Introduction to Multilevel Modelling

SGSSS Summer School 2022

Patricio Troncoso & Ana Morales-Gómez

Heriot-Watt University

University of Edinburgh

Scottish Centre for Administrative Data Research (SCADR)

Housekeeping

Who are we?

Patricio Troncoso

- Research Fellow, I-SPHERE, Heriot-Watt University
- Works (mostly) in Educational research
- Twitter: @ptronc
- Email: p.troncoso@hw.ac.uk

Ana Morales-Gómez

- Research Fellow, School of Law, The University of Edinburgh
- Works (mostly) in Criminological research
- Twitter: @A_moralesgomez
- Email: Ana.Morales@ed.ac.uk

We both work in the Scottish Centre for Administrative Data Research (SCADR). Find out more here

Today's programme

- 10:00-10:30 - A brief overview of linear regression
- 10:30-11:00 - Multilevel data structures and examples
- 11:00-11:15 - Break
- 11:15-11:45 - Variance components and group-specific estimates
- 11:45-12:30 - Practical 1
- 12:30-13:30 - Lunch
- 13:30-13:50 - Accounting for individual and group characteristics: fixed effects
- 14:00-14:30 - Practical 2
- 14:30-15:00 - Multilevel modelling for binary responses
- 15:00-15:45 - Practical 3
- 15:45-16:00 - Break
- 16:00-16:20 - Differential processes between groups: random effects
- 16:20-16:50 - Practical 4
- 16:50-17:00 - Wrap up and Finish

Today's learning goals

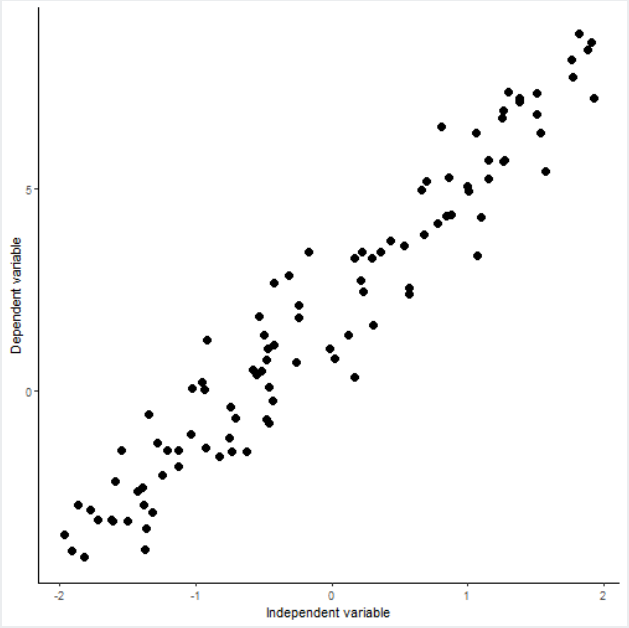
- Understand the general concept of multilevel modelling and its applications in social science research questions
- Understand a range of multilevel models and when to use them
- Specify multilevel models for continuous and binary responses using R
- Interpret the R output of standard multilevel models

Part One:

In the beginning, there was… linear regression

Correlation

The easiest way to see the relationship between two continuous variables is to plot the values of one of them against the other.



This is presupposing that you have a theoretical reason to think they are indeed related

Why is there a correlation?

- There might be a causal link:
  - X causes Y or vice versa
- Both variables are affected by another variable
  - We call this “confounding”
- Both variables measure the same thing but in different ways
- Last but not least...

A correlation can occur purely by chance!

Correlation doesn't mean...

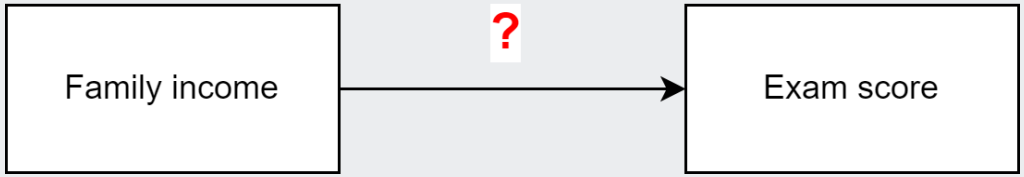
- Causation
  - Trends over time can look as if there is causation, but they can be a natural process.
    - Reading comprehension and age
- Also, a low correlation (close to zero) does not mean there is no association either
  - A "non-linear" relationship might exist
  - There can be groups/clusters with varying associations (we'll get back to this)

What is Linear Regression? (1)

It's the process by which we systematise the relationship between two (or more) variables.

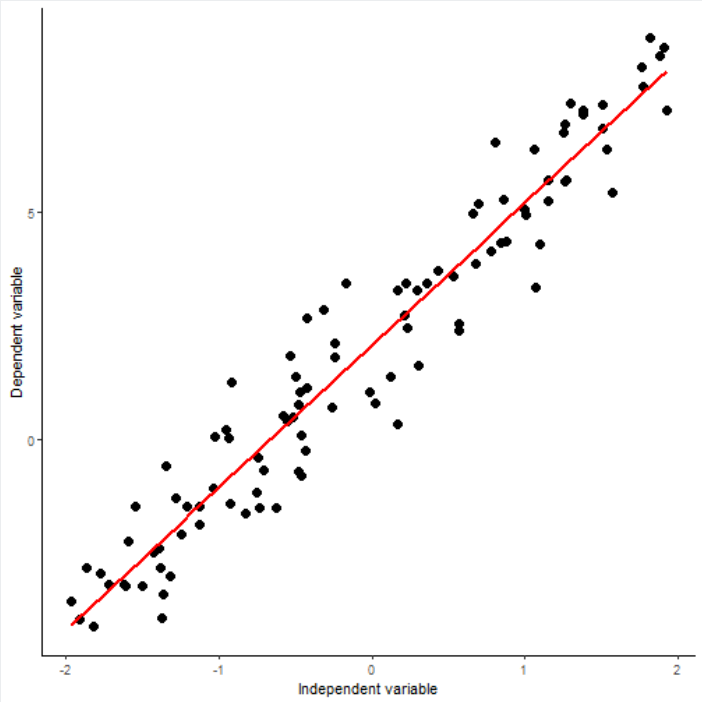
We look for the "best fit" to the data we have.

We obtain an equation that describes how much our dependent variable (y) changes as our independent variable (x) changes.



What is Linear Regression? (2)

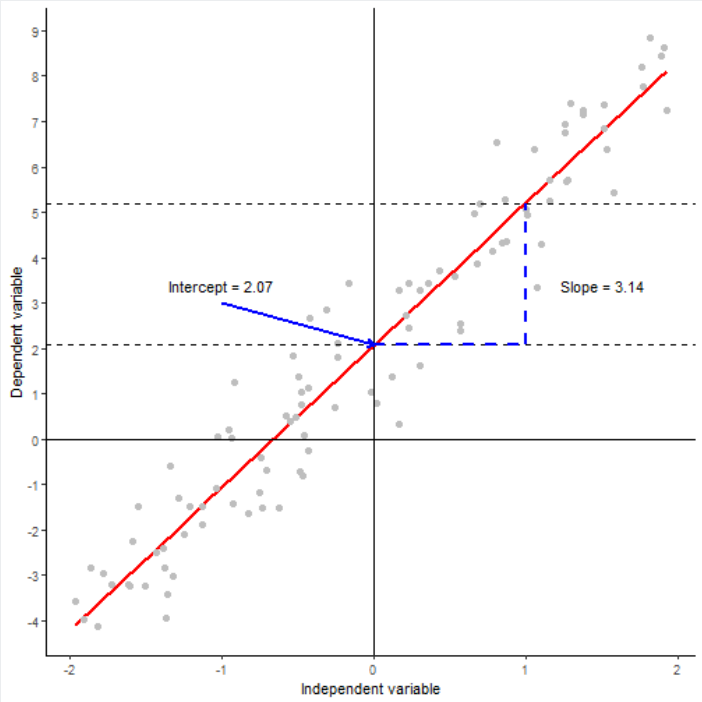
In the presence of an association between two variables, we can fit a straight line to define the relationship



This is what we typically call "the line of best fit", because it reduces the error to the minimum

What is Linear Regression? (3)

In the presence of an association between two variables, we can fit a straight line to define the relationship



The regression equation would be:

$$y = 2.07 + 3.13x + e$$

What is Linear Regression? (4)

A Simple Linear Regression model has the following form:

$$y = \beta_0 + \beta_1 x + e$$

Where:

y is the value of the dependent variable

β_0 is the intercept (point at which the line crosses the y axis)

β_1 is the slope (expected increase in y given a one unit increase in x)

x is the value of the independent variable

e is the error term

How do you do this in R?

Assuming we have a dataset called "data", containing variables "y" and "x":

```
summary(lm(y ~ x, data = data))
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.24324 -0.55252  0.01285  0.65260  2.07716
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.07135    0.09477   21.86   <2e-16 ***
## x            3.13895    0.08588   36.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9451 on 98 degrees of freedom
## Multiple R-squared:  0.9317,    Adjusted R-squared:  0.931
## F-statistic:  1336 on 1 and 98 DF,  p-value: < 2.2e-16
```

How do you interpret R output?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.07135    0.09477   21.86   <2e-16 ***
x            3.13895    0.08588   36.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9451 on 98 degrees of freedom
Multiple R-squared:  0.9317,    Adjusted R-squared:  0.931
F-statistic:  1336 on 1 and 98 DF,  p-value: < 2.2e-16
```

- The overall mean or intercept is *2.07*
- A one unit increase in *x* is associated with a *3.14* increase in *y*
- Both coefficients are statistically significant (*p<0.05*)
- The regression model explains *93.1%* of the variance in *y*

Multiple Linear Regression

As hinted in its name, multiple linear regression is when we have more than one independent variable in our model

- We rarely ever use models with only one "x"
- How is this done? Simply add more x's to the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + e$$

- Interpretation remains largely similar, but:
  - Effect of one IV on DV refers to when the other IVs remain constant
  - This is why you see in papers phrases like:
    - "the effect of x1 on y while adjusting/controlling for x2 and x3..."
  - Intercept is now the value of DV when all IVs are *zero*

MLR example output

This is a subset of a large dataset of examination results in London schools

- The DV is "normexam": normalised exam score at age 16
- The IVs are
  - "standlrt": score at age 11 on the London Reading Test (LRT)
  - "sex": coded as girl or boy
- We used this code to run the model:

```
lm(normexam ~ standlrt + factor(sex), data= tutorial)
```

```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     -0.1031842 0.01990450 -5.183966 2.277999e-07
## standlrt         0.5905958 0.01268159 46.571132 0.000000e+00
## factor(sex)girl  0.1699601 0.02570919  6.610869 4.317291e-11
```

Question time!

- What is the intercept?
- What is the effect of the score at age 11 on the score at age 16?
- What is the effect of sex on the exam score at age 16?

Model selection

How do you decide what variables to put in a model?

- First of all: Theory!
- Model nesting: start with what is known and add your own hypotheses after that

If you have multiple models with different variables, how do you decide which one is better?

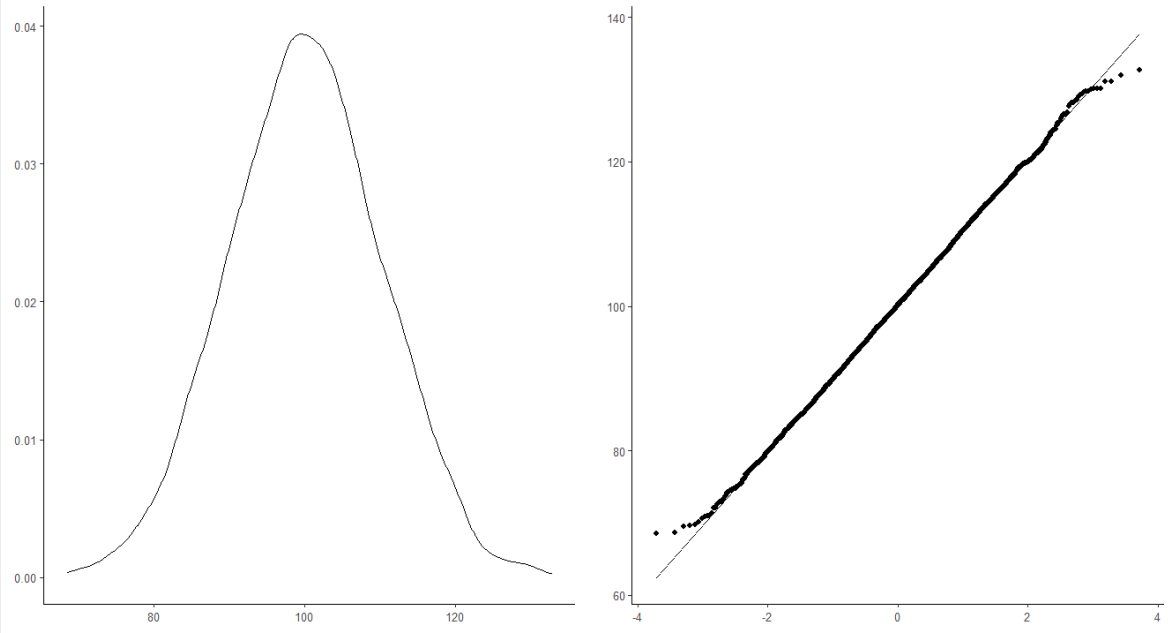
- Theory!
- Statistical criteria: R Squared comparison, comparing goodness of fit criteria, the likelihood-ratio test, AIC and BIC values, etc.
    - We'll get back to this as this is vital in MLM

MLR Assumptions (1)

Normality of residuals

The residual term is a continuous random variable, and as such, it should have a normal distribution.

This roughly means that whatever is left unexplained by our model can be thought of as “random white noise”

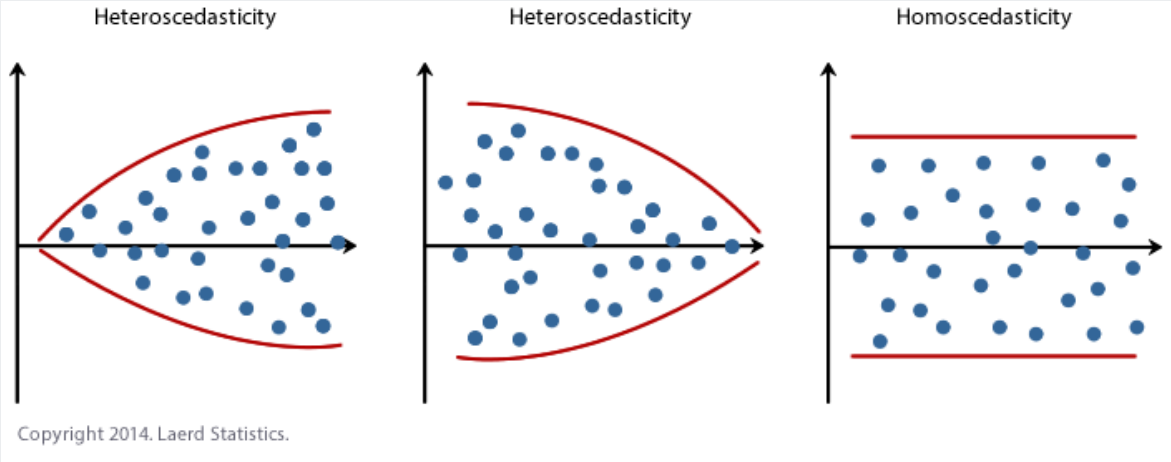


MLR Assumptions (2)

Homoscedasticity

Easier said than done (!)

The variation of the residuals should remain constant across the predicted values of our model



Source: Laerd Statistics

MLR Assumptions (3)

Independence of observations

Observations (or individuals) should not be related to one another or somehow "clustered"

This is a difficult assumption to meet. Some examples:

- Surveys usually sample whole groups
- Individuals live in households, neighbourhoods, attend schools, etc.
    - Those are all clusters that can make them be related to each other
- This is where MLM comes into play!

Part Two:

Multilevel data structures

Why MLM?

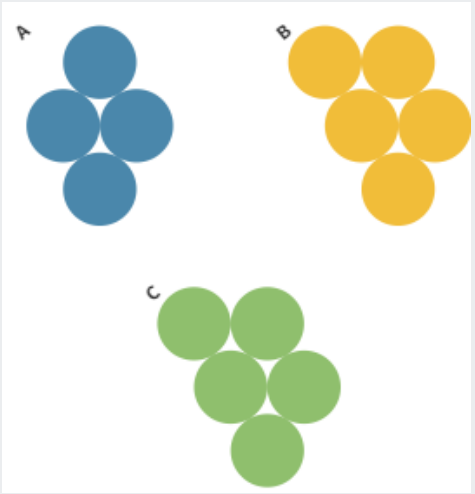
Data in the real world tend to violate the assumptions of:

- Independence
- Homogeneity of residual variance

Often, in reality, data is **hierarchically structured**

Data structures (1)

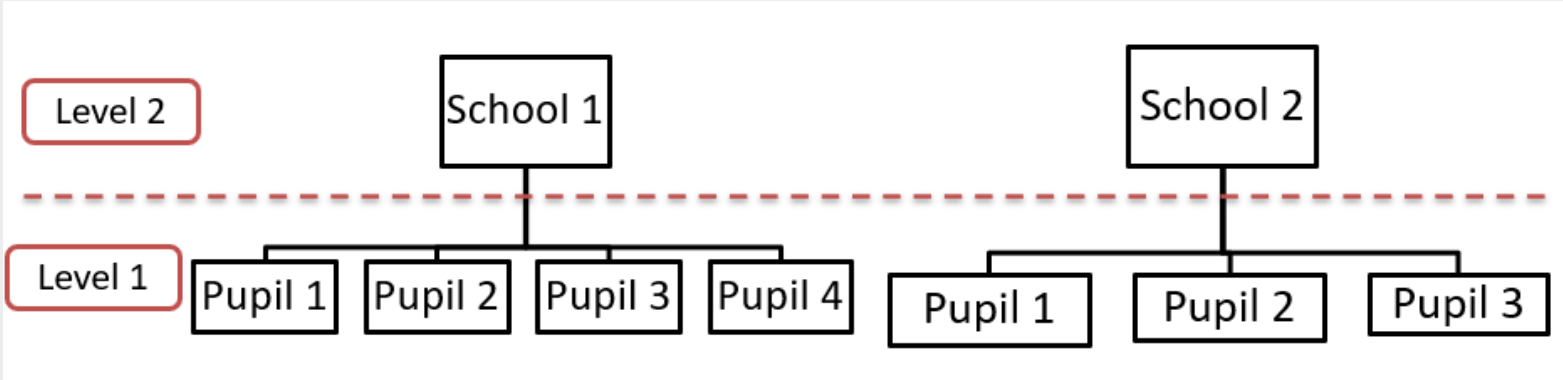
Individuals usually belong to groups:



- Schools
- Neighbourhoods
- Hospitals
- Prisons

Data structures (2)

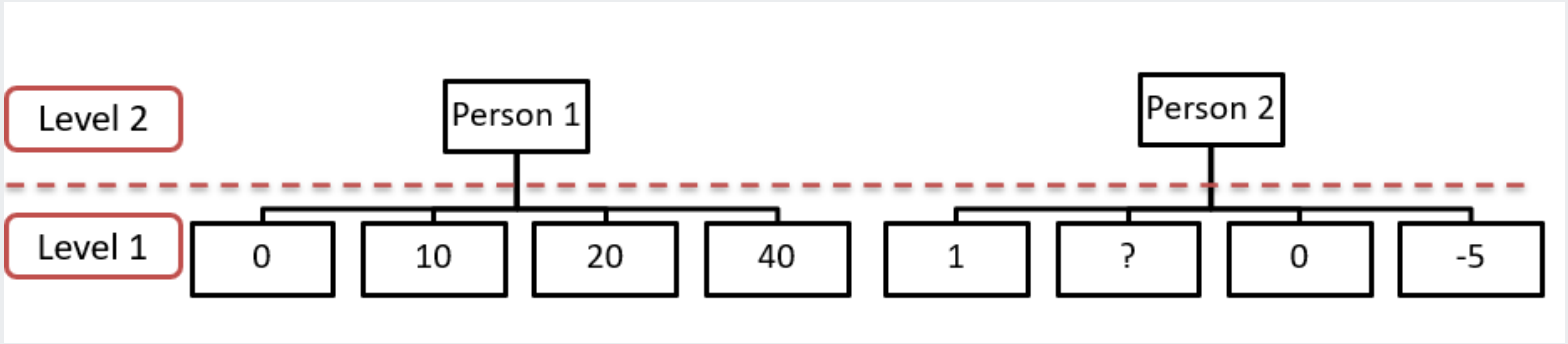
The "classic" pupils in schools structure:



- This is what we call a "2-level model"
- This is the typical structure of the "school value-added models"

Data structures (3)

But the structure doesn't need to be composed of individuals in groups:

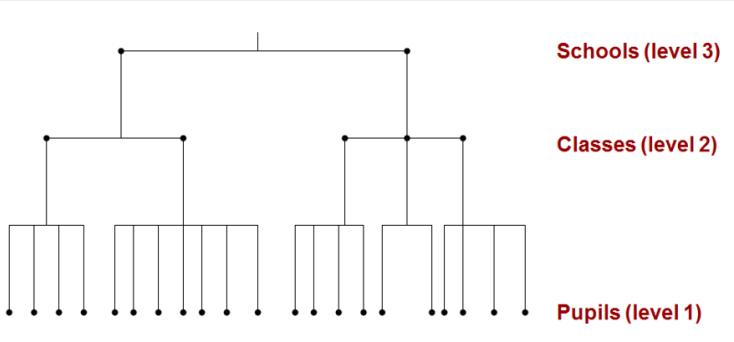


It can also be repeated measures nested within individuals

- This is a longitudinal model structure
- It can be referred to as a "growth curve model" or as "the multilevel model for change"

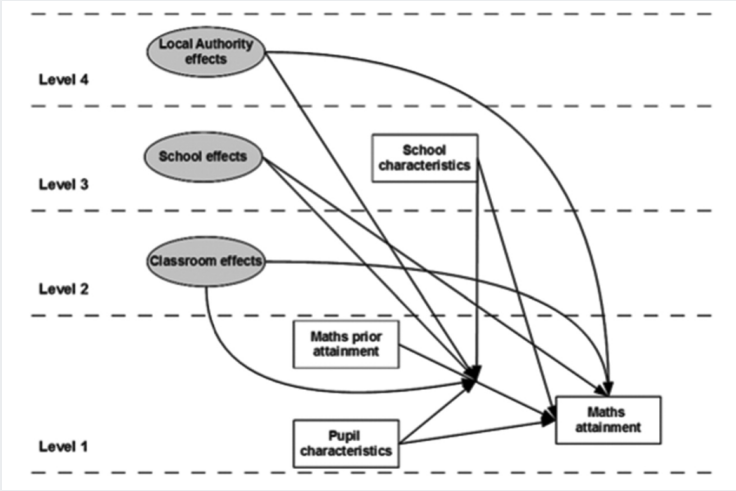
Data structures (4)

But we are not limited to 2 levels only:



- This is a 3-level structure of pupils nested within classrooms, nested within schools.

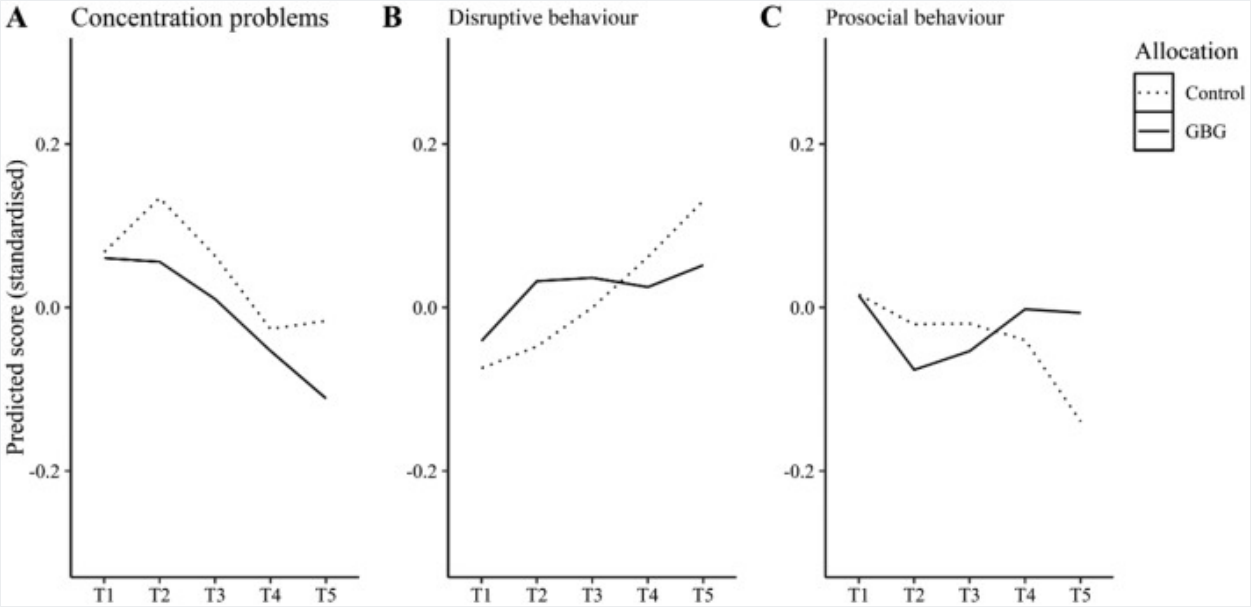
We can go even further, as it was done here:



- In this 4-level structure, schools are nested within a higher level of local authorities.

Data structures (5)

- We are not limited to one outcome either

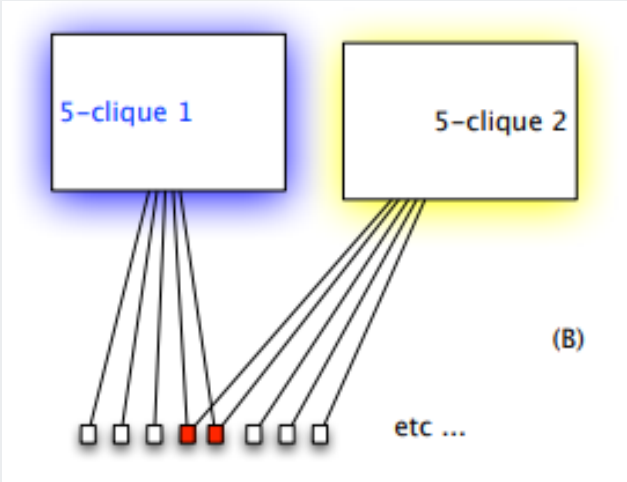


- The structure here is longitudinal and multivariate: measures of children's mental health over time

Image Source: Troncoso and Humphrey, 2021

Data structures (6)

- We are not limited to level-1 units (individuals) belonging to one level-2 unit (groups) only:



- The structure here is people nested in multiple "cliques": Multiple Membership and Multiple Classification (MMMC) Models and Multilevel Social Networks

Image Source: Tranmer, 2010

Data structures (7)

To sum up:

Hierarchical structures are generated by:

a) Data collection mechanism

- Survey data rarely comes from a simple random sample (SRS).
- Surveys often have multi-stage designs: clustered data.

b) ‘Natural’ structures within the population

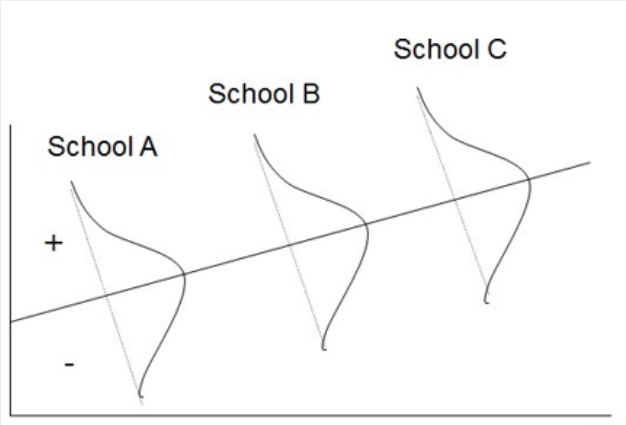
- Individuals are clustered according to geography, household, etc.
- Observation are therefore not independent.

Why is non-independence a problem?

If data has been collected using a two-stage design, carrying out an individual level analysis is equivalent to assuming it is a simple random sample.

But independence assumption is unrealistic; for example:

- we could expect a positive correlation between exam scores from pupils from the same school.



- The very fact that two pupils go to the same schools makes them have a "shared environment"

What is a level? (1)

A study samples 100 areas from the UK. Within each area a sample of citizens is selected and asked questions. Below are shown the first six rows of the resulting dataset.

Citizen	Voting intention	Area	Age	Gender	Ethnicity
511	Labour	19	64	Female	White
1204	Labour	36	36	Female	Asian
3389	Liberal Democrat	98	18	Male	White
2521	Conservative	72	72	Male	White
1193	Labour	35	49	Female	Mixed
262	Labour	11	28	Male	Other

Source: LEMMA course Bristol University

Question time:

- Which of the variables can be sensibly treated as random classifications and taken as levels of units?

What is a level? (2)

We are interested in assessing to what extent the difference between boys' and girls' educational achievement varies across secondary schools for 16 year old students.

Which of the following designs would be most effective?

1. Randomly sample 5 schools and take achievement scores for 100 boys and 100 girls aged 16 in each school.

2. Randomly sample 30 schools and within each school take a random sample of 10 boys and 10 girls aged 16 and take these children's achievement scores.

3. Sample 1000 schools and take 1 boy and 1 girl aged 16 from each school.

What is a level? (3)

Should a variable be treated as a level in a multilevel structure or as a categorical explanatory variable?

Consider a structure with students nested in schools with information on school type (state vs private). How should we include school type in our model?

1) As a random classification (i.e. level) if units can be regarded as a random sample from a wider population of units, for example schools.

- Interested in generalising to population of schools.

2) As a fixed classification (i.e. categorical variable) if small fixed number of categories. For example, if state and private schools were not two types sampled from a larger number of types.

- Not interested in generalising to a wider population of school types.

Schools can be thought of as a random classification but school type should be a fixed classification

Why does clustered data matter?

Standard analysis assumes independence and estimates standard errors of model parameters accordingly.

If observations within clusters are positively correlated this will underestimate standard errors.

- Result: variables may appear significant when in fact they are not.
- What to do then? Need to take account of clustering.

Why are standard errors too small if we ignore clustering?

Suppose we have 5000 individuals in 100 groups.

In a single level model standard errors calculated on the assumption that the sampled individuals provide 5000 independent pieces of information.

But when outcomes are clustered, the number of independent observations (the effective sample size - ESS) will be fewer than 5000 and standard errors from standard regression will be too small.

ESS depends on the amount of clustering. For example, if all individuals in a group have the same y value, ESS = 100.

Underestimation most severe for coefficients of level 2 variables.

What's the problem with standard regression (OLS)?

Technical

Assumption of independence of residuals will be invalid if there are dependencies between individuals in the same group (area, school etc). This will lead to underestimation of standard errors, and therefore p-values that are too small.

Substantive

We are often interested in estimating the amount of variation between groups, and the extent to which it can be explained by group-level explanatory variables.

Methods for hierarchical structures

Traditional approaches to analysing clustered data treat clustering as a nuisance that must be accounted for.

Parameters estimated in the usual way but standard error estimates are adjusted for impact of clustering.

Multilevel modelling takes account of hierarchical structure and regards structure of substantive interest.

General framework notation (1)

- Level one/microlevel unit: i (e.g., individual)
- Level two/macrolevel unit: j (e.g., area, school)
- There are $i = 1, 2, \ldots n_j$ level one units within level-2 units
- and $j = 1, 2, \ldots J$ level-2 units.
- Response variable y_{ij} (continuous) is a function of:
  - individual variables: $x_{0ij}, x_{1ij}, \ldots x_{pij}$
  - and contextual variables: $z_{1j}, z_{2j}, \ldots z_{Qj}$
- Error terms:
  - e_{ij} at the individual level
  - u_{0j} at level-2

Multilevel modelling software

- R packages: lme4, nlme, R2MLwiN, rstanarm, brms, MCMCglmm (and others)
- Stata: mixed (previously known as xtmixed)
- SAS: PROC MIXED
- SPSS: MIXED and VARCOMP commands
- MLwiN (specialist software)
- Mplus, Latent Gold and others!

Part Three:

Variance components and group-specific estimates

Single level regression for the mean (1)

Consider the simplest possible statistical model:

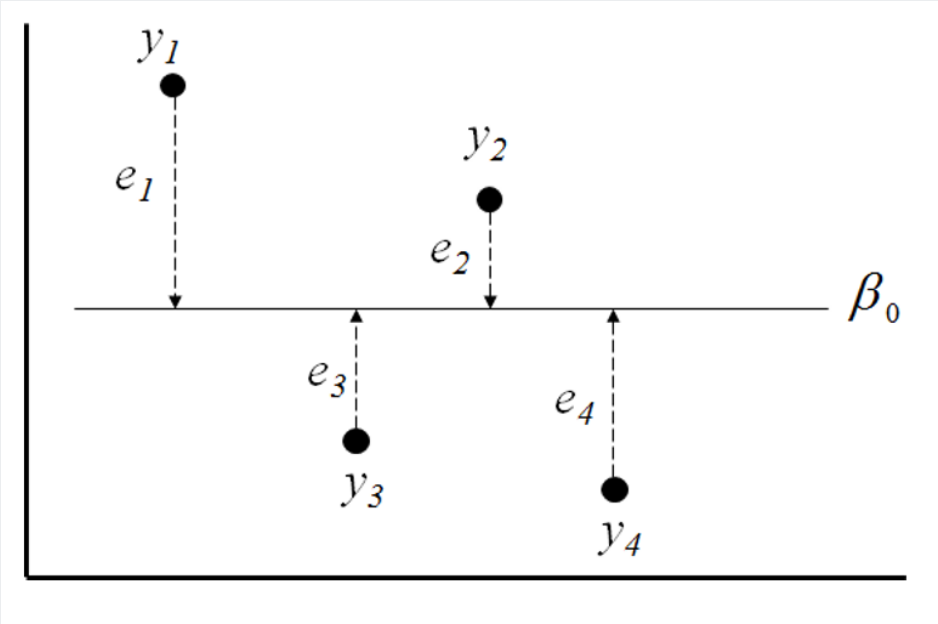
$$y_i = \beta_0 + e_i$$

where β_0 is the mean of y in the population, and e_i is the residual for the i-th individual $(i = 1, 2, \ldots n)$

- Usually assume $e_i \sim N(0, \sigma^2)$.
    - Normal distribution
- The variance σ^2 summarizes the variability around the mean;
    - if this is zero all the points would lie on the $y = \beta_0$ line

Single level regression for the mean (2)

These are the residuals for four data points in single model for the mean:



Multilevel model for group means (1)

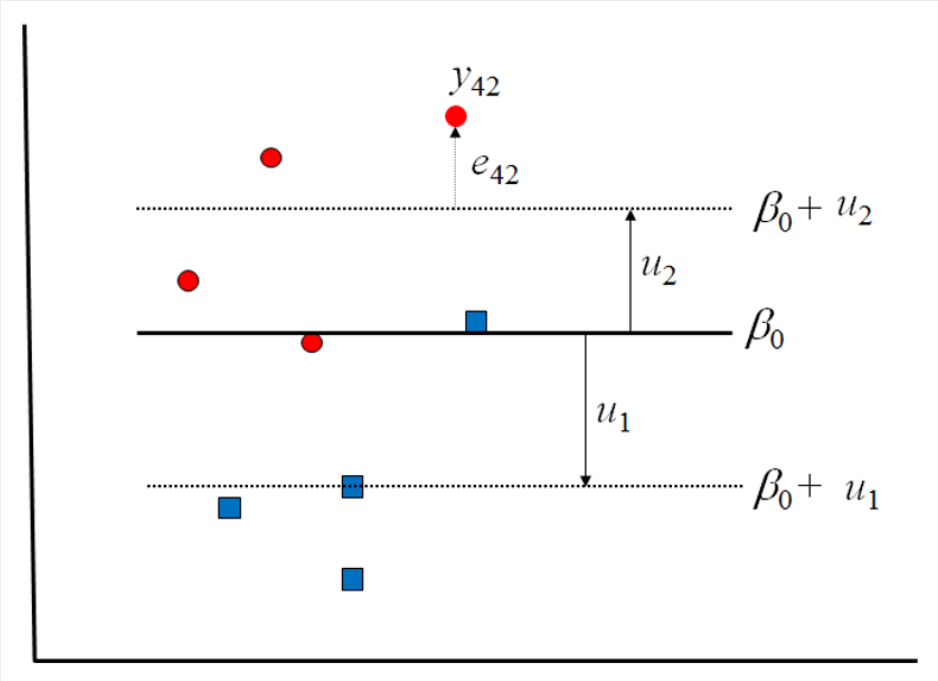
y_{ij} is the value of y for the i-th individual in the j-th group. A model that allows for (random) group effects is:

$$y_{ij} = \beta_0 + u_{0j} + e_{ij}$$

- β_0 is the overall mean of y (across all groups).
- $\beta_0 + u_j$ is the mean of y for group j.
- u_j is the difference between group j's mean and the overall mean.
- e_{ij} is the difference between the y-value for the i-th individual and that individual's group mean:
  - $e_{ij} = y_{ij} - (\beta_0 + u_j)$

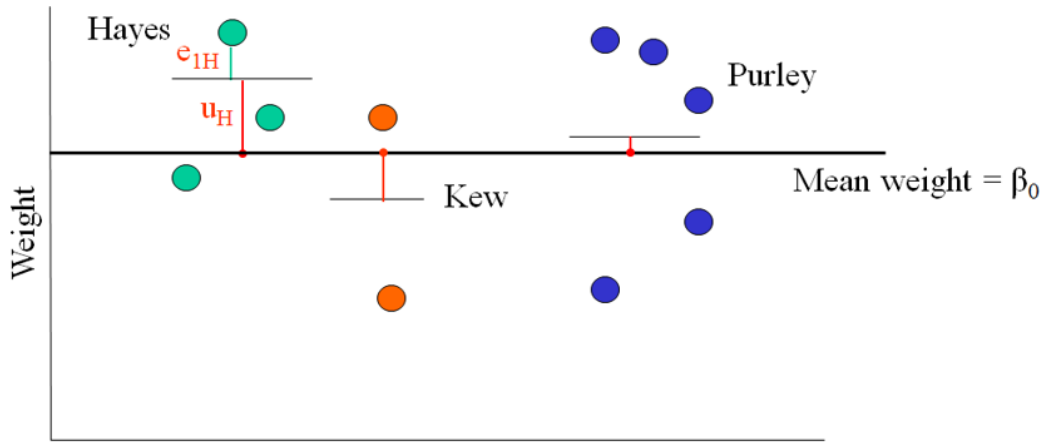
Multilevel model for group means (2)

Individual and group residuals in a two level model for the mean:



Multilevel model for group means (3)

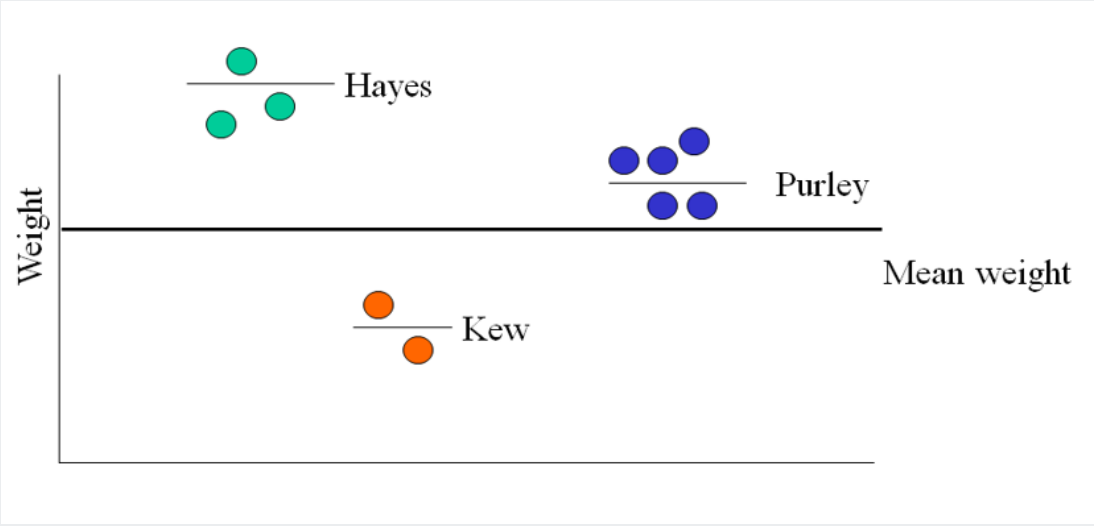
Example: Weight in small areas in London:



$$Weight_{ij} = \beta_0 + u_j + e_{ij}$$

Multilevel model for group means (4)

Example: Weight in small areas in London:



What's changed from the previous example?

Variance Partitioning (1)

Assume $u_j \sim N(0, \sigma_u^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$

σ_u^2 is the between group variance and σ_e^2 is the within group variance.

The variance partition coefficient is the proportion of total variance due to differences between groups.

$$VPC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

- VPC is also denoted as ρ (rho)
- Also known as the Intra-class correlation (ICC)
- $VPC = 0$ if no group effects
- $VPC = 1$ if no within-group differences

Variance Partitioning (2)

Example "tutorial" data: subset of a large dataset of examination results in London schools

- Empty MLM results:

```
Random effects:
 Groups   Name        Variance Std.Dev.
 school   (Intercept) 0.1686   0.4107
 Residual             0.8478   0.9207
Number of obs: 4059, groups:  school, 65

Fixed effects:
            Estimate Std. Error t value
(Intercept) -0.01317    0.05363  -0.246
```

VPC:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

$$\rho = \frac{0.169}{0.169 + 0.848}$$

$$\rho = 0.166$$

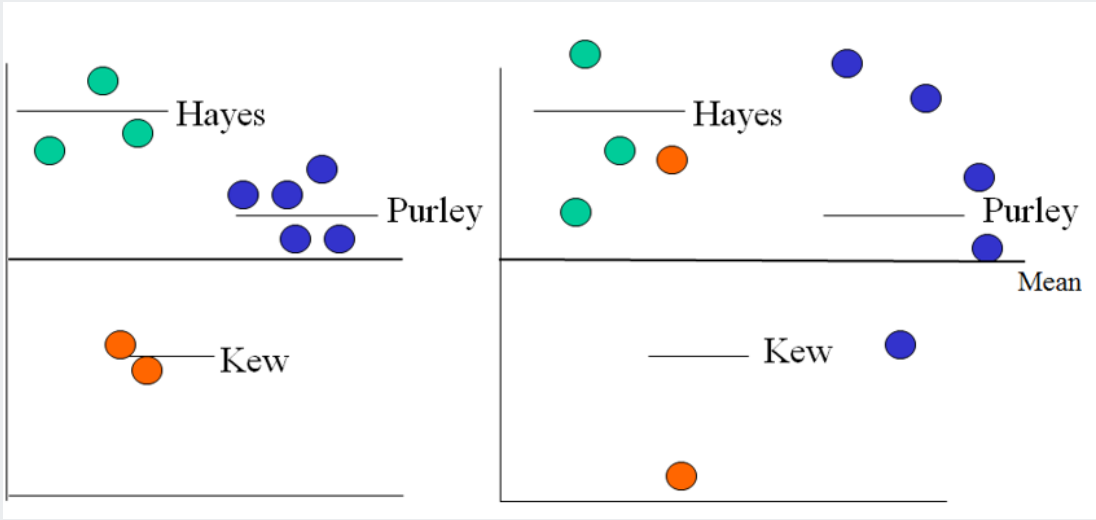
16.6% of the total variation in exam scores is due to differences between schools.

Assuming `R2MLwiN` and `lme4` are installed and loaded, this is to replicate the above example:

```
data(tutorial)
summary(mlm1 <- lmer(normexam ~ 1 + (1|school), data=tutorial, REML=F))
```

Variance Partitioning (3)

- Which model has the highest VPC?



- Answer: The figure on left-hand side has less within-group variation, hence VPC is larger

Testing for group effects (1)

- Test $H_0: \sigma_u^2 = 0$ by comparing single level and multilevel model in a likelihood ratio test.
- Use the deviance statistic:

$$D = 2(logL_1 - logL_2)$$

- Where:
  - L_1 is the likelihood of the single level model
  - L_2 is the likelihood of the multilevel model
- The test statistic D is compared with a χ^2 distribution with df = number of extra parameters in the more complex model
- Rejection of the null hypothesis implies “real group differences", in which case the multilevel model is preferred.

Testing for group effects (2)

Example "tutorial" data

- First we run a single-level model and extract the loglikelihood:

```
single <- lm(normexam ~ 1, data=tutorial)
(L1 <- logLik(single))
```

```
## 'log Lik.' -5754.682 (df=2)
```

- Then we run the multilevel model and extract the loglikelihood:

```
mlm1 <- lmer(normexam ~ 1 + (1|school), data=tutorial, REML=F)
(L2 <- logLik(mlm1))
```

```
## 'log Lik.' -5505.324 (df=3)
```

Apply the formula and compare to a χ^2 distribution with df being the number of extra parameters:

$$D = 2 * (L_2 - L_1)$$

Quick tip: for a difference of 1 parameter, χ^2 values over 3.84 are statistically significant

Practical 1

Part Four:

Accounting for individual and group characteristics: fixed effects

Random Intercepts model (1)

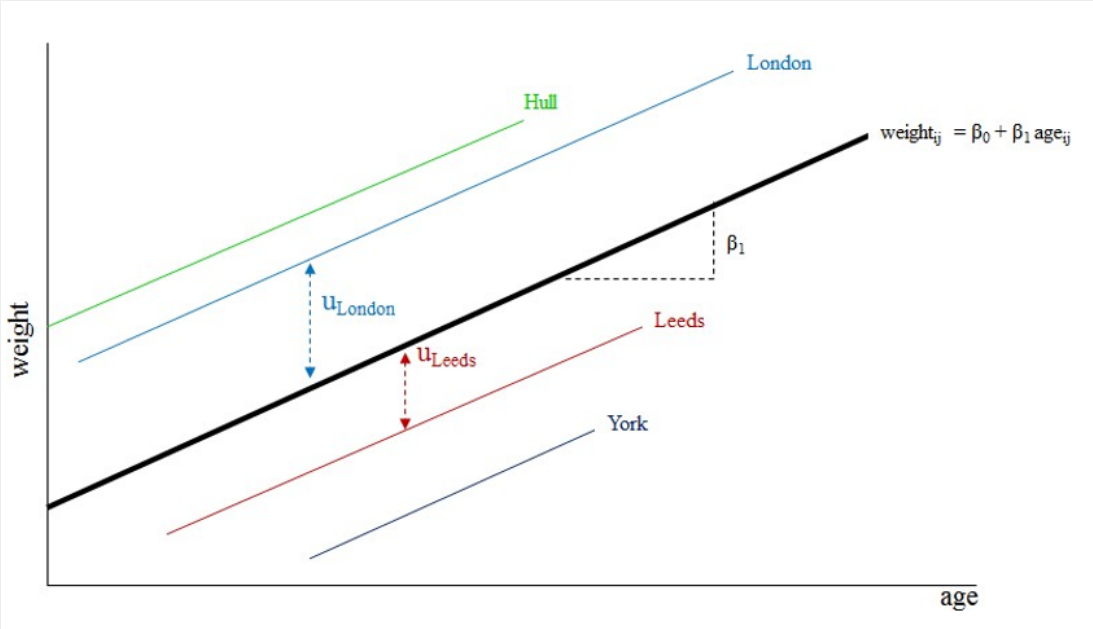
- We will now add a level 1 explanatory variable to the model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}$$

- The overall relationship between y and x is represented by a straight line with intercept β_0 and slope β_1
- There are two components in this model:
  - Fixed part: $\beta_0 + \beta_1 x_{ij}$
  - Random part: $u_j + e_{ij}$

Random intercepts model (2)

Same slope in all areas (same relationship between age and weight in all areas)



Random intercepts model (3)

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}$$

- Suppose now that x is a dichotomous predictor: only takes values 0 and 1
- β_0 is the overall mean of y for individuals with $x = 0$
- $\beta_0 + u_j$ is the mean for individuals with $x = 0$ in group j
- The slope β_1 is the difference in the mean for $x = 1$ relative to $x = 0$ (in any group)

Random intercepts model (4)

Example "tutorial" data:

```
Random effects:
 Groups   Name        Variance Std.Dev.
 school   (Intercept) 0.09213  0.3035
 Residual             0.56573  0.7522
Number of obs: 4059, groups:  school, 65

Fixed effects:
            Estimate Std. Error t value
(Intercept) 0.002391   0.040023    0.06
standlrt    0.563371   0.012465   45.20
```

- Coefficient for standlrt is 0.563
  - For each one-unit increase in standlrt we can expect a 0.563 standard deviations increase in normexam
- Is the coefficient statistically significant?
  - Where are the p-values?
  - Tip: t-values over 1.96 correspond to p-values below 0.05

To reproduce this example:

```
summary(mlm2 <- lmer(normexam ~ standlrt + (1|school), data=tutorial, REML=F))
```

Effect of a level-1 variable on the variance

- Adding a level 1 explanatory variable to the model will always reduce the level 1 variance and the total variance.
- However, the level 2 variance may stay the same, increase or decrease.
- This depends on the association between the level 1 explanatory variable and the level 2 outcomes.

Adding level 2 explanatory variables (1)

MLM allows us to explore the effects of group-level variables while simultaneously allowing for **unmeasured** group characteristics that influence the outcome.

- But it also allows for **measured** group characteristics:

When modelling individuals at level 1 and groups at level 2:

- level 2 variables are often called **contextual variables**
- and their effects on an individual's y-value are **contextual effects**.

Note: it is particularly important to use MLM to estimate contextual effects because their standard errors may be severely underestimated when a single-level model is used.

Adding level 2 explanatory variables (2)

Example "tutorial" data:

```
Random effects:
 Groups   Name        Variance Std.Dev.
 school   (Intercept) 0.0800   0.2828
 Residual             0.5658   0.7522
Number of obs: 4059, groups:  school, 65

Fixed effects:
                      Estimate Std. Error t value
(Intercept)           -0.08704    0.05112  -1.703
standlrt               0.56379    0.01246  45.258
factor(schgend)boysch  0.09688    0.10891   0.890
factor(schgend)girlsch 0.24511    0.08497   2.885
```

- The reference category for schgend is mixed school
- What type of school obtained better scores?

To reproduce this example:

```
summary(mlm3 <- lmer(normexam ~ standlrt + factor(schgend) +
        (1|school),
        data=tutorial, REML=F))
```

Contextual effects: Example research questions

- How do teacher characteristics (for example, the number of years in teaching or measures of their teaching style) affect student attainment?
    - Is a student's attainment affected by the ability of peers, and does any effect depend on a student's own ability?
- Is living in a deprived area associated with poorer health?
- Is this association independent of personal deprivation?
- What is the role of family background on child health?

Sources of contextual data

- Data referring to organisations may be collected routinely by government authorities,
    - Administrative data
    - GIS data
- Contextual data may also derive from level 1 data that is aggregated to form level 2 variables.
    - e.g. School averages
- Data may be collected at level 2, for example surveys in which key members of a group are interviewed.

Multilevel model with contextual effects

- Include a level 2 variable in exactly the same way as a level 1 variable.
- Suppose x_{ij} is defined at level 1 and x_{2j} at level 2. The random intercept model is:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{2j} + u_j + e_{ij}$$

Note: Sometimes z's may be used to denote level-2 explanatory variables, but you can't go wrong with x.

Practical 2

Part Five:

Multilevel modelling for binary responses

This session's goals

- Introduce the MLM for binary responses
- Provide examples of real-world research
- Examine statistical output
- Practice fitting a logistic MLM

When to use Multilevel models for binary response? (1)

Recidivism Study in Chile

Data: criminal history data of a cohort of individuals released from prison

Outcome: Whether a *person* got a new conviction or not

Levels: *Individuals* (level 1) nested within *prisons* (level 2)

Morales-Gomez, A. (2018). Individual and Structural Factors Affecting Recidivism: The Role of Prisoners, Prisons and Places in the Chilean Context

When to use Multilevel models for binary response? (2)

Controlled Delivery of Drug Parcels in Scotland

Data: Drug parcels seized by the UKBF en route to Scotland

Outcome: Whether a *parcel* was adopted for a controlled delivery or not

Levels: *drug parcels* (level 1) nested within *Local Authorities* (level 2)

Morales-Gómez, A., McVie, S. & Pantoja, F. (2022). Controlled Delivery of Illegal Drug Parcels in Scotland: Does Policing Practice Align With a Public Health Approach Focused on Drug-Related Harm?

Multilevel models for binary response

The multilevel linear model is generally appropriate when the outcome is continuous and normally distributed

Other types of data do not satisfy the assumption of normality: Count data, categorical data, ordinal data, etc.

Just as in the well-known single level models, we use logistic regression in multilevel modeling when the outcome variable is binary (Yes/No)

Examples

- Whether someone reoffends or not
- Whether someone was victim of a crime or not
- Whether someone intends to vote in the next elections

Logistic regression (Recap)

Assume we want to analyse employment patterns in a sample of people:

We have binary variable Y indicating employment status where:

- 1 = employed
- 0 = unemployed

We can't use linear regression as many of the assumption for linear regression are not satisfied

- Values are bounded by 0 and 1
- We can't assume normality

We model the probability that $y = 1$ (i.e. employed)

$$Pr(y_i = 1) = p_i$$

Logistic regression (Recap)

We use a logistic or **logit** transformation of the outcome to *link* the dependent variable to a set of explanatory variables

$$logit(p_i) = \log\left(\frac{p_i}{1 - p_i}\right)$$

We can write the model:

$$logit(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i$$

The logit link function keeps values in a range between 0 and 1

- This is necessary as probabilities must be between 0 and 1
- We can write the equation in terms of odds:

$$\frac{p_i}{1 - p_i} = \exp(\beta_0 + \beta_1 x_i)$$

Two level random intercept model for binary response

Using the same example, imagine that we now have an additional variable indicating local authorities where respondents live

Consider a two-level structure where n individuals i are nested within groups J

- Group can be any level-2 unit (school, cities, etc.)

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 x_{ij} + u_j$$

The left side is the nonlinear transformation (log odds)

the right side takes the form of a linear model

where

$$u_j \sim N(0, \sigma_u^2)$$

The group effects or level-2 residuals u_j are assumed to be independent and follow a normal distribution

Interpretation of main parameters

β_0 is the *overall intercept*

- β_0 is the log-odds that $y = 1$ when $x = 0$ and $u = 0$

β_1 is usually referred to as the *cluster-specific* effect

- β_1 is the effect of a 1-unit change in x in the log-odds that $y = 1$, while adjusting for group effect u

u_j is the *group random effect* or level two residual

- $\beta_0 + u_j$ is the intercept for a given group j
- $var(u_j) = \sigma^2_u$ is the between-group variance and represents the variability across groups

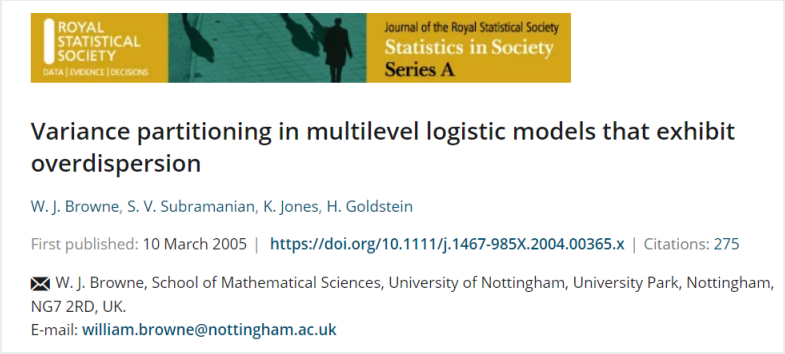
Variance Partition Coefficient (VPC)

Measures the proportion of the total variance that is due to differences between groups

There are several ways of defining VPC for binary data:

- Model linearisation
- Simulation
- Latent Variable approach

For more details: click here



VPC Latent Variable Approach

Recall VPC formula for continuous response:

$$VPC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

Assume that the underlying response variable is *continuous* but we can only observe a **binary response** indicating whether the underlying variable is greater or less than a given threshold.

This underlying continuous variable comes from a logistic distribution with a variance of $\frac{\pi^2}{3} \approx 3.29$

Replacing on the previuos formula:

$$VPC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2 *}$$

where $\sigma_e^2 * = \frac{\pi^2}{3}$

Example

Antenatal care in Bangladesh

We want to analyse whether a woman received antenatal care from a medically-trained provider at least once before her most recent live birth.

We also want to explore whether antenatal care varies across communities

Variables	Description	Level
comm	Community identifier	2
antemed	Received antenatal care at least once (1 = yes, 0 = no)	1
mage	Mother's age at the child's birth (in years)	1
urban	Type of region of residence (1 = urban, 0 = rural)	2

There are level 1 variable (individual level) and level 2 variables (area level).

- antemed: Outcome
- comm: Level 2 identifier

R output

```
fit <- glmer(antemed ~    # Outcome variable
             (1 | comm),  # level two specification
             family = binomial("logit"),  ##link function
             data = mydata)
```

Fixed part

```
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.14809    0.07178   2.063   0.0391 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation

The log-odds of receiving antenatal care in an ‘average’ community $(u_j = 0)$ is estimated as $\beta_0 = 0.148$.

R Output (2)

Random part

```
Random effects:
 Groups Name        Variance Std.Dev.
(comm)   (Intercept) 1.464    1.21
Number of obs: 5366, groups:  comm, 361
```

Interpretation

The intercept for community j is $0.148 + u_j$, where the variance of u_j is estimated as $\sigma_u^2 = 1.464$

Adding explanatory variables (1)

We add maternal age (level 1 predictor)

```
fit2 <- glmer(antemed ~ magec +
                (1 | comm),
              family = binomial("logit"),
              data = mydata)
```

```
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.144604   0.071781   2.015    0.044 *
magec       -0.032357   0.005235  -6.181 6.37e-10 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Adding explanatory variables (2)

```
fit3 <- glmer(antemed ~ magec + urban +
                (1 | comm),
              family = binomial("logit"),
              data = mydata)
```

We add Urban (level 2 predictor)

```
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.346247   0.074256  -4.663 3.12e-06 ***
magec       -0.032481   0.005221  -6.221 4.94e-10 ***
urban        1.494405   0.132870  11.247  < 2e-16 ***
```

A quick note on estimation of logistic MLM

- The models presented here are all fitted with frequentist methods
    - But...
- Researchers have previously found that Bayesian estimation performs better at estimating random effects with binary responses
- For some arguably dense details, see Browne and Draper, 2002

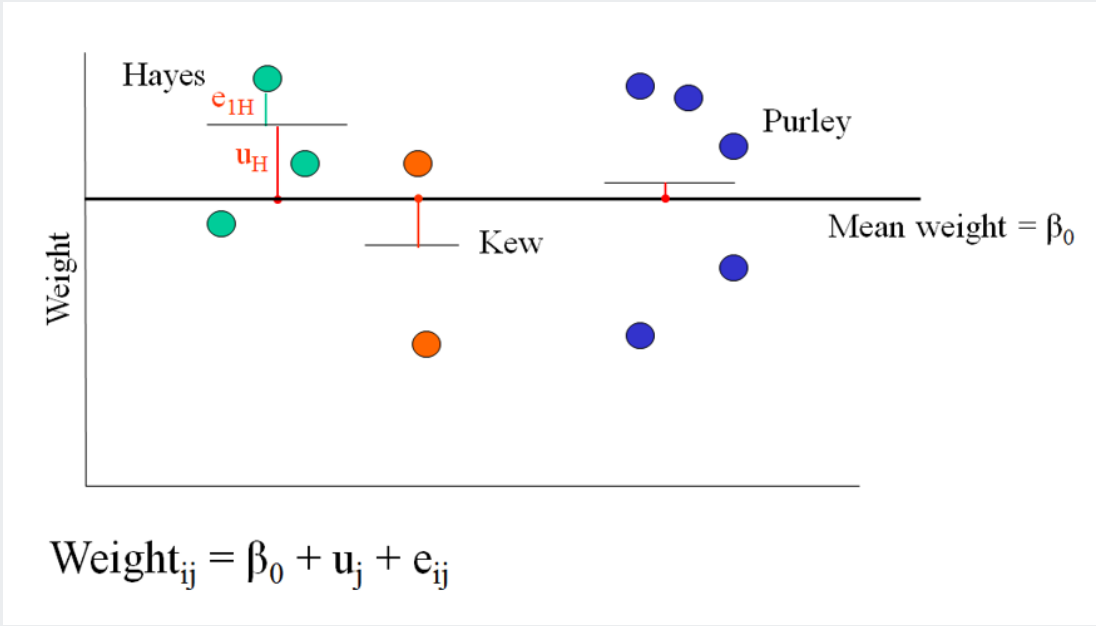
Practical 3

Part Six:

Differential processes between groups: random effects

Quick revision (1)

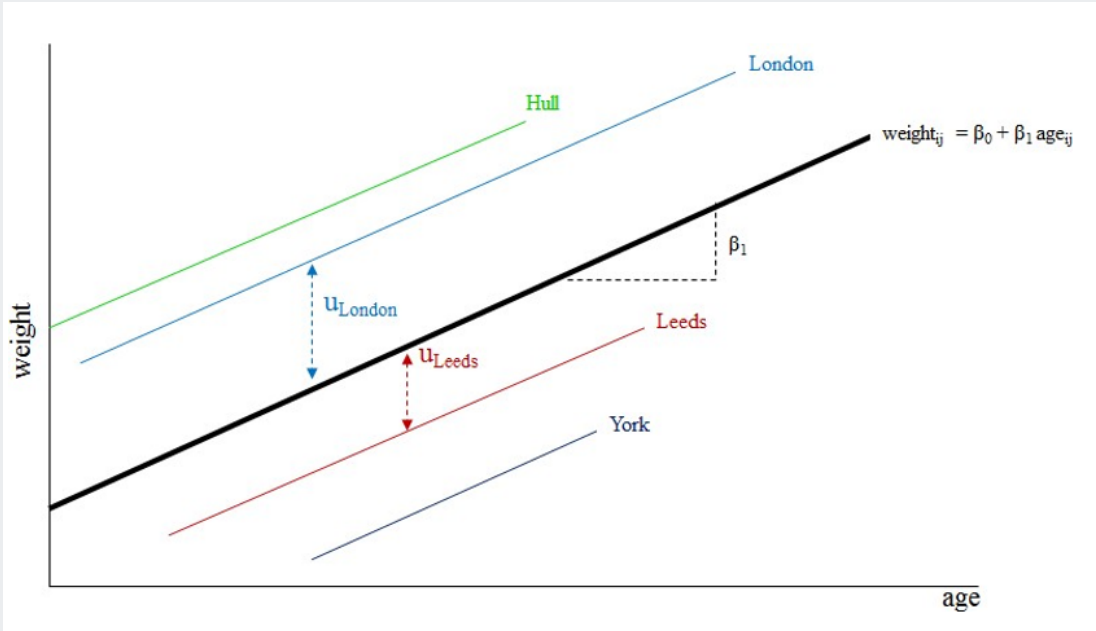
Variance components



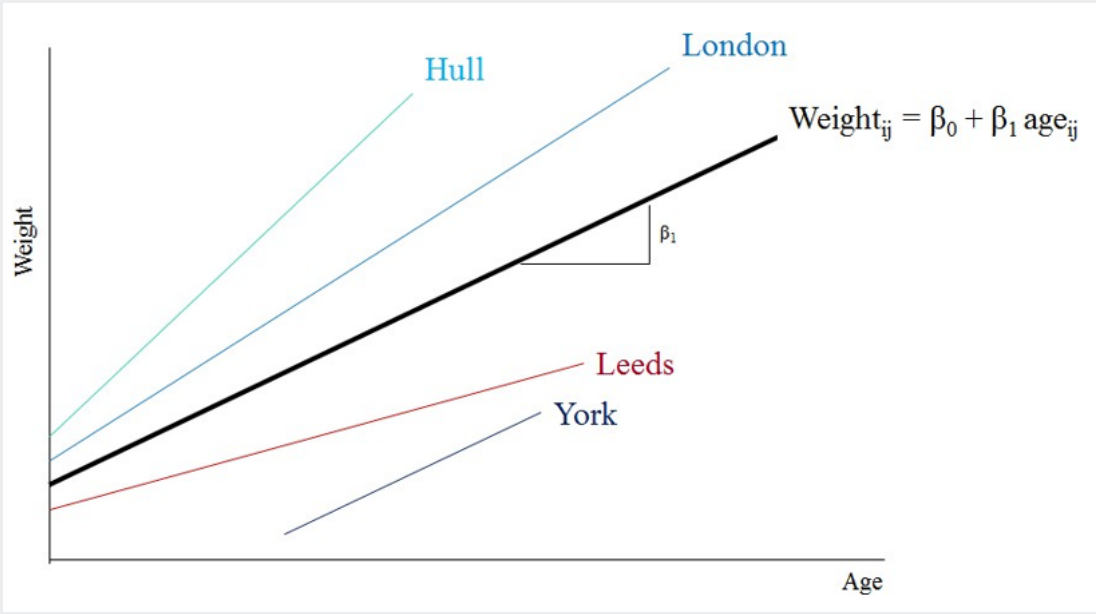
$$Weight_{ij} = \beta_0 + u_j + e_{ij}$$

Quick revision (2)

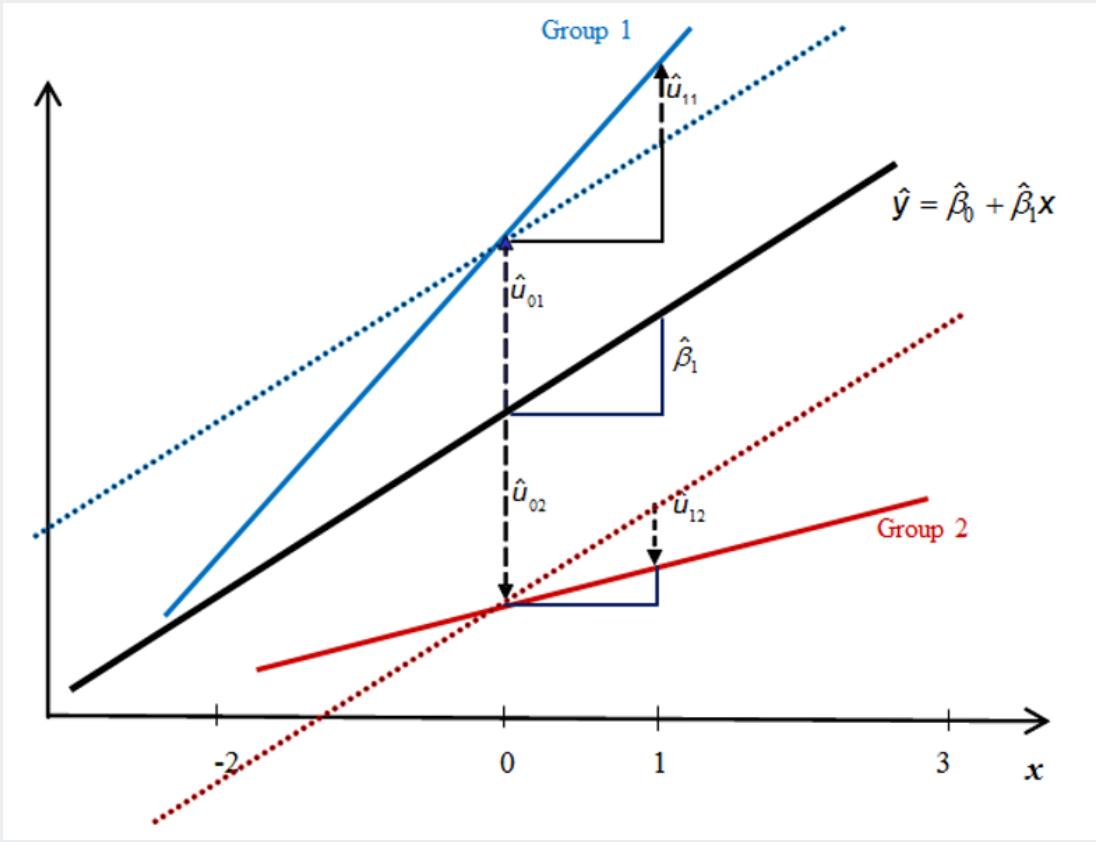
Random intercepts



Random slopes model (1)



Random slopes model (2)



Random slopes model (3)

- Allow both intercept and slope to vary randomly across groups:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0j} + u_{1j} x_{ij} + e_{ij}$$

- Comparing with the random intercepts model, a new term $u_{1j} x_{ij}$ has been added and we have two random effects: u_{0j} and u_{1j}
- β_1 is the slope of the average regression
- $\beta_1 + u_{1j}$ is the slope of the line for group j
- Assume u_{0j} and u_{1j} follow a bivariate normal distribution with variances σ^2_{u0} and σ^2_{u1}, and covariance σ_{u01}

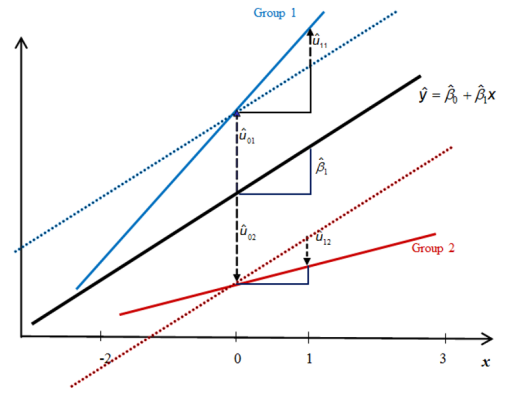
$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} MVN \sim \begin{pmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix}$$

Note: Every time we add one random slope to our model, we have 2 additional coefficients.

Interpretation of intercept-slope covariance

- Positive values of σ_{u01} imply that groups with high intercept residuals u_{0j} tend to have slope residuals u_{1j}
- When $\sigma_{u01} > 0$ groups with high intercepts (high $\beta_0 + u_{0j}$) tend to have steeper than average slopes (high $\beta_1 + u_{1j}$). Groups with low intercepts have flatter than average slopes.
    - This will lead to a “fanning out” of the group prediction lines.
- When $\sigma_{u01} < 0$, there is a “fanning in” pattern of group lines.

Interpretation of differential slopes



- Suppose this is relationship between attainment at 16 (y) and attainment on entry to the school (x).
- Across range of x, school 1 is more effective than school 2.
- But difference between the two schools widens as x increases.
- So choice of school is particularly important among children with high prior attainment.
- Although school 2 is less effective than school 1, it's flatter line means that school 2 has decreased differences in the outcomes of children with different prior attainments.
- Prior attainment is more predictive of subsequent performance in school 1.

Random slopes: example

Example "tutorial" data:

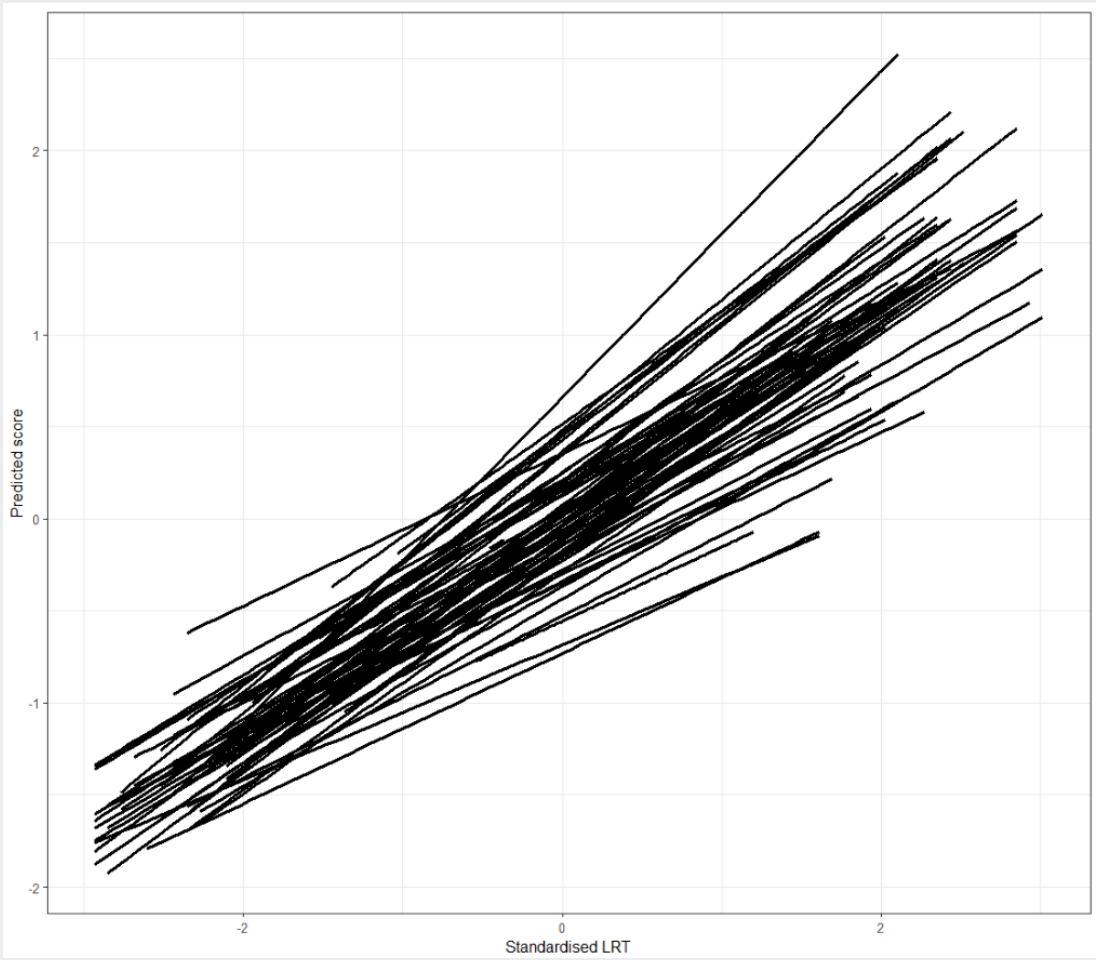
```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 school   (Intercept) 0.07842  0.2800
          standlrt    0.01453  0.1205   0.60
 Residual             0.55382  0.7442
Number of obs: 4059, groups:  school, 65

Fixed effects:
                     Estimate Std. Error t value
(Intercept)           -0.10997    0.04851  -2.267
standlrt               0.55845    0.01990  28.068
factor(schgend)boysch  0.10438    0.09720   1.074
factor(schgend)girlsch 0.26756    0.07591   3.525
```

- What is pattern of predicted school lines?
- What is the correlation between slope and intercept?

To reproduce this example:

```
summary(mlm4 <- lmer(normexam ~ standlrt + factor(schgend) +
                (1 + standlrt|school), data=tutorial, REML=F))
```

School predicted lines



Is the addition of a slope meaningful?

Example "tutorial" data

- First we run a random intercepts model:

```
mlm3 <- lmer(normexam ~ standlrt
    (1|school),
    data=tutorial, REML=F)
```

- Then we run the multilevel model with the random slope:

```
mlm4 <- lmer(normexam ~ standlrt
    (1 + standlrt|school),
    data=tutorial, REML=F)
```

- Then we compare using the **anova** function

```
anova(mlm3, mlm4)
```

```
Data: tutorial
Models:
mlm3: normexam ~ standlrt + factor(schgend) + (1 | school)
mlm4: normexam ~ standlrt + factor(schgend) + (1 + standlrt | school)
     npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
mlm3    6 9361.4 9399.3 -4674.7   9349.4
mlm4    8 9321.8 9372.2 -4652.9   9305.8 43.643  2  3.335e-10 ***
```

- The df for this test is 2 and the χ^2 value is 43.643
- p-value is much smaller than 0.001

Centring and Standardising

- An explanatory variable with a mean of zero is achieved by subtracting the sample mean of x from the raw values,
  - $x_i - \bar{x}$
  - This type of centring is sometimes called "grand mean centring".
  - After centring, the intercept is interpreted as the predicted mean of y at $\bar{x}$
- Another way of achieving a mean of zero is by *standardising*
  - $\frac{x_i - \bar{x}}{s_x}$
  - You can do the same with y and all continuous x's
  - In this case, coefficients will be *standardised*, which means that interpretation is in units of standard deviations
- Why is this important?
  - σ^2_{u0} is the variance at $x = 0$, which may be outside the range of x

Practical 4

Wrapping up

What hasn't been covered today?

- There is a *plethora* of applications and model specifications we haven't covered. Just to name a few:
    - Longitudinal models
    - Bayesian models (all we've practiced today has been from a frequentist approach)
    - Non-hierarchical models: cross-classified, multiple membership
    - Spatial models
    - Multilevel social networks

MLMs are everywhere. Actually, once you see multilevel structures in your data, you cannot unsee them...

-'I see MLMs!'

So where to go next?

- LEMMA (University of Bristol): Free online course. Click here
    - This course has tutorials in R, Stata and MLwiN
- Tutorials in R by Rens van de Schoot. Click here and here
- An online tutorial covering this from a Bayesian perspective using R. Click here
- Online resources by UCLA's Statistical Methods and Data Analytics: Click here and here
    - These are MLM textbook examples using R, Stata, MLwiN, Mplus and others.

Some applications in social sciences (1)

Education:

- Prior, L., Goldstein, H. & Leckie, G. (2021). School value-added models for multivariate academic and non-academic outcomes: exploring implications for performance monitoring and accountability
- Troncoso, P. (2019). A two-fold indicator of school performance and the cost of ignoring it
- Troncoso, P., Pampaka, M., Olsen, W. (2016). Beyond traditional school value-added models: a multilevel analysis of complex school effects in Chile
- O'Hanlon, F., Paterson, L. & McLeod, W. (2013). The attainment of pupils in Gaelic-medium primary education in Scotland

Some applications in social sciences (2)

Criminology:

- Morales-Gómez, A., McVie, S. & Pantoja, F. (2022). Controlled Delivery of Illegal Drug Parcels in Scotland: Does Policing Practice Align With a Public Health Approach Focused on Drug-Related Harm?
- Ben Matthews, Ben Collier, Susan McVie, S. (2021). Understanding digital drug markets through the geography of postal drug deliveries in Scotland
- Morales-Gomez, A. (2018). Individual and Structural Factors Affecting Recidivism: The Role of Prisoners, Prisons and Places in the Chilean Context. (Prison effects: pp. 106-157). (Area effects: pp. 158-201)
- Pina-Sánchez, J., Linacre, R. (2013). Sentence Consistency in England and Wales: Evidence from the Crown Court Sentencing Survey

For more applications:

You could visit the University of Bristol's Gallery of Multilevel Papers

General multilevel modelling books

- Goldstein, H. (2011). Multilevel statistical models (4th ed.). John Wiley and Sons
- Hox, J., Moerbeek, M., van de Schoot, R. (2017). Multilevel Analysis: Techniques and Applications (3rd Ed). Routledge
- Snijders, T., Bosker, R. (2012). Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling (2nd ed.). Sage

Thank you!