Lecture 3

Multi-level Models – Review of Big Ideas

1. An individual's health is caused by myriad factors that operate on different spatial, temporal, and interaction scales
   - Genes, cells, organs, persons, families, neighborhoods, states, countries
   - 100 Hz, seconds, hours, days, weeks, months, years, generations
   - Intra/inter-specific, nested/crossed, Markov,…

Multi-level Models – Review of Big Ideas

2. An individual's health state, trajectory, and intervention effects are best predicted by treating her as a representative of a population of people **who share the processes** that give rise to health and disease

3. Multi-level models are a key methodology for learning about individuals from relevant population data and vice versa

Multi-level model: individual ⇔ population

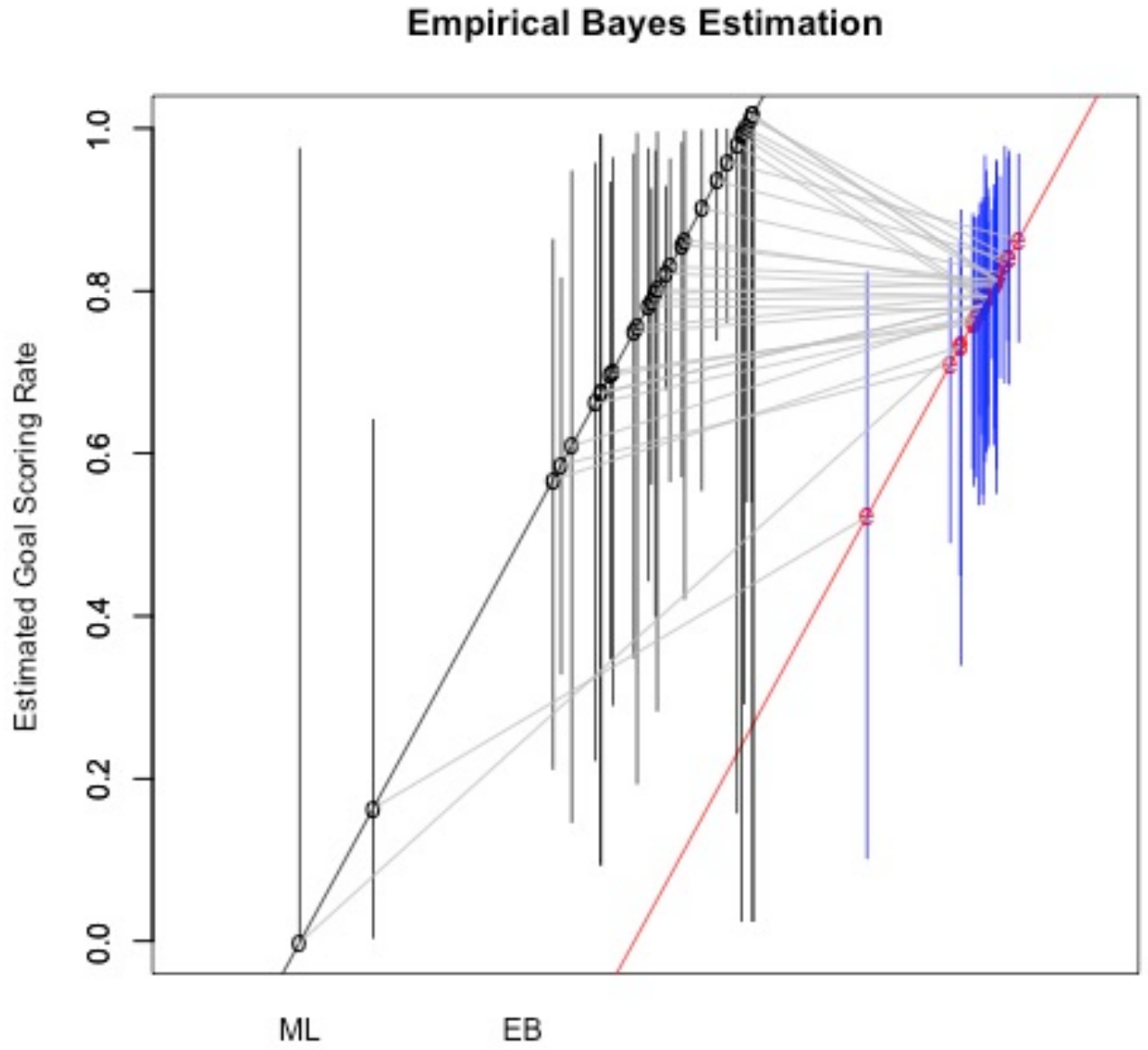
Predicting an Individual's Penalty Shot Scoring by Borrowing Strength from the Experience of Others

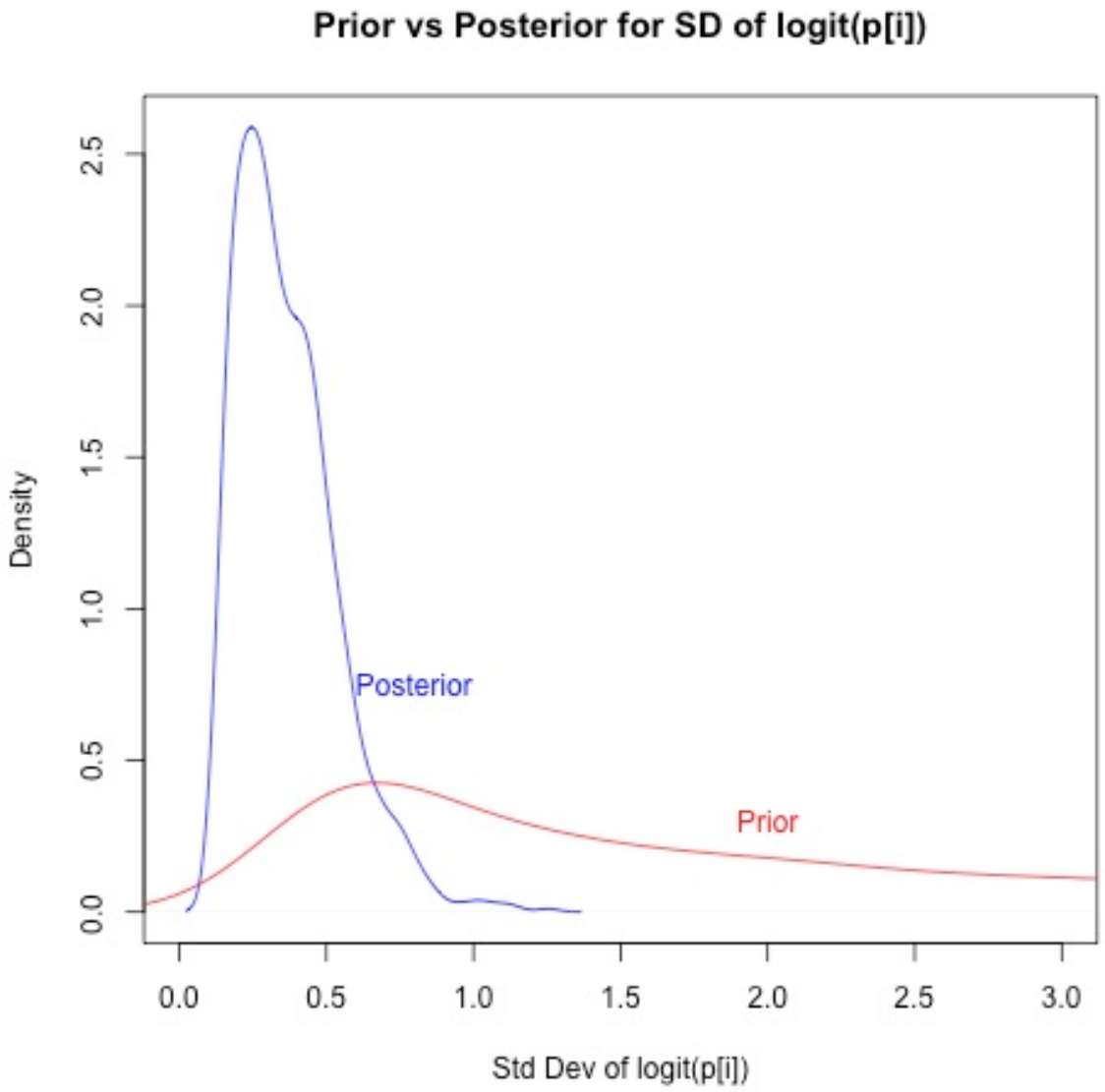

KUMHO
10
21
24

	shots	goals	phat	se	alpha	phat.eb
Valdivia	1	0	0.00	0.499	0.045	0.51
Forlan	6	2	0.33	0.204	0.222	0.49
Hernandez	9	4	0.44	0.166	0.300	0.50
Cavani	17	11	0.65	0.121	0.447	0.58
Lukaku	5	3	0.60	0.223	0.192	0.54
Samara	3	0	0.00	0.288	0.125	0.46
Benzema	3	2	0.67	0.288	0.125	0.55
Ozil	6	1	0.17	0.204	0.222	0.45
Giroud	10	5	0.50	0.158	0.323	0.52
Dempsey	7	4	0.57	0.189	0.250	0.54
Odemwinge	4	2	0.50	0.250	0.160	0.53
Johannsson	8	2	0.25	0.176	0.276	0.45
Martinez	9	6	0.67	0.166	0.300	0.57
Hulk	23	12	0.52	0.104	0.523	0.53
Robben	10	7	0.70	0.158	0.323	0.59
Ruaz	5	3	0.60	0.223	0.192	0.54
van Persie	17	5	0.29	0.121	0.447	0.42
Messi	41	22	0.54	0.078	0.661	0.53
Rodriguez	7	4	0.57	0.189	0.250	0.54
Muller	14	8	0.57	0.133	0.400	0.55
Aguero	10	5	0.50	0.158	0.323	0.52
Hazard	19	13	0.68	0.114	0.475	0.60
Vidal	21	11	0.52	0.109	0.500	0.53
Neymar	6	5	0.83	0.204	0.222	0.60
Guardado	6	3	0.50	0.204	0.222	0.52
Borges	1	1	1.00	0.499	0.045	0.55
Salpingidis	1	0	0.00	0.499	0.045	0.51
Moses	3	3	1.00	0.288	0.125	0.59
Drmie	2	2	1.00	0.353	0.087	0.57
Shaquiri	1	0	0.00	0.499	0.045	0.51

Multi-level Model – hint of Bayes

- Level I. $[y_i \mid t_i, n_i]$: likelihood of the data given unknowns
  - Given true rate t_i for player $i = 1,\ldots,m,$ the observed number of scored penalty shots, y_i out of n_i attempts is *binomial*(t_i, n_i)
- Level II. $[t_i \mid n_i, m, v]$: prior distribution of unknowns
  - The true rates t_i are independent and identically distributed random variables with distribution *Beta*(m, v)
- Multi-level model calculates $[t_i, m, v \mid n_i, y_i]$: posterior of the unknowns given the data – using Bayes theorem


Empirical Bayes Estimation
Estimated Goal Scoring Rate
0.0
0.2
0.4
0.6
0.8
1.0
ML
EB


Prior vs Posterior for SD of logit(p[i])
Density
0.0
0.5
1.0
1.5
2.0
2.5
Posterior
Prior
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Std Dev of logit(p[i])

Gaussian-Gaussian Model from Prior Lecture

Lecture 3

The goals of this lecture are to

1. Describe the various types of variables that you may work with in multi-level data with 2 levels
2. Describe various scientific questions of interest in two level models
3. Present EDA for common goals within linear models

We will do all of this within the context of a data example; Inner-London School Data

Inner-London School Data

- At age 16, students take Graduate Certificate of Secondary Education (GCSE) exams
  - Scores derived from the GCSE are used for school comparisons
  - However, schools should be compared based upon their “value added”; the difference in GCSE score between schools after controlling for achievements before entering the schoo
- One measure of prior achievement is the London Reading Test (LRT)
  - taken by these students at age 11
- Dataset represents 2-level data
  - Students (level 1) nested within Schools (level 2)

Variable Types in 2-level data

- Outcome measured at level-1: GCSE score for each student
  - This defines the multi-level data problem
  - NOTE: If the outcome is measured at level 2, you really only have one measure per cluster, don't need special methods
- Level-1 exposures
  - Characteristics of the level-1 units
  - e.g. LRT score for each student , student gender
- Level-2 exposures
  - Characteristics of the level-2 units
  - EG. Type of school: mixed gender, all boys, all girls

Variable Types in 2-level data

- Level-2 calculated variables
  - These are variables that represent summary information for level-1 exposures
  - School mean LST score
  - School variance of LST score
  - Proportion of female students

Possible Questions within 2-level Data

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
   - Within a school, quantify the relationship between GCSE score and LRT score
   - Across schools, quantify the relationship between school-specific mean GCSE and mean LRT

2. 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.

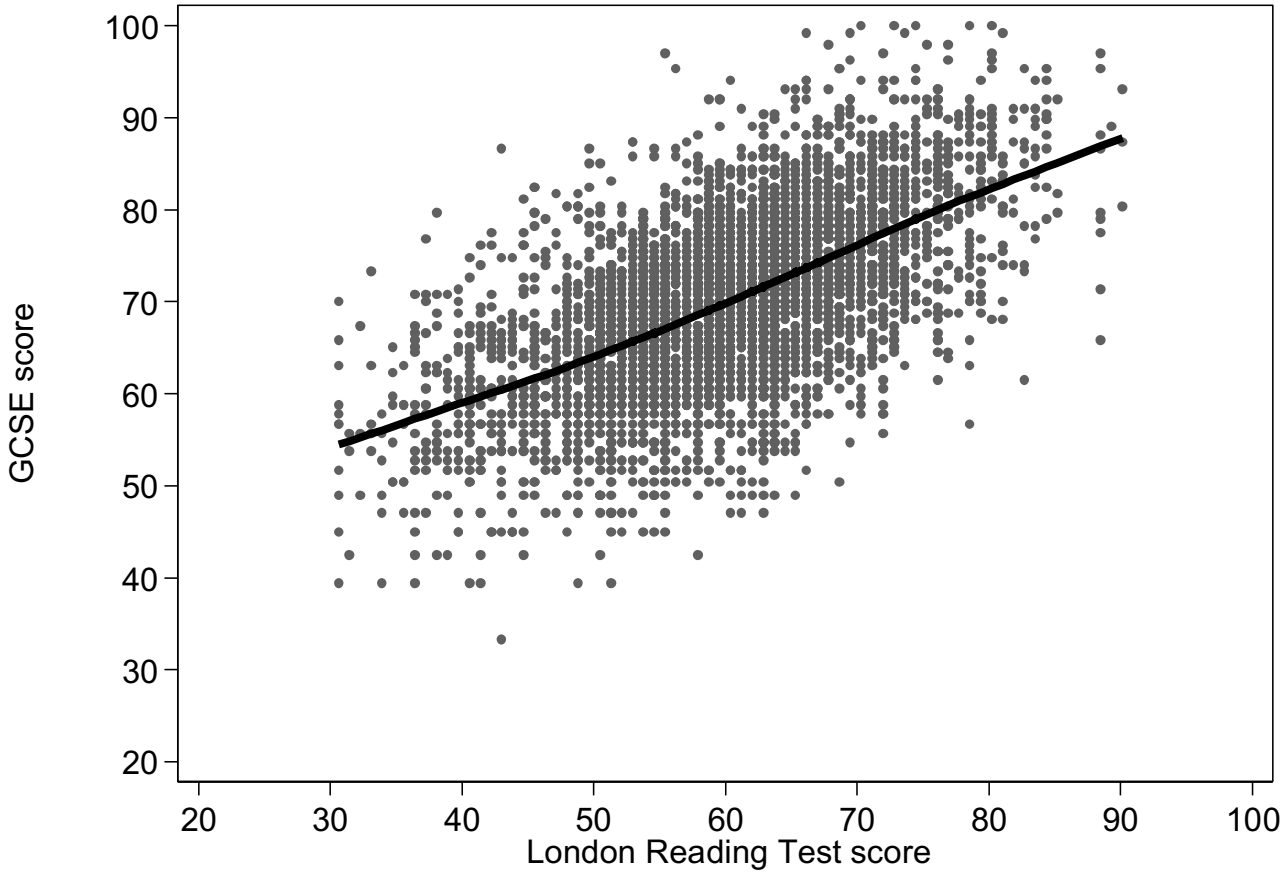
Possible Questions within 2-level Data

4. Are there differences in performance across school type (mixed gender, girls only and boys only)?
5. Within a school, does student gender modify the relationship between GCSE score and LRT score?
6. Does the type of school modify the within-school relationship between GCSE score and LRT score?

Visualization of the data

Quantify the relationship between GCSE score and LRT score

---- > TOTAL EFFECT, have ignored cluster membership



Visualization of the data

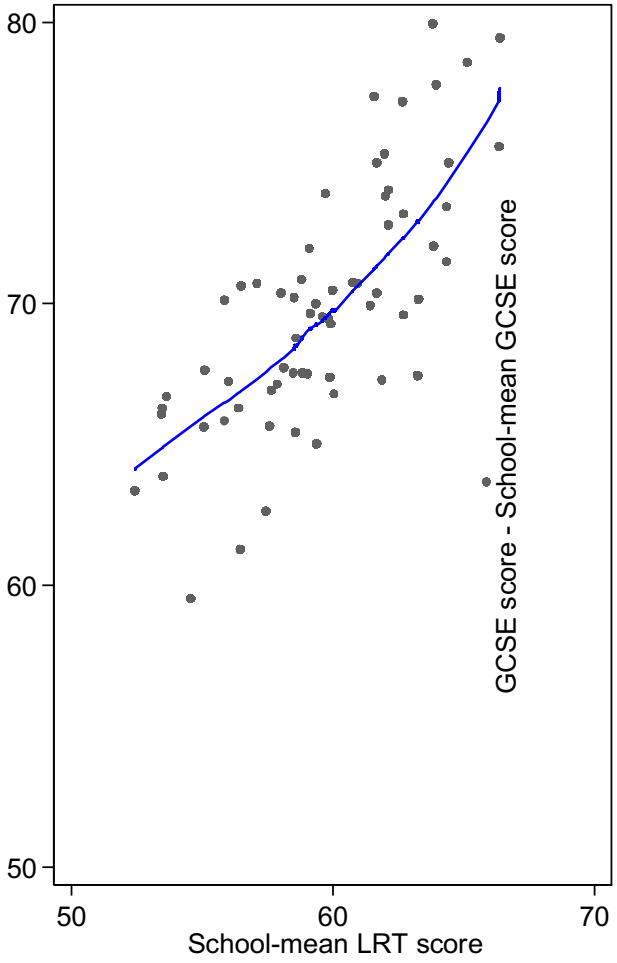
2. Within a school, quantify the relationship between GCSE score and LRT score
3. Does the "context" of the school matter? I.e. do students from schools with higher school-average LRT scores fair better than similar students in schools with lower school-average LRT scores

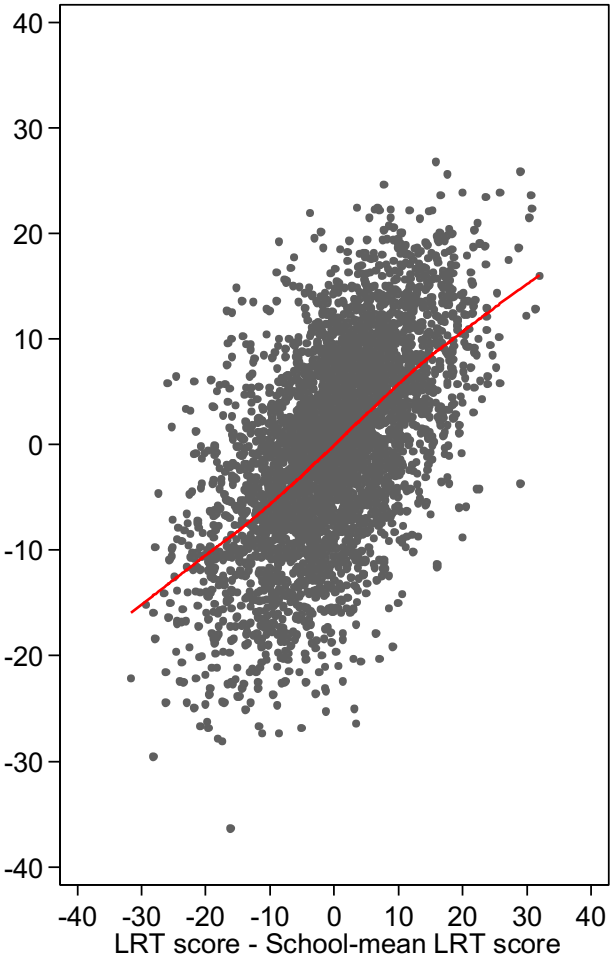
Remember the decomposition of the total effect:

$$X_{ij} = \left(X_{ij} - \bar{X}_i\right) + \bar{X}_i$$

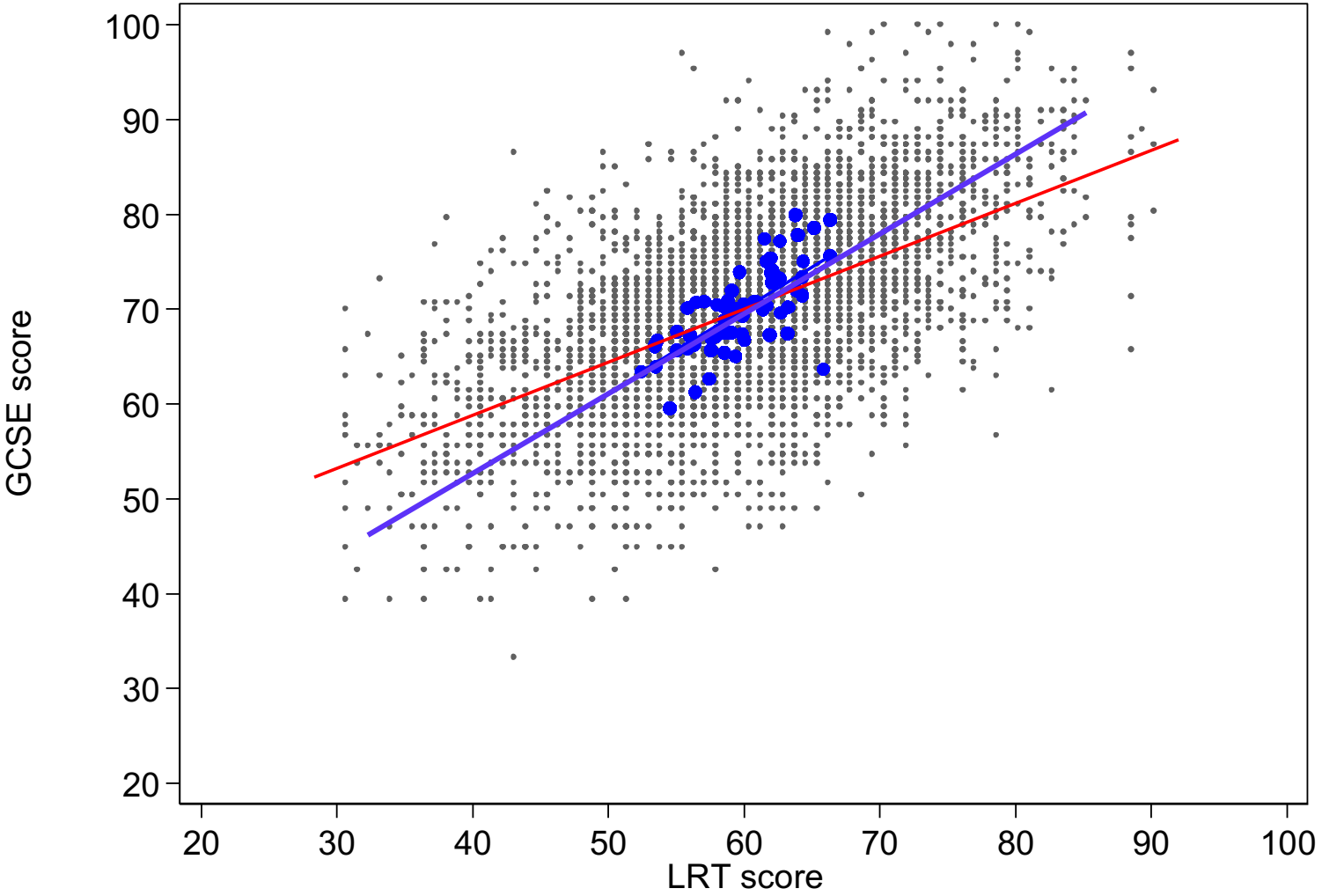
total effect = within + between

Separation of Between and Within Effects





Separation of Between and Within Effects



Ecological Fallacy

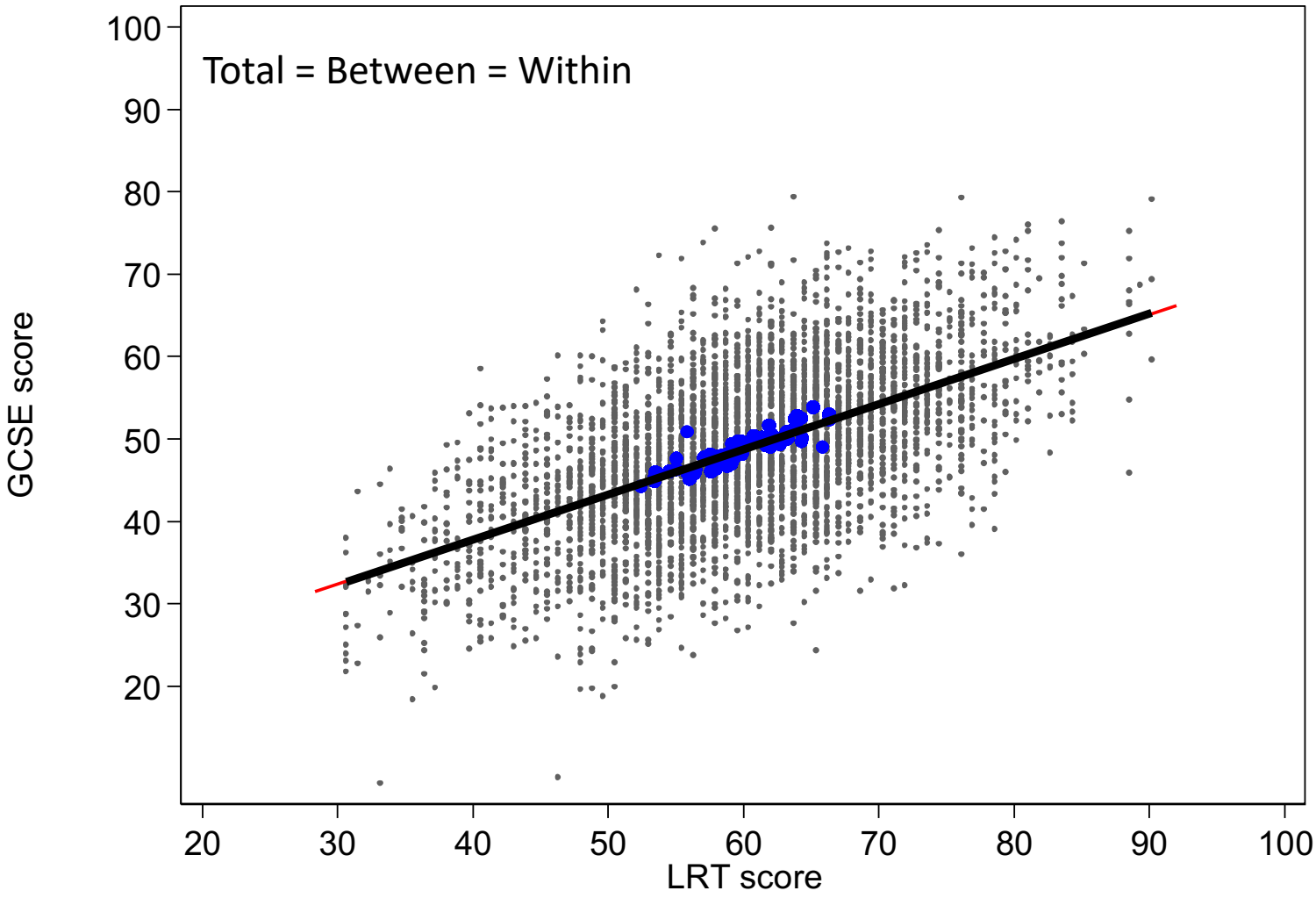
- The graph on the prior slide is a visualization of the ecological fallacy.

- Definition: a logical fallacy in the interpretation of statistical data where inferences about the nature of individuals are deduced from inference for the group to which those individuals belong.

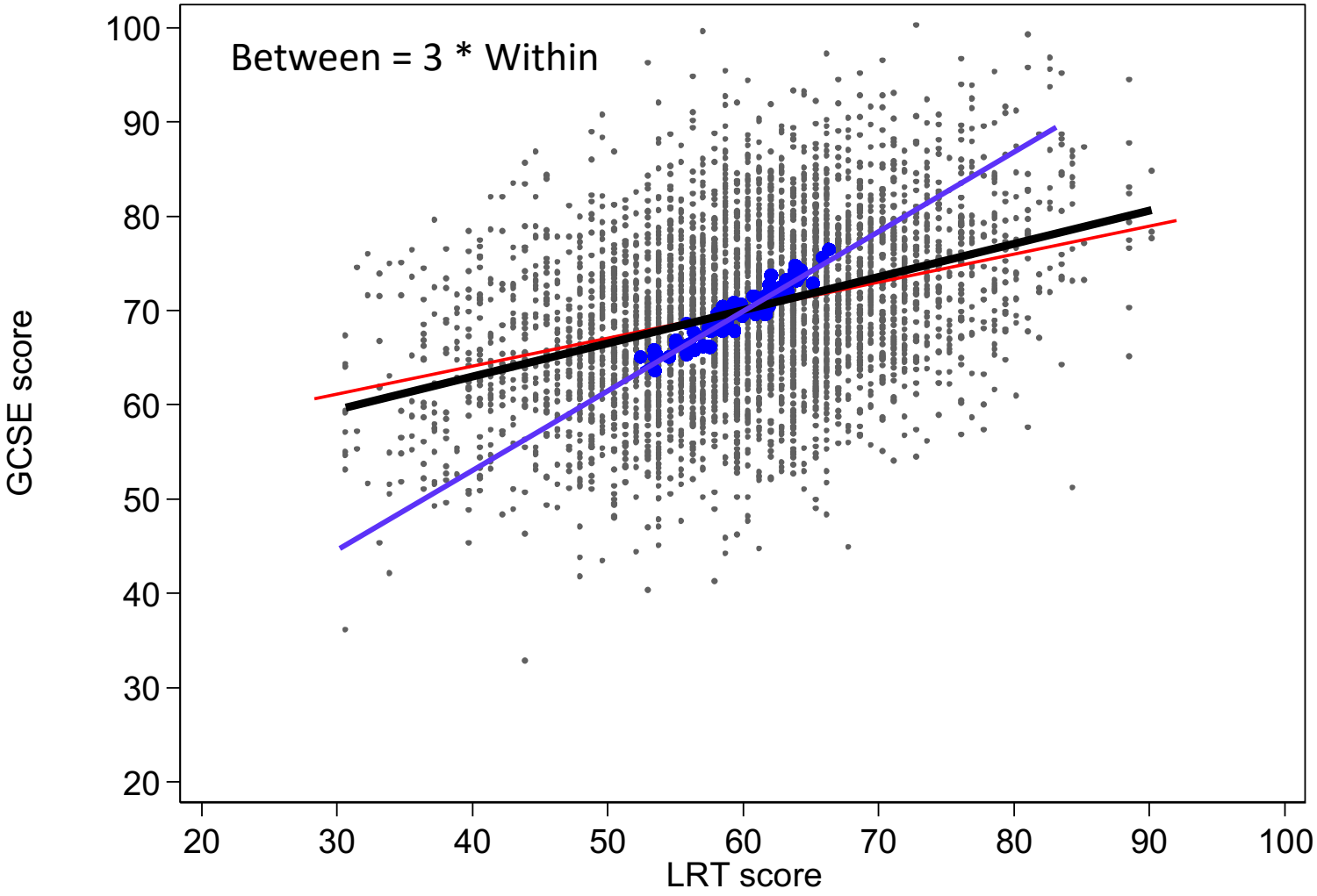
- The data demonstrate a different “between” and “within” association (although we have not tested this statistically yet)

- What would the data look like if there was no ecological fallacy?

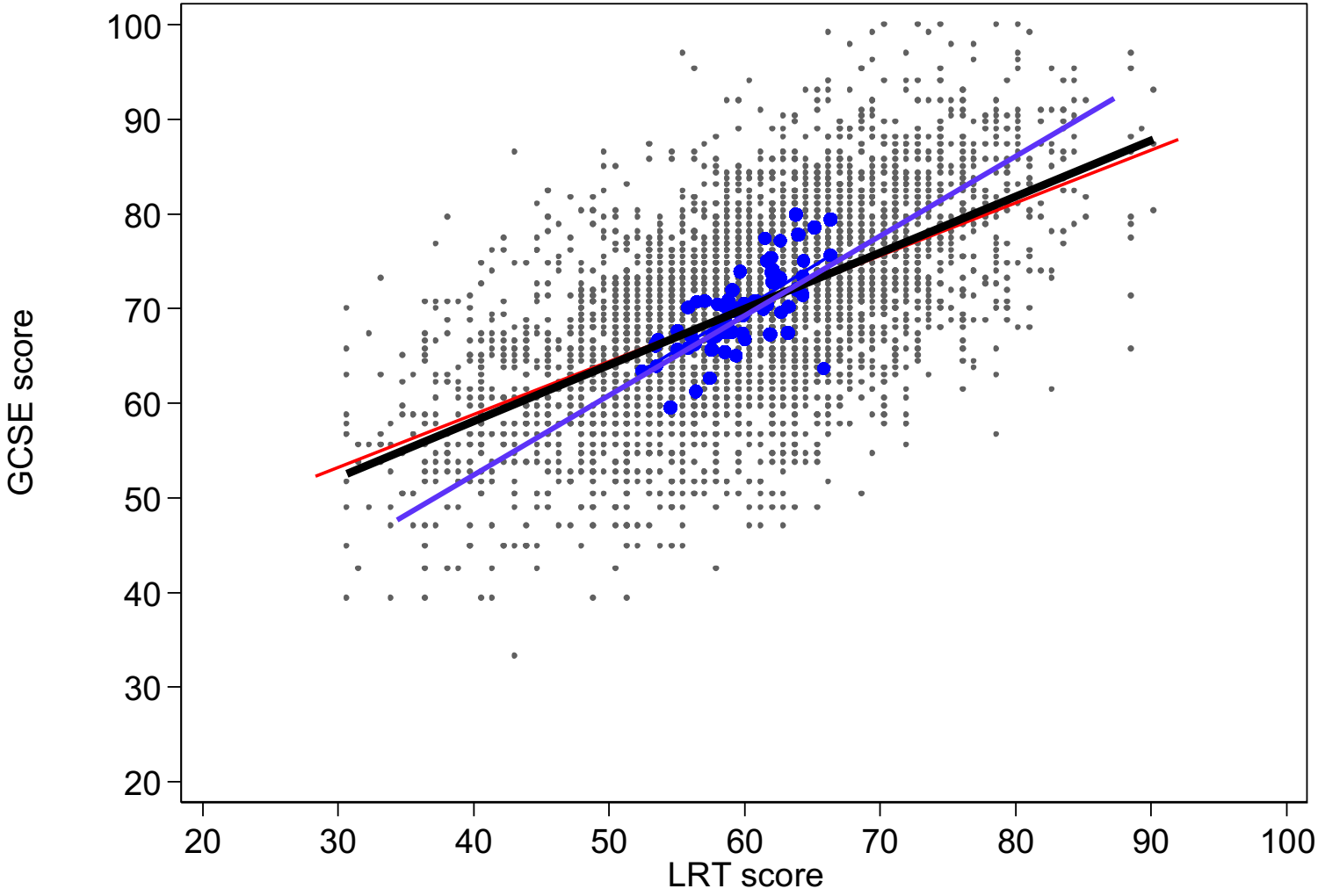
Ecological Fallacy



Large ecological fallacy



Total Effect = Weighted average of Between and Within



Visualization of the data

3. Does the “context” of the school matter? I.e. do students from schools with higher school-average LRT scores fair better than similar students in schools with lower school-average LRT scores

When would the “context” of the school matter?

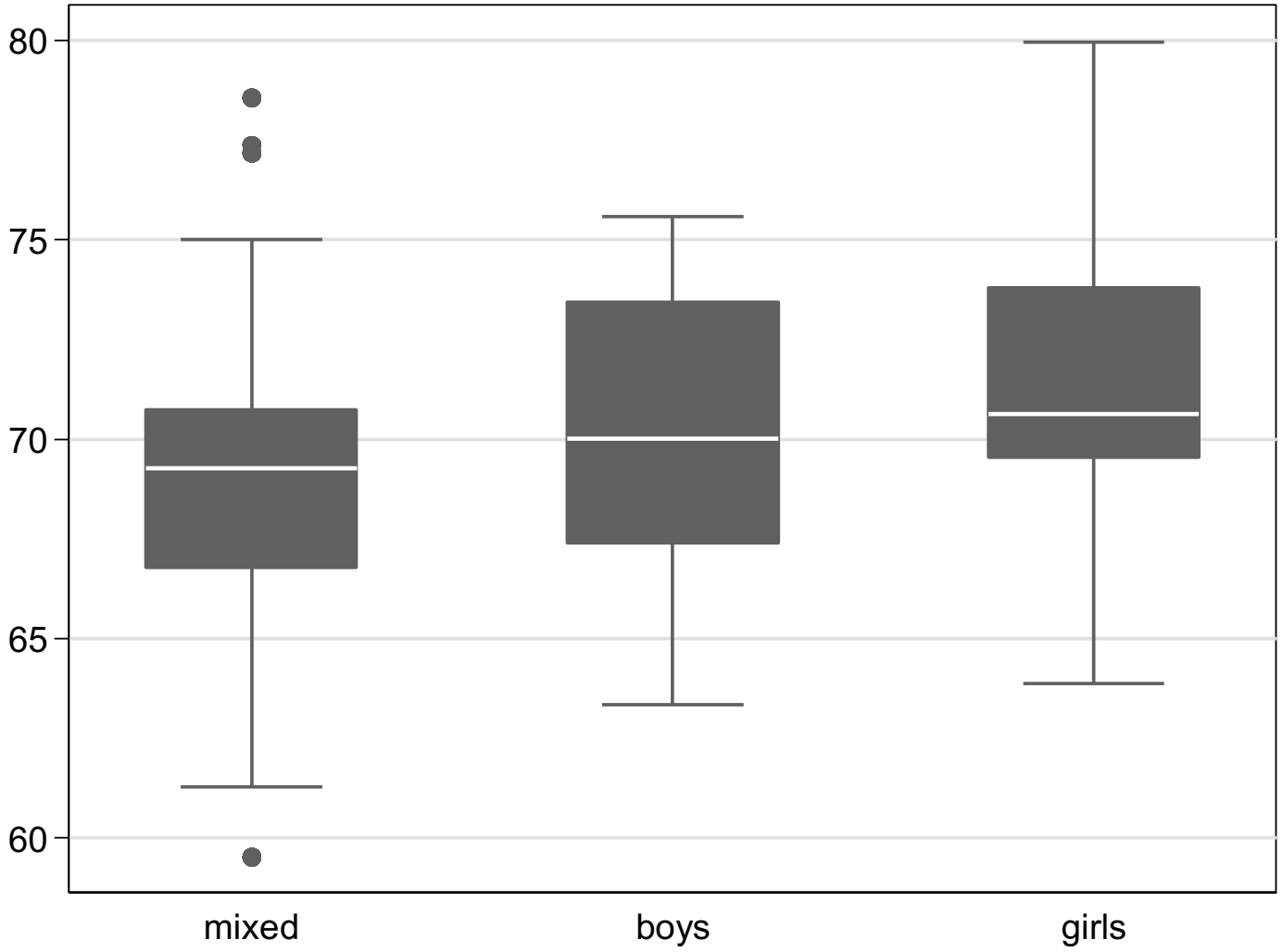
Possible Questions within 2-level Data

4. Are there differences in performance across school type (mixed gender, girls only and boys only)?

   School type is measured at level-2

   This is basically a question about the average school performance as a function of school type.

Exploration of Level-2 Exposure

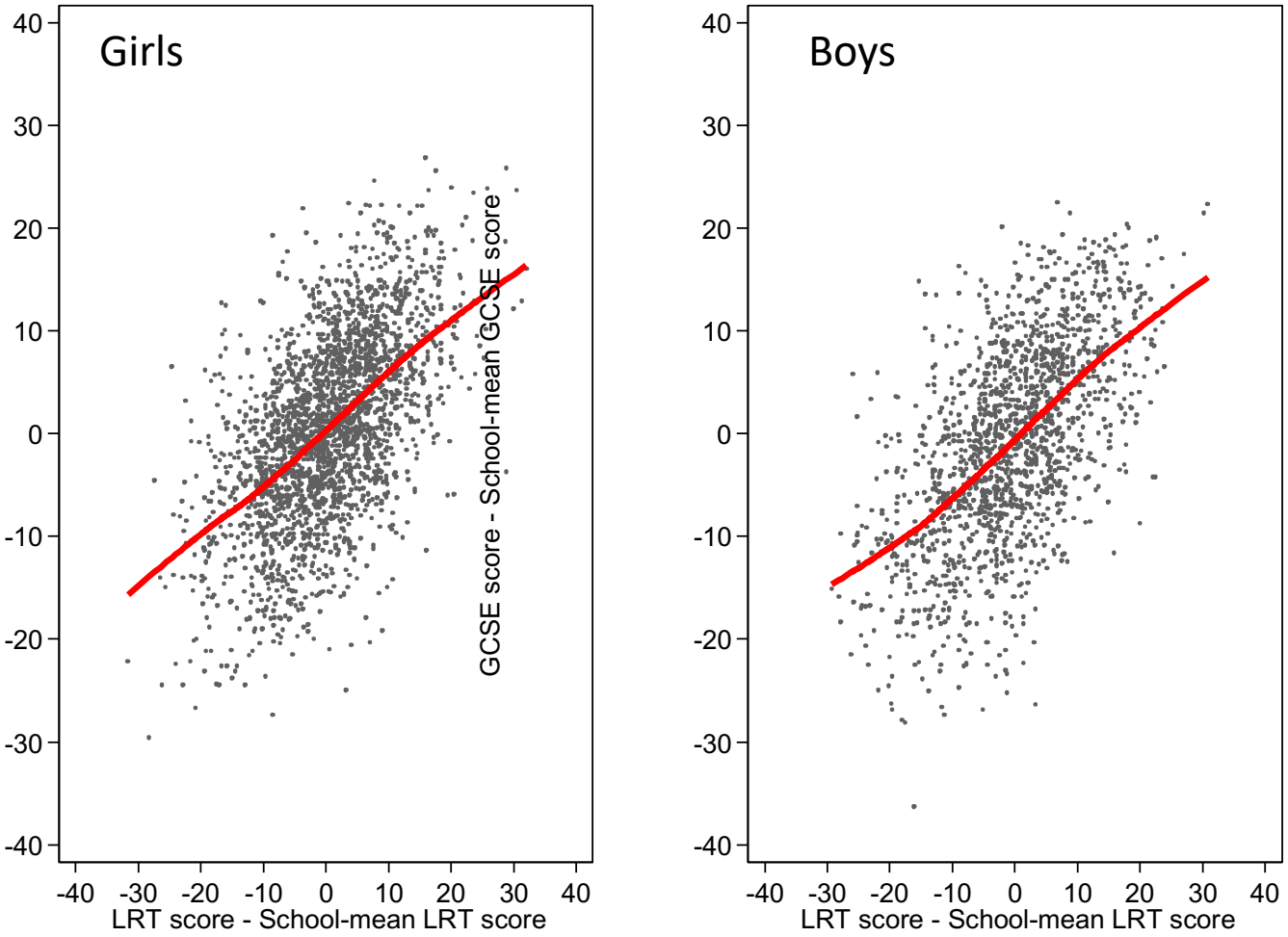


Possible Questions within 2-level Data

5. Within a school, does student gender modify the relationship between GCSE score and LRT score?

Here we will assess the relationship between GCSE and school mean centered LRT scores separately for male and female students.

Interaction of two level – 1 covariates



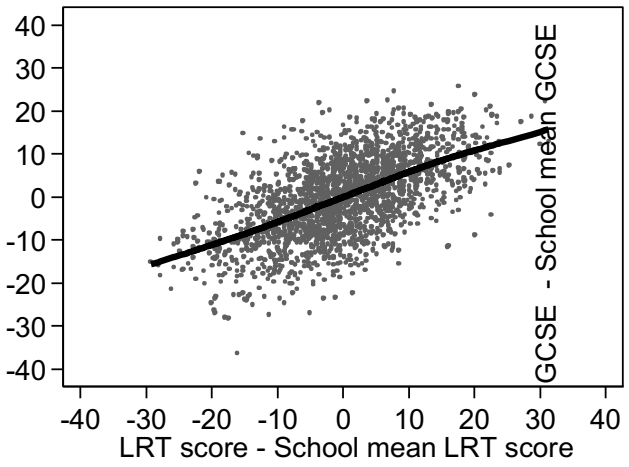
Possible Questions within 2-level Data

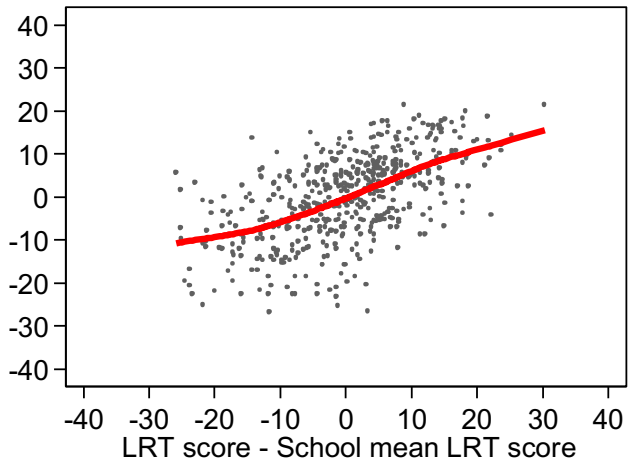
6. Does the type of school modify the within-school relationship between GCSE score and LRT score?

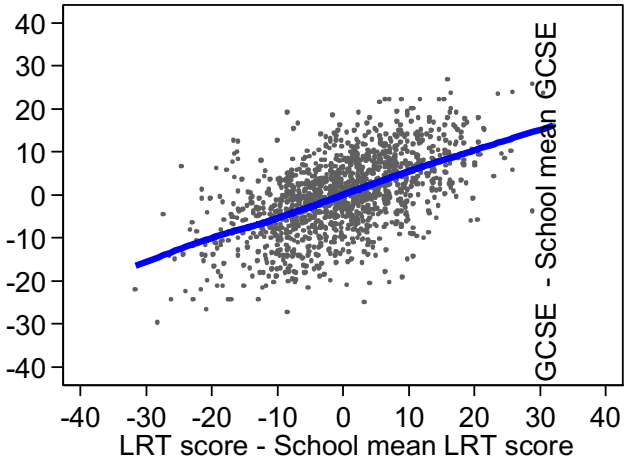
Similar to interaction of level-1 variable:

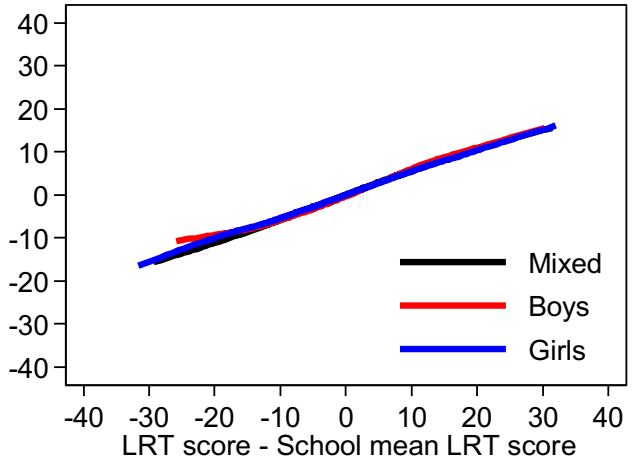
Interaction of level-1 with level-2

Level-1 Level-2 Interaction









Summary of Lecture 3

Within the context of a data example; Inner-London School Data

- Described various variable types
  - Level 1
  - Level 2 (measured and calculated)
- Described scientific questions of interest related to the fixed effects
- Presented some visual displays of the data to address each of the scientific questions
  - Discussed the separation of between and within effects in clustered data
  - How this relates to the ecological fallacy
  - In Lectures 4 and 5, we will fit linear mixed models to address each of these questions and review interpretation of the findings!