears to be variation in this association across schools ?

- Lecture 6

- For now, we want to estimate the average of these slopes!

Possible Models to Consider

G L

$$Model\ 1:\ E\left(Y_{ij} \middle| X_{ij}\right) \qquad = \alpha_{1i} + \beta_1 X_{ij}$$
$$Model\ 2:\ E\left(Y_{ij} \middle| X_{ij}, \bar{X}_{i.}\right) = \alpha_{2i} + \beta_2 X_{ij} + \gamma_2 \bar{X}_{i.}$$
$$Model\ 3:\ E\left(Y_{ij} \middle| X_{ij}, \bar{X}_{i.}\right) = \alpha_{3i} + \beta_3\left(X_{ij} - \bar{X}_{i.}\right) + \gamma_3 \bar{X}_{i.}$$
$$Model\ 4:\ E\left(Y_{ij} \middle| X_{ij}, \bar{X}_{i.}\right) = \alpha_{4i} + \beta_4\left(X_{ij} - \bar{X}_{i.}\right)$$
$$Model\ 5:\ E\left(Y_{ij} \middle| X_{ij}, \bar{X}_{i.}\right) = \alpha_{5i} + \gamma_5 \bar{X}_{i.}$$

where

$$Y_{ij} = E\left(Y_{ij} \middle| ..\right) + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma_k{}^2)$$
$\alpha_{ki} \sim N(\underline{\alpha_k}, \underline{\tau_k}^2)$, for each model k $= 1, \ldots, 5$

Notes on the possible models

$Model\ 1\text{:}\ \ E(Y_{ij}|X_{ij}) = \alpha_{1i} + \beta_1 X_{ij}$

- Ignores the clustering of the data when it is estimating β_1, this is the "total effect"

  - does allow for heterogeneity in intercepts.

$Model\ 2\text{:}\ \ E(Y_{ij}|X_{ij}, \bar{X}_{i.}) = \alpha_{2i} + \beta_2 X_{ij} + \gamma_2 \bar{X}_{i.}$

$Model\ 3\text{:}\ \ E(Y_{ij}|X_{ij}, \bar{X}_{i.}) = \alpha_{3i} + \beta_3 (X_{ij} - \bar{X}_{i.}) + \gamma_3 \bar{X}_{i.}$

- These two models are ~~mathematically~~ assume equivalent values for $E(Y_{ij}|$
- In Model 3, we have chosen to "center" the level-1 covariate
  - "cluster-mean" or "cluster-specific" or "within-cluster" centering
- The choice ~~of~~ to center~~ing~~ or not will change the interpretation of the intercept and also the interpretation of γ_2 and γ_3
- We will also discuss the option to "grand-mean center" later in the lecture and the impact this can have.

Notes on the possible models

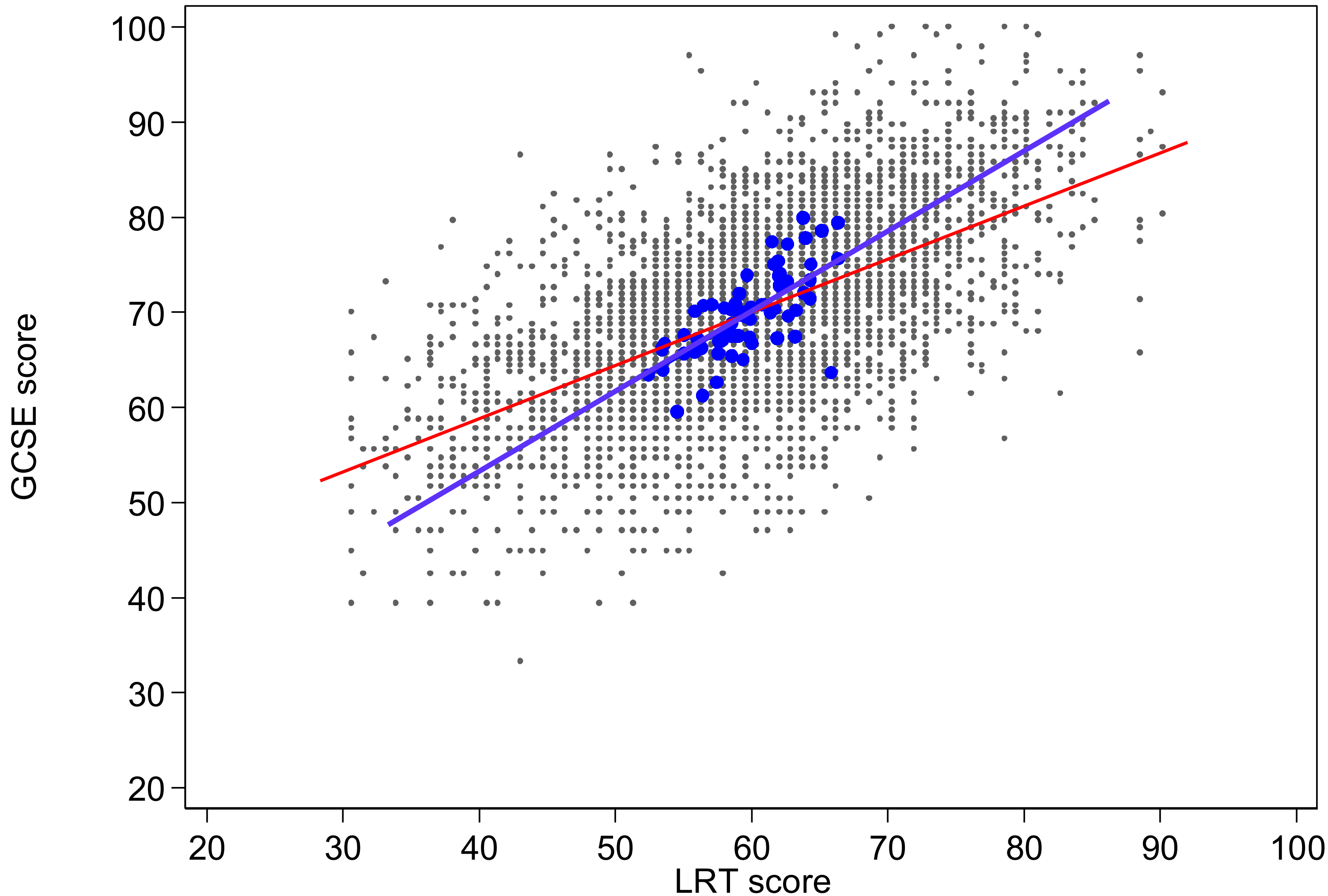
$Model\ 4$: $E\left(Y_{ij}\middle|X_{ij},\bar{X}_{i.}\right) = \alpha_{4i} + \beta_4\left(X_{ij} - \bar{X}_{i.}\right)$ $+ 0 \cdot \bar{X}_{i.}$

- This model ignores the cluster mean covariate
- Includes only the cluster-mean centered level-1 covariate

$Model\ 5$: $E\left(Y_{ij}\middle|X_{ij},\bar{X}_{i.}\right) = \alpha_{5i} + \gamma_5\bar{X}_{i.}$ $+ 0 \cdot (X_{ij} - \bar{X}_{i.})$

- This model ignores the cluster-mean centered level-1 covariate
- Includes only the cluster-mean covariate

Recall the EDA: Separation of Between and Within Effects



Fit the models using (xt)mixed in Stata

```
* Generate the cluster-mean variable and cluster-mean centered
* lrt score
bys school: egen mean_lrt = mean(lrt)
gen lrt_within  = lrt – mean_lrt

***** Model 1
mixed gcse lrt || school:
***** Model 2
mixed gcse lrt mean_lrt || school:
***** Model 3
mixed gcse lrt_within mean_lrt || school:
***** Model 4
mixed gcse lrt_within || school:
***** Model 5
mixed gcse mean_lrt || school:
```

Results

Model	β_k	γ_k
1	0.563 (0.538, 0.587)	
2	0.559 (0.534, 0.583)	0.357 (0.142, 0.573)
3	0.559 (0.534, 0.583)	0.916 (0.702, 1.131)
4	0.559 (0.534, 0.583)	
5		0.925 (0.712, 1.138)

We note first the difference in estimates in the $\boldsymbol{\beta_k}$ column.

- The estimate from Model 1 (0.563) is the total effect
- This estimate basically ignores the cluster membership
- This estimate is a distorted estimate of the within cluster effect due to confounding by the cluster-average LRT
  - Can show that cluster-average LRT is correlated with both individual LRT and the GCSE score.

Interpretation of Models 2 and 3

$$Model\ 2:\ \ E\left(Y_{ij} \middle| X_{ij}, \bar{X}_{i.}\right) = \ \alpha_{2i} + \beta_2 X_{ij} + \gamma_2 \bar{X}_{i.}$$

$$Model\ 3:\ \ E\left(Y_{ij} \middle| X_{ij}, \bar{X}_{i.}\right) = \alpha_{3i} + \beta_3\left(X_{ij} - \bar{X}_{i.}\right) + \gamma_3 \bar{X}_{i.}$$

Model	β_k	γ_k
2	0.559 (0.534, 0.583)	0.357 (0.142, 0.573)
3	0.559 (0.534, 0.583)	0.916 (0.702, 1.131)

β_2 and β_3 have the same estimated values and interpretation.

The effect of LRT score on GCSE within a given cluster:

G

Within a school, the student's average GCSE scores differ by 0.559 points per additional point on the LRT.

L

Interpretation of Models 2 and 3

$$Model\ 2:\ \ E\left(Y_{ij}\middle|X_{ij},\bar{X}_{i.}\right) = \ \alpha_{2i} + \beta_2 X_{ij} + \gamma_2 \bar{X}_{i.}$$

$$Model\ 3:\ \ E\left(Y_{ij}\middle|X_{ij},\bar{X}_{i.}\right) = \alpha_{3i} + \beta_3\left(X_{ij} - \bar{X}_{i.}\right) + \gamma_3 \bar{X}_{i.}$$

Model	$\boldsymbol{\beta_k}$	$\boldsymbol{\gamma_k}$
2	0.559 (0.534, 0.583)	0.357 (0.142, 0.573)
3	0.559 (0.534, 0.583)	0.916 (0.702, 1.131)

γ_2 and γ_3 have different values, what do they represent?

Model 2: γ_2 represents the "contextual effect"!

Model 3: γ_3 represents the "between effect!
(among) 'school"

Interpretation of Models 2 and 3

$Model\ 2$: $E(Y_{ij}|X_{ij}, \bar{X}_{i.}) = \alpha_{2i} + \beta_2 X_{ij} + \gamma_2 \bar{X}_{i.}$

Model	β_k	γ_k
2	0.559 (0.534, 0.583)	0.357 (0.142, 0.573)

Model 2: Holding X_{ij} fixed, the mean difference in Y_{ij} per unit increase in $\bar{X}_{i.}$

Consider two students with the same LRT score but who come from schools that differ in school average LRT score by 1 point.

The student from the school with higher average LRT score is expected to have a GCSE score that is 0.357 points higher than the other student.

Interpretation of Models 2 and 3

$$Model\ 3\text{:}\ \ E\left(Y_{ij}\middle|X_{ij},\bar{X}_{i.}\right) = \alpha_{3i}+\beta_3\left(X_{ij}-\bar{X}_{i.}\right)+\gamma_3\bar{X}_{i.}$$

Model	$\boldsymbol{\beta_k}$	$\boldsymbol{\gamma_k}$
3	0.559 (0.534, 0.583)	0.916 (0.702, 1.131)

Consider the full model:

$$Y_{ij} = \alpha_{3i}+\beta_3\left(X_{ij}-\bar{X}_{i.}\right)+\gamma_3\bar{X}_{i.}+\varepsilon_{3ij}$$

Take the cluster mean:

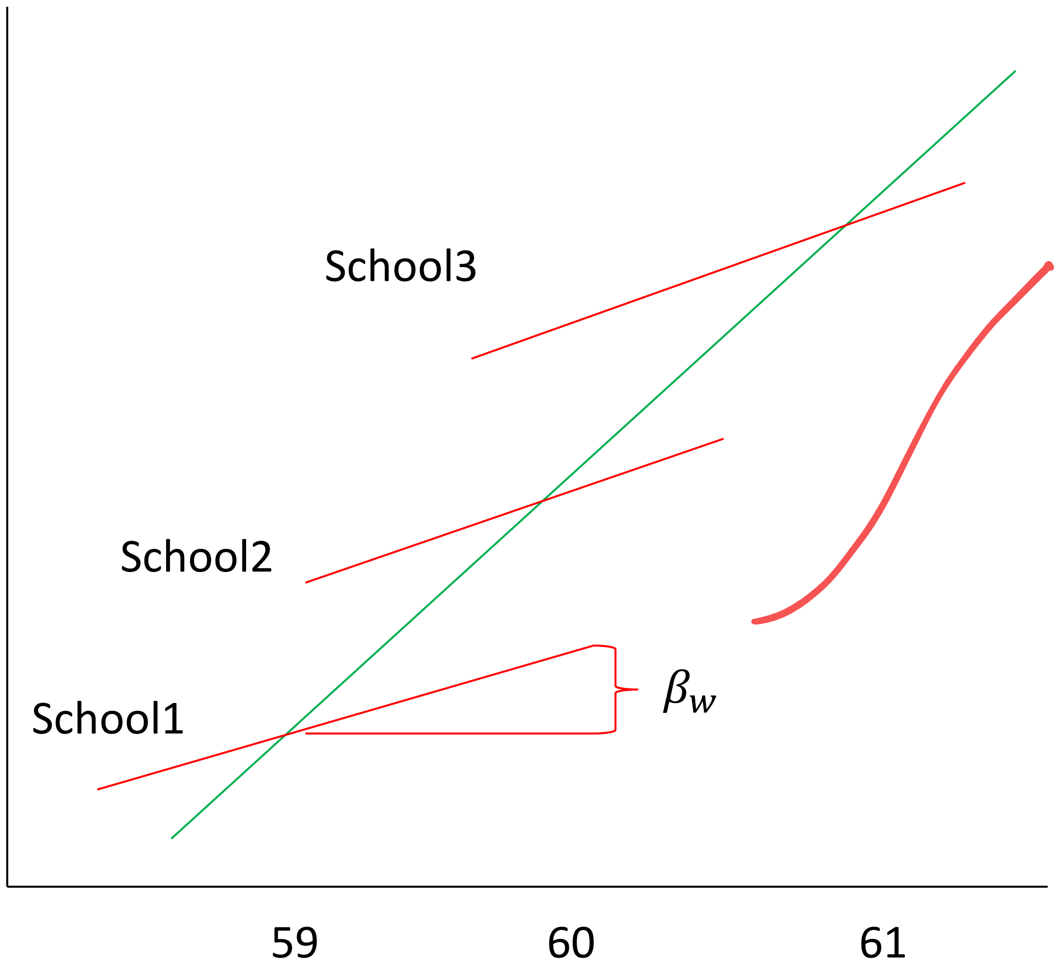
$$\bar{Y}_{i.} = \alpha_{3i}+\beta_3(\bar{X}_{i.}-\bar{X}_{i.})+\gamma_3\bar{X}_{i.}+\bar{\varepsilon}_{3i.}$$
$$\bar{Y}_{i.} = \alpha_{3i}+\gamma_3\bar{X}_{i.}+\bar{\varepsilon}_{3i.}$$

This is the between effect: The difference in school average GCSE per unit increase in school average LRT score.

Contextual Effects

- Why do these effects occur?
  - Normative effects associated with the cluster/organization/level-2 factor
    - i.e. persons within the cluster tend to be much more like each other than otherwise similar persons from other clusters
  - The mean X within a cluster may act as a proxy for other important cluster level characteristics that are not measured
  - They may signal a statistical artifact if the mean X within a cluster may carry some information if X is measured with error
- Posted a few references for with todays lecture that explore these ideas.

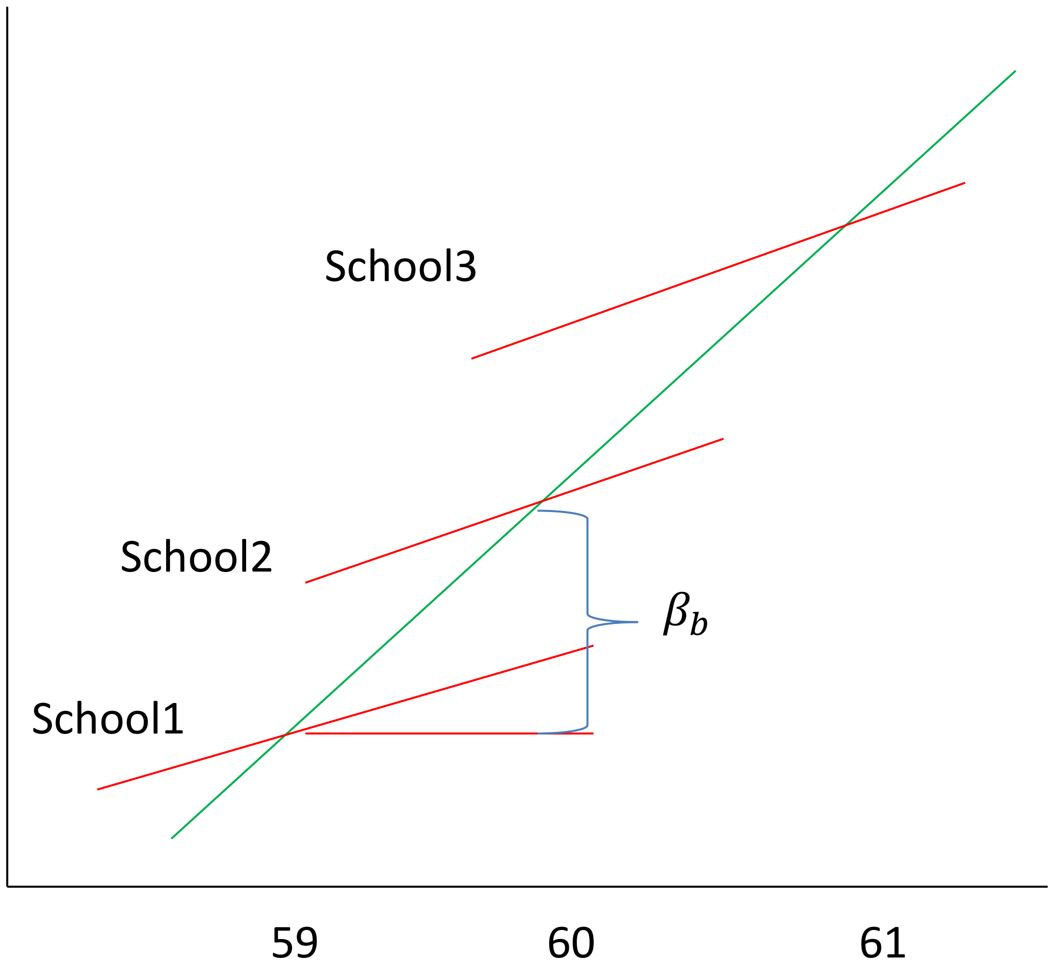
Contextual Effects



Within-effect: Expected difference in Y between two **subjects** from the same **cluster** but who differ in X by one unit

Inner London School EX: expected difference in GCSE score between two students from the same school but who differ in LRT by one point

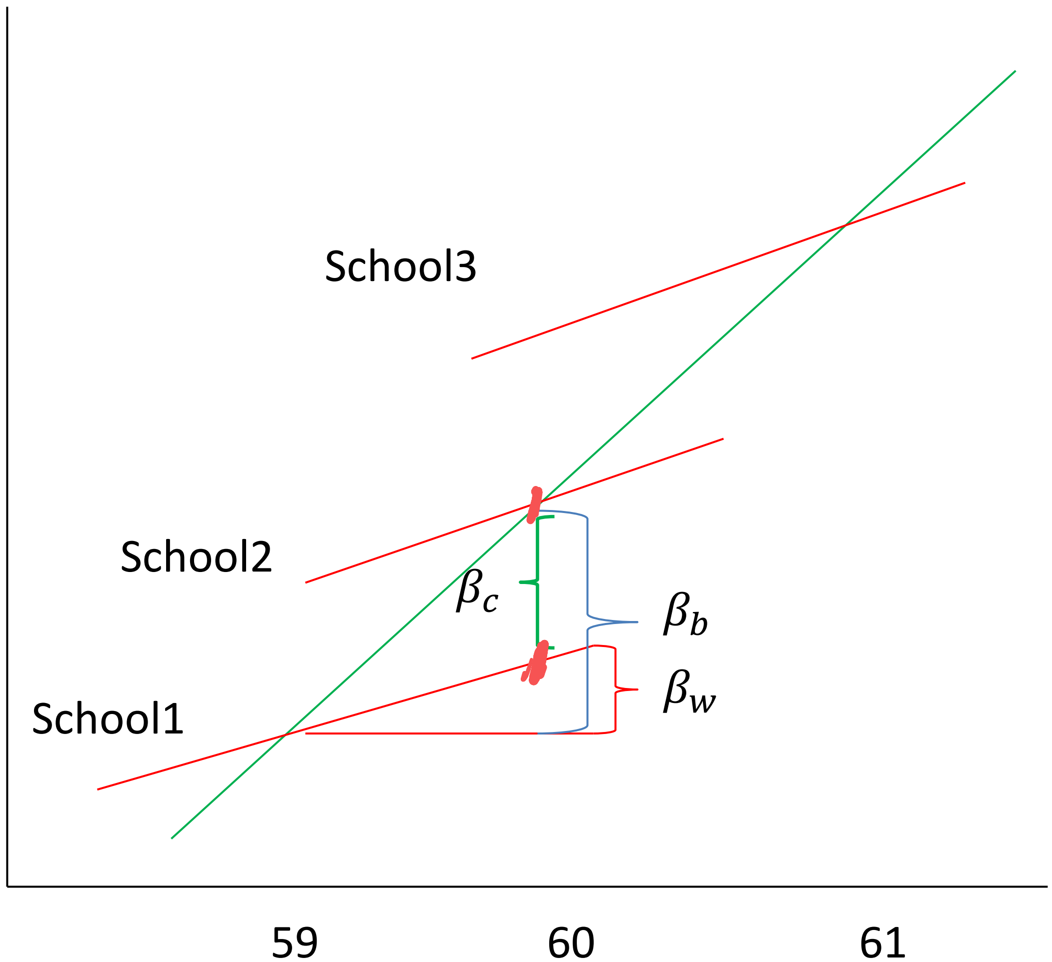
Contextual Effects



Between-effect: Expected difference in mean Y between two **clusters** that differ in average X by one unit

Inner London School EX: expected difference in mean GCSE score between two schools that differ in average LRT by one point

Contextual Effects



Contextual-effect: Expected difference in Y between two **subjects** who have the same value of X but who come from **clusters** that differ by one unit in mean X

Inner London School EX: expected difference in GCSE scores between two students with the same LRT score but who come from two schools that differ in average LRT by one point

Interpretation of Models 2 and 3

$$Model\ 2\text{:}\ \ E\left(Y_{ij} \middle| X_{ij}, \bar{X}_{i.}\right) = \ \alpha_{2i} + \beta_2 X_{ij} + \gamma_2 \bar{X}_{i.}$$

$$Model\ 3\text{:}\ \ E\left(Y_{ij} \middle| X_{ij}, \bar{X}_{i.}\right) = \alpha_{3i} + \beta_3 \left(X_{ij} - \bar{X}_{i.}\right) + \gamma_3 \bar{X}_{i.}$$

Model	$\boldsymbol{\beta_k}$	$\boldsymbol{\gamma_k}$
2	0.559 (0.534, 0.583)	0.357 (0.142, 0.573)
3	0.559 (0.534, 0.583)	0.916 (0.702, 1.131)

γ_2 is the contextual effect; estimated directly within Model 2

β_3 is the within cluster effect and γ_3 is the between cluster effect.

The contextual effect can be estimated within Model 3 by taking:

$$\gamma_2 = \gamma_3 - \beta_3$$

Interpretation of Models 4 and 5

$$Model\ 3\text{:}\ \ E\left(Y_{ij}\middle|X_{ij},\bar{X}_{i.}\right) = \alpha_{3i} + \beta_3\left(X_{ij} - \bar{X}_{i.}\right) + \gamma_3\bar{X}_{i.}$$
$$Model\ 4\text{:}\ \ E\left(Y_{ij}\middle|X_{ij},\bar{X}_{i.}\right) = \alpha_{4i} + \beta_4\left(X_{ij} - \bar{X}_{i.}\right)$$
$$Model\ 5\text{:}\ \ E\left(Y_{ij}\middle|X_{ij},\bar{X}_{i.}\right) = \alpha_{5i} + \gamma_5\bar{X}_{i.}$$

Model	$\boldsymbol{\beta_k}$	$\boldsymbol{\gamma_k}$
3	0.559 (0.534, 0.583)	0.916 (0.702, 1.131)
4	0.559 (0.534, 0.583)	
5		0.925 (0.712, 1.138)

Models 3 and 4, $\boldsymbol{\beta_k}$ are both estimating the within effect.

Models 3 and 5, $\boldsymbol{\gamma_k}$ are both estimating the between effect.

What does this imply about $\left(X_{ij} - \bar{X}_{i.}\right)$ and $\bar{X}_{i.}$?

Interpretation of Models 4 and 5

What does this imply about $(X_{ij} - \bar{X}_{i.})$ and $\bar{X}_{i.}$?

Level-2 covariates and cluster-mean centered level-1 covariates are independent!

```
 corr gcse lrt lrt_within mean_lrt girl schgend
(obs=4059)

             |        1         2         3         4         5         6
-------------+------------------------------------------------------------
1       gcse |   1.0000
2        lrt |   0.5917   1.0000
3 lrt_within |   0.5277   0.9484   1.0000
4   mean_lrt |   0.2879   0.3170  -0.0000   1.0000
5       girl |   0.1144   0.0532   0.0425   0.0407   1.0000
6    schgend |   0.1115   0.0067  -0.0000   0.0210   0.4365   1.0000
```

Estimating level-2 effects while adjusting* for level-1 covariates

* "Compare like to like"

- One of the most common applications of MLM!
- Here it is commonly assumed that no contextual effect exists.
- In this setting, cluster-mean centering is not appropriate
- If you want to center the level-1 covariates, you should grand-mean center

Example: School Type

- Estimate the difference in school-average GCSE score across school-type adjusting for the composition of the schools
- Composition measured by gender and LRT score

Unadjusted model: $Y_{ij} = \beta_{0i} + \varepsilon_{ij}$

$$\beta_{0i} = \beta_0 + \beta_1 I(boys_i) + \beta_2 I(girls_i) + b_i$$

Adjusted models add main effects to the level-1 equation

Adjusted 1: $Y_{ij} = \beta_{0i} + \alpha_{11} LRT_{ij} + \alpha_{21} girl_{ij} + \varepsilon_{ij}$

Adjusted 2: $Y_{ij} = \beta_{0i} + \alpha_{12}\left(LRT_{ij} - \overline{LRT}_{..}\right) + \alpha_{22}\left(girl_{ij} - \overline{girl}_{..}\right) + \varepsilon_{ij}$

Adjusted 3: $Y_{ij} = \beta_{0i} + \alpha_{13}\left(LRT_{ij} - \overline{LRT}_{i.}\right) + \alpha_{23}\left(girl_{ij} - \overline{girl}_{i.}\right) + \varepsilon_{ij}$

Adjusted Model	All Boys (β_1)	All Girls (β_2)
Unadjusted	0.644 (-2.282, 3.569)	2.562 (0.275, 4.849)
1, main effects	1.176 (-0.392, 3.945)	1.575 (-0.133, 3.284)
2, grand mean centered	1.176 (-0.392, 3.945)	1.575 (-0.133, 3.284)
3, cluster mean centered	0.617 (-2.311, 3.546)	2.550 (0.255, 4.845)

- Unadjusted model:
  - The school average GCSE score is 0.644 points higher among all boys schools compared to mixed gender schools.
- Adjusted model:
  - After accounting for the composition of the school, the school average GCSE score is 1.176 points higher among all boys schools compared to mixed gender schools.

NOTE: The adjustment for the level-1 covariates occurs only in the models with no centering or with grand-mean centering.

Lecture 4 Summary

- In linear models, we considered 5 different models that allowed us to estimate
  - The total effect
  - Within-cluster effect
  - Among-cluster effect
  - Contextual effect
- We reviewed the interpretation of these effects within an example; Inner London School data
- We noticed that cluster-mean centered level-1 covariates are independent of level-2 covariates
  - Explored the impact of this observation for studies where our goal is to estimate the association between a level-1 outcome and level-2 covariate but adjusting for the composition of the clusters.

Lecture 5 Introduction

- Our example at the end of the lecture brings up an important topic for MLMs
  - CENTERING!
- In the next lecture, we will consider:
  - the choice of centering and how this choice can affect your solutions in MLMs
  - how the choice of centering can affect your estimates of random intercept and slope variances!