Week4 Mixed Models 1: Introduction to Multilevel Models

Assign	
Status	Epidemiology and Big Data

Mixed Model

random effect

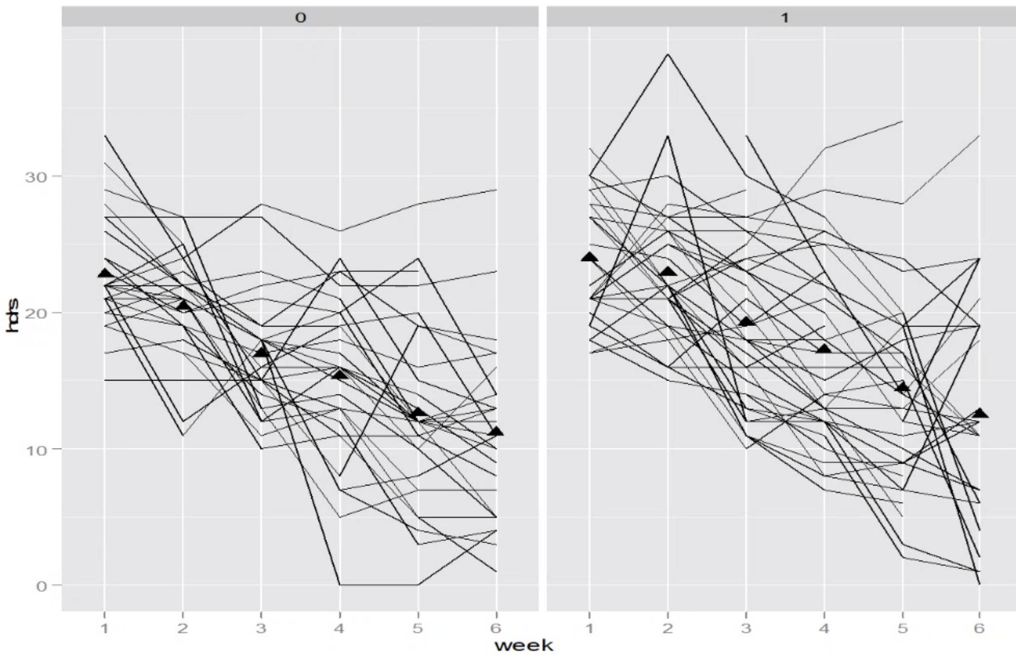
Reisby Data, After several trials

- std increase
- correlation decrease

ex vs indogenous

Example: Reisby Data

"Spaghetti Plot" (R, using ggplot2)



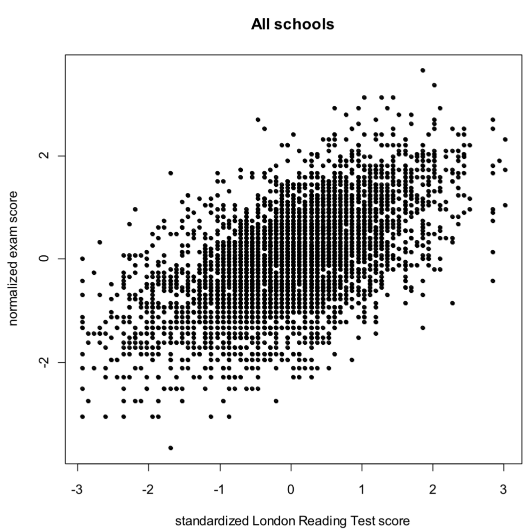
11

Introduction to multilevel data

Example: multilevel data (children within schools)

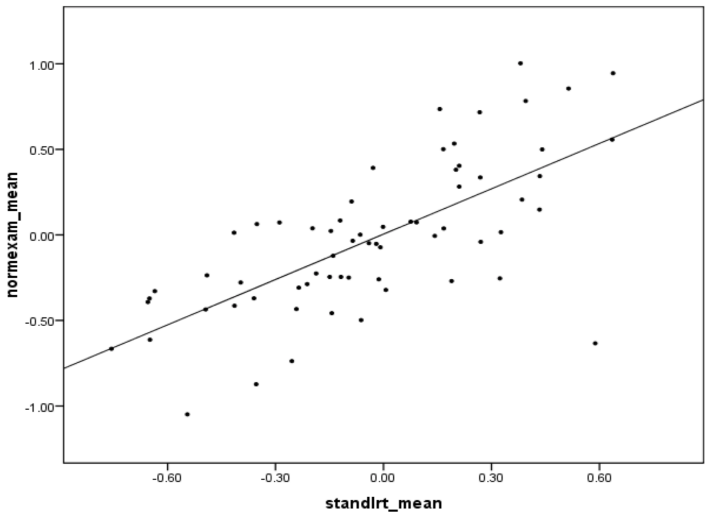
- Hierarchical structure of data
- Variation at all levels
- “Units” within a level expected to be correlated
- Variables can be measured at different level

Example: London Schools



Variables in dataset:

o School ID
o Student ID
o **Normalised exam score** (outcome variable)
o **Standardised LR test score**
o Student gender
o School gender
o School average of intake score
o Student level Verbal Reasoning (VR) score category at intake
o Category of students’ intake score (averaged)

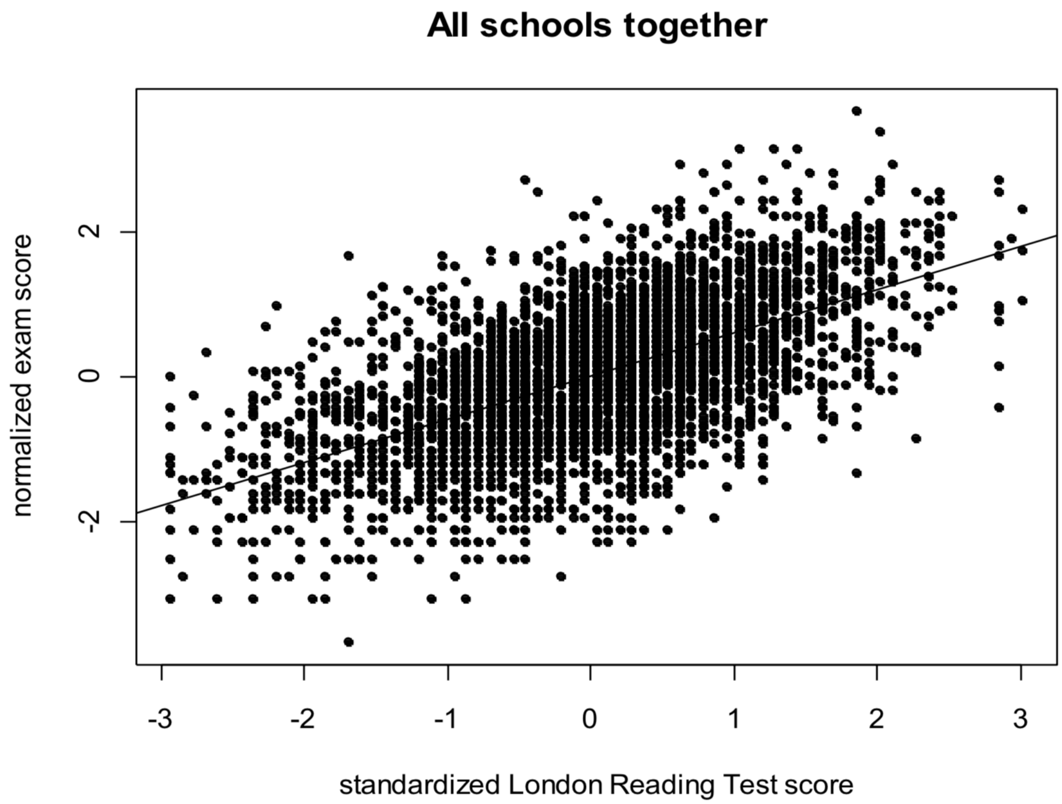


1. linear regression, mean exam per school vs mean LRT ("aggregated data")

- estimate for intercept: 0.005 (se 0.040)
- estimate for slope: 0.884 (se 0.116)

Disadvantages:

- every school (regardless of sample size) given equal weight
- N = 65
- school-level variables possible, but not child-level variables
- we can only make inference at school level, not child-level
- possibility of "ecological fallacy"

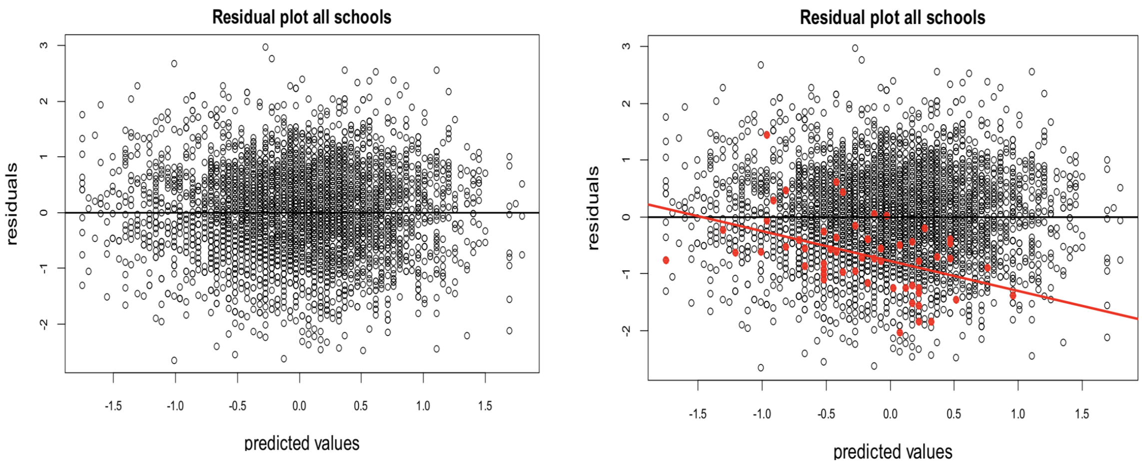


2. linear regression, all schools together ("disaggregated data")

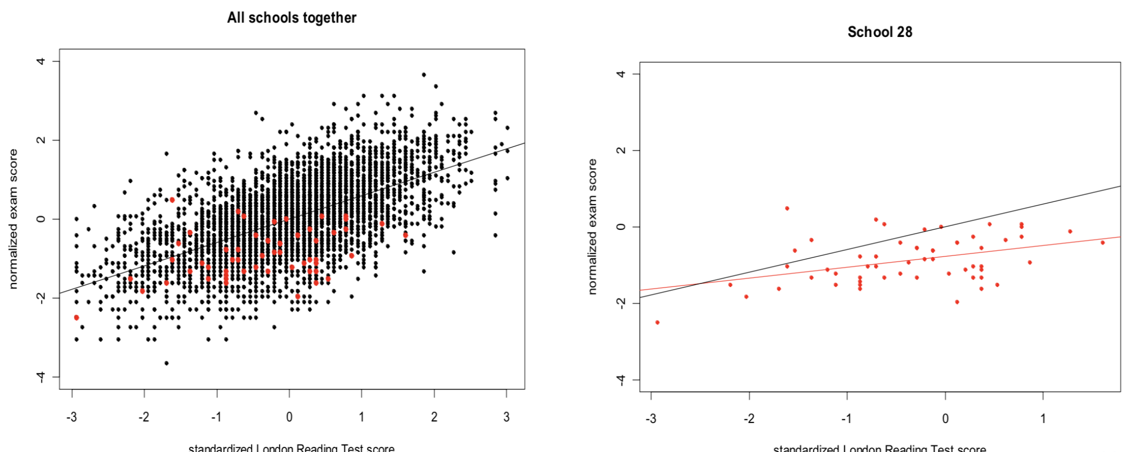
- estimate for intercept: - 0.001 (se 0.013)
- estimate for slope: 0.595 (se 0.013)

Disadvantages:

- inflates sample size, especially for level-2 variables
  - SE's of level-2 variables tend to be underestimated → p-values too small, CI's too narrow (type I error inflated)
  - SE's of level-1 variables may be over or underestimated
- ignore correlated residuals (correlation of children within schools)



?



If we split the School_28 separately, we will find that the slope is totally different

3. linear regression per school

summary intercepts:

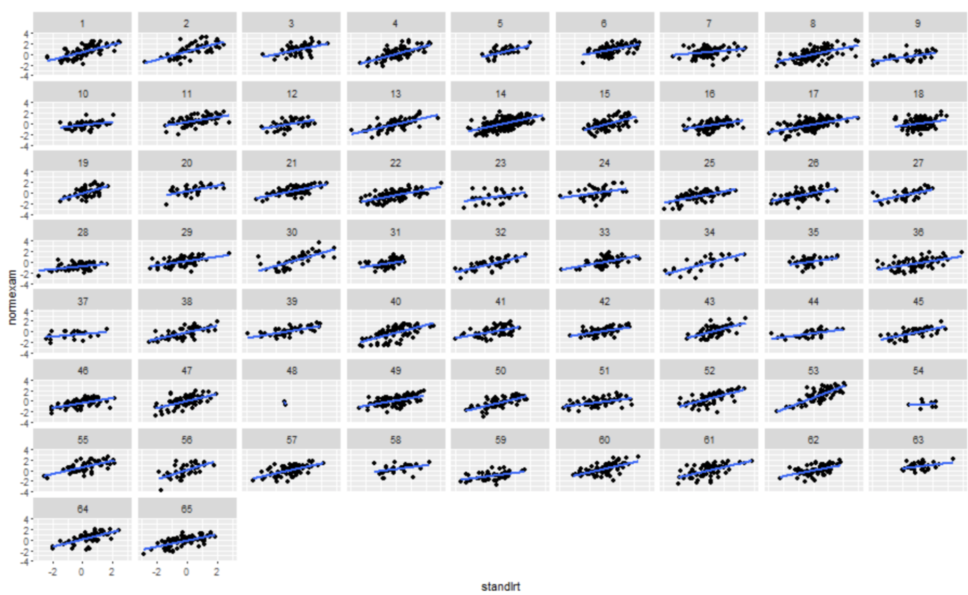
- mean=-0.068;sd=0.519;sem=0.064

summary slopes:

- mean=0.425;sd=0.939;sem=0.116

Disadvantages:

- 65 different regressions, how to combine the results?
    - mean slope: every school has equal weight
    - standard error of parameter estimate correct?
- child-level variables possible, but not school-level variables



4. linear regression, all schools together, regression with main effect and interactions to allow for different intercepts and slopes

Advantage over previous analysis:

- now we can include both child- and school-level variables
- residuals probably normally distributed (with a constant variance?) around individual lines

Disadvantages:

- We wanted 1 intercept and 1 slope for LRT, but:
- 65 schools, so 1 reference category and 64 estimates for intercepts (main effects per school) + 64 estimates for interactions (slopes per school)!
- Which school is the reference?
    - We can't generalize beyond these 65 schools
    - This model uses 128 extra df for all those intercepts & slopes

5. Linear Mixed Model

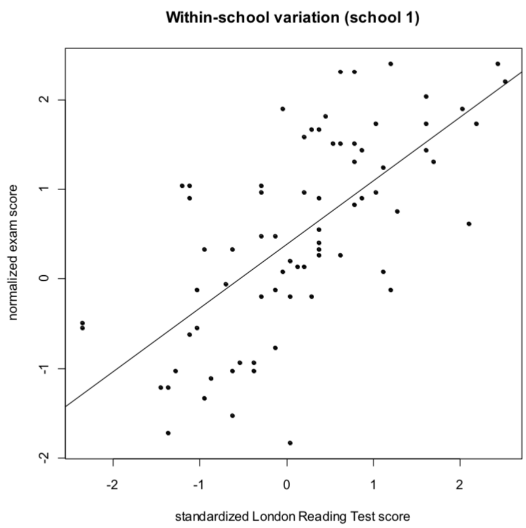
Advantages:

- sample size correct, account for correlation of children within schools
    - so: correct SE's/p-values/CI's
- no need for 64 main effects and interactions
    - differences between schools captured one or more 'variance components'
- both child-level and school-level variables simultaneously
    - so: inference for both children and schools
    - interactions between child- and school-level variables possible
- examine variation at different levels
- models work well in presence of missing outcomes (longitudinal)

Mixed Models = random effects models (inaccurately) = random coefficient models

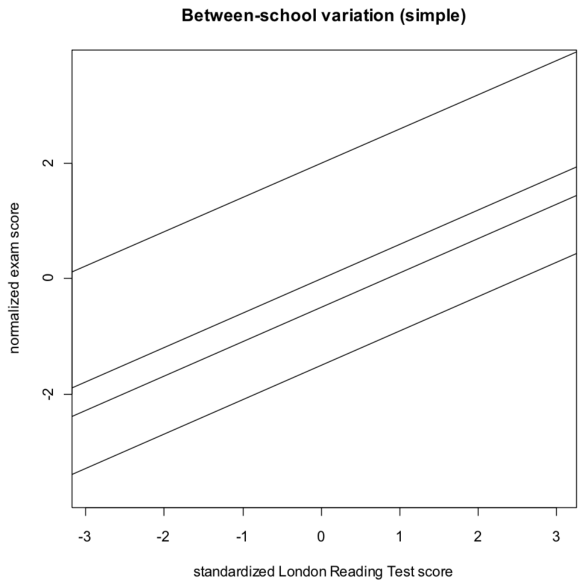
- Mixed models = fixed effects + random effects
- Some variables (or: their coefficients) can be included as both "fixed" (of interest) and "random" (random variation across the **level-2 units**)

Random intercept only vs Random intercept + random slope

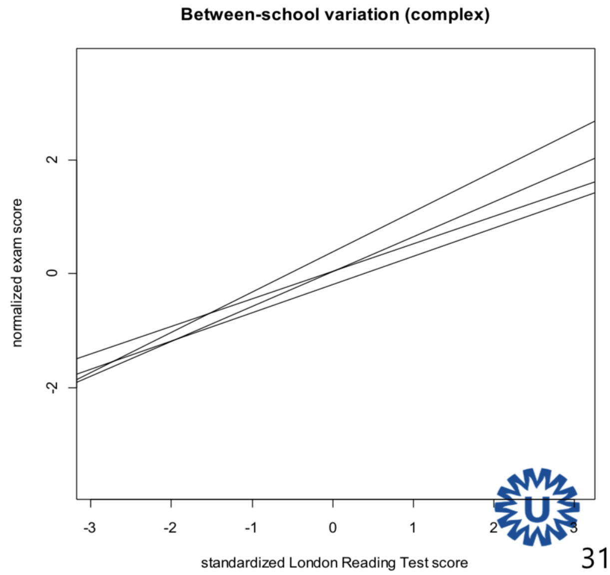


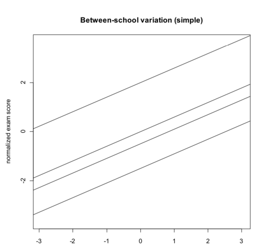
$$y_{1j} = \beta_0 + \beta_1 X_{11j} + \varepsilon_{1j}$$

Random intercept only:



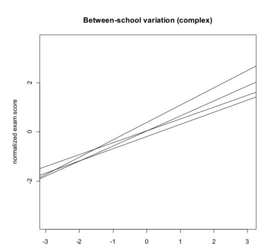
Random intercept + random slope:





$$y_{1j} = \beta_0 + \upsilon_{01} + \beta_1 X_{11j} + \varepsilon_{1j}$$
$$y_{2j} = \beta_0 + \upsilon_{02} + \beta_1 X_{12j} + \varepsilon_{2j}$$

$$y_{ij} = \beta_0 + \upsilon_{0i} + \beta_1 X_{1ij} + \varepsilon_{ij}$$



$$y_{1j} = \beta_0 + \upsilon_{01} + \beta_1 X_{11j} + \upsilon_{11} X_{11j} + \varepsilon_{1j}$$
$$y_{2j} = \beta_0 + \upsilon_{02} + \beta_1 X_{12j} + \upsilon_{12} X_{12j} + \varepsilon_{2j}$$

$$y_{ij} = \beta_0 + \upsilon_{0i} + \beta_1 X_{1ij} + \upsilon_{1i} X_{1ij} + \varepsilon_{ij}$$

fixed effect

individual effect (random effect)

Interlude: some notation

intercept slope

- level-1 (child) model: $y_{ij} = b_{0i} + b_{1i} \cdot x_{1ij} + \varepsilon_{ij}$
- level-2 (school) model: $b_{0i} = \boxed{\beta_0} + \boxed{\upsilon_{0i}}$; $b_{1i} = \boxed{\beta_1} + \boxed{\upsilon_{1i}}$
- combine the two: $y_{ij} = \beta_0 + \upsilon_{0i} + \beta_1 \cdot x_{1ij} + \upsilon_{1i} \cdot x_{1ij} + \varepsilon_{ij}$
  - rewrite: $y_{ij} = (\beta_0 + \upsilon_{0i}) + (\beta_1 + \upsilon_{1i}) \cdot x_{1ij} + \varepsilon_{ij}$

- y_{ij} : outcome (exam score) for j^th child in i^th school
- x_{1ij}: 1st explanatory var (LRT score) at level 1 (j^th child in i^th school)
- $\beta_0, \beta_1, \ldots$: regression coeffients for overall effects of explanatory vars ("fixed effects")
- υ_{0i} : individual effect of i^th school on intercept ("random effect")
- υ_{1i} : individual effect of i^th school on slope (for LRT) ("random effect")
- ε_{ij} : level-1 residual (j^th child in i^th school)

30

- A random intercept per school allows schools to have different intercepts
- vice versa, A random slope ⇒ different slope

- All variables of interest are added as fixed (The target should be fixed)

- Model assumptions:
  - $\varepsilon_{ij} \sim N(0, \sigma_e^{\,2})$; $\upsilon_{0i} \sim N(0, \sigma_{\upsilon 0}^{\,2})$; $\upsilon_{1i} \sim N(0, \sigma_{\upsilon 1}^{\,2})$
  - ε_{ij} independent
  - $cov(\upsilon_{0i}, \upsilon_{1i}) = \sigma_{\upsilon 01}$
  - $cov(\varepsilon_{ij}, \upsilon_{0i}) = cov(\varepsilon_{ij}, \upsilon_{1i}) = 0$

The parameter μ is the mean or expectation of the distribution (and also its median and mode), while the parameter σ is its standard deviation.[1] The variance of the distribution is σ^2

- normal distribution mean = 0, variance = 1
- **(mean, standard deviation, variance)**

The problem, and some possible solutions

The mixed model solution

Adding random effects (random intercept, random slope)

Adding fixed effects (school- and child-level) to the model

Interpretation of mixed models

Summary

Q&A

p18

sem?

child params

LRT? = London Reading test

level2 unit

variable of interest = 哪些是你關注的變量 (目標變量)

Mixed Models: what is a "random effect"?

- A random intercept per school allows schools to have different intercepts
- A random effect for LRT per school allows the effect of LRT on exam score to differ per school ("random slope for LRT" = different slope for exam-LRT relation for each school)
- Random effect ("slope") can also be for a categorical variable
  - difference between boys and girls on exam score could differ per school
  - treatment effect on an outcome can be thought to vary per center in a multi-center study
- All variables of interest are added as fixed
- Depending on theory, none/one/some fixed variables may also be modelled as random

28

Random effect ("slope") can also be for a categorical variable, (intercept can be too?)