Multilevel Data Analysis

AQ801-3-M & Version 1.1

The Random Intercept Model Part B

Contents & Structure

- Variable intercepts: fixed or random parameters? When to use random coefficient models
- Definition of the random intercept model
- More explanatory variables
- Within-and between-group regressions
- Parameter estimation

Recap:

- What is the main assumption under the Random Intercept Model?

Learning Outcomes

- **At the end of this topic, You should be able to:**
  - Estimate and interpret random intercept model

Key Terms You Must Be Able To Use

- **If you have mastered this topic, you should be able to use the following terms correctly in your assignments and exams:**
- ***Fixed effects***
- ***Random effects***
- ***Residuals***
- ***Fixed effect model***
- ***Ordinary least square (OLS)***
- ***Mixed model***
- ***Random Intercept model***
- ***Hierarchical linear model***
- ***Random intercept***
- ***Empty model***
- ***Intraclass correlation coefficient***

- *Within- and between-group regression coefficients.*
- *Cross-level interaction effect*
- *Parameter estimation*
- *Posterior confidence intervals*
- *Comparative standard errors, or posterior standard deviation*
- *Diagnostic standard error*

Learning Outcome 1

- Estimate and interpret random intercept model

The Empty Model

- Although this topic follows an approach along the lines of regression analysis, the simplest case of the HLM is the random effects analysis of variances (ANOVA) model, in which the explanatory variables X and Z, do not figure.
- This model only contains random groups and random variation within groups.

- It can be expressed as a model – the same model encountered before in formula (3.1) – where the dependent variable is the sum of a general mean, γ_{00}, a random effect at the group level, U_{0j}, and a random effect at the individual level, R_{ij}:
- $Y_{ij} = \gamma_{00} + U_{0j} + R_{ij}$ -----(4.6)

- Groups with a high value of U_{0j} tend to have, on average, high responses whereas groups with a low value of U_{0j} tend to have, on average, low responses.
- The random variables U_{0j} and R_{ij} are assumed to have a mean of 0 (the mean of Y_{ij} is already represented by γ_{00}), to be mutually independent, and to have variances $var(R_{ij}) = \sigma^2$ and $var(U_{0j}) = \tau_0^2$

- In the context of multilevel modeling (4.6) is called the empty model, because it contains **not** a single explanatory variable.
- It is important because it provides the basic partition of the variability in the data between the two levels.
- Given model (4.6), the total variance of Y can be decomposed as the sum of the level-two and level-one variances,

$$\text{var}(Y_{ij}) = \text{var}(U_{0j}) + \text{var}(R_{ij}) = \tau_0^2 + \sigma^2$$

Example: Empty model for language scores in elementary schools

- In this example, a data set is used that will recur in examples in next few subsequent topics.
- The data set is concerned with grade 8 students (age about 11 years) in elementary schools in the Netherlands.
- After deleting 258 students with missing values, the number of students is M=3758, and the number of schools is N=211.

- Class sizes in the original data set range from 5 to 36.
- In the data set reduced by deleting cases with missing data, the class sizes range from 4 to 34.
- One class per school is included, so the class and the school level are the same in this data set.

- The dependent variable is the score on a language test.
- Most of the analyses of this data set in this module are concerned with investigating how the language test score depends on the pupil's intelligence and his or her family's SES, and on related class variables.
- Fitting the empty model yields the parameter estimates presented in Table 4.1 in the next slide.

Coding to obtain Table 4.1 Using SAS University Ed./ Studio

- *options validvarname=any;*
- *libname mdadat XLSX '/folders/myfolders/MDA/MDA.XLSX';*
- *proc mixed data=mdadat.sheet1;*
- *class schoolnr;*
- *model lang_post= /solution;*
- *random intercept/sub=schoolnr;*
- *run;*

Table 4.1: Estimates for empty model.

Fixed effect	Coefficient	S.E.
γ_{00} = Intercept	41.00	0.32
Random part	Variance component	S.E.
Level-two variance:		
$\tau_0^2 = \text{var}(U_{0j})$	18.12	2.16
Level-one variance:		
$\sigma^2 = \text{var}(R_{ij})$	62.85	1.49
Deviance	26,595.3	

- The estimates $\hat{\sigma}^2 = 62.85$, and $\hat{\tau}_0^2 = 18.12$ yield an intraclass correlation coefficient of $\hat{\rho}_I = \frac{18.12}{80.97} = 0.22$.
- This is on the high side but not unusual, compared to other results in educational research (values between 0.10 and 0.25 are common)

- For the overall distribution of the language scores, these estimates provide a mean of 41.00 and a standard deviation of $\sqrt{18.12 + 62.85} = 9.00$
- The mean of 41.00 should be interpreted as the expected value of the language score for a random pupil in a randomly drawn class.

One explanatory variable

- The following step is the inclusion of explanatory variables.
- These are used to try to explain part of the variability of Y; this refers to variability at level two as well as level one.
- With just one explanatory variable X, model (4.5) is obtained (repeated here for convenience):
- $Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + U_{0j} + R_{ij}$

- The population variance of the lower-level residuals R_{ij} is assumed to be constant across the groups, and is again denoted by σ^2; the population variance of the higher-level residual U_{0j} is denoted by τ_0^2
- Thus, model (4.5) has four parameters: the regression coefficients γ_{00} and γ_{10} and the variance components σ^2 and τ_0^2

- The random variables U_{0j} can be regarded as residuals at the group level, or group effects that are left unexplained by X.
- The fixed intercept γ_{00} is the intercept for the average group.
- The residual variance (i.e. the variance conditional on the value of X) is

$$\text{var}(Y_{ij} \mid x_{ij}) = \text{var}(U_{0j}) + \text{var}(R_{ij}) = \tau_0^2 + \sigma^2$$

Example 4.2: Random intercept and one explanatory variable: IQ

- As a variable at the pupil level that is essential for explaining language score, we use the measure for verbal IQ taken from the ISI test (Snijders and Welten, 1968).
- The IQ score has been centered (standardized), so that its mean is zero.

- This facilitates interpretation of various parameters.
- Its standard deviation in this data set is 2.04 (this is calculated as a descriptive statistic, without taking the grouping into account)
- The result are presented in Table 4.2 in next slide.
- In the model presented in Table 4.2 each class, indexed by the letter j, has its own regression line given by
- $Y = 41.06 + U_{oj} + 2.507IQ$

Coding to obtain Table 4.2

- options validvarname=any;
- libname mdadat XLSX '/folders/myfolders/MDA/MDA.XLSX';
- proc mixed data=mdadat.sheet1;
- class schoolnr;
- model lang_post = IQ_Verb / solution ddfm=bw;
- random intercept/sub=schoolnr;
- run;

Table 4.2

Table 4.2: Estimates for random intercept model with effect for IQ.

Fixed effect	Coefficient	S.E.
γ_{00} = Intercept	41.06	0.24
γ_{10} = Coefficient of IQ	2.507	0.054
Random part	Variance component	S.E.
Level-two variance:		
$\tau_0^2 = \text{var}(U_{0j})$	9.85	1.21
Level-one variance:		
$\sigma^2 = \text{var}(R_{ij})$	40.47	0.96
Deviance	24,912.2	

- The U_{0j} are class-dependent deviations of the intercept and have a mean of 0 and a variance of 9.85 (hence, a standard deviation of $\sqrt{9.85} = 3.14$)
- Figure 4.2 depicts 15 such random regression lines.
- This figure can be regarded as a random sample from the population of schools defined by Table 4.2
- Within group variance is 40.47 and therefore a standard deviation of $\sqrt{40.47} = 6.36$.

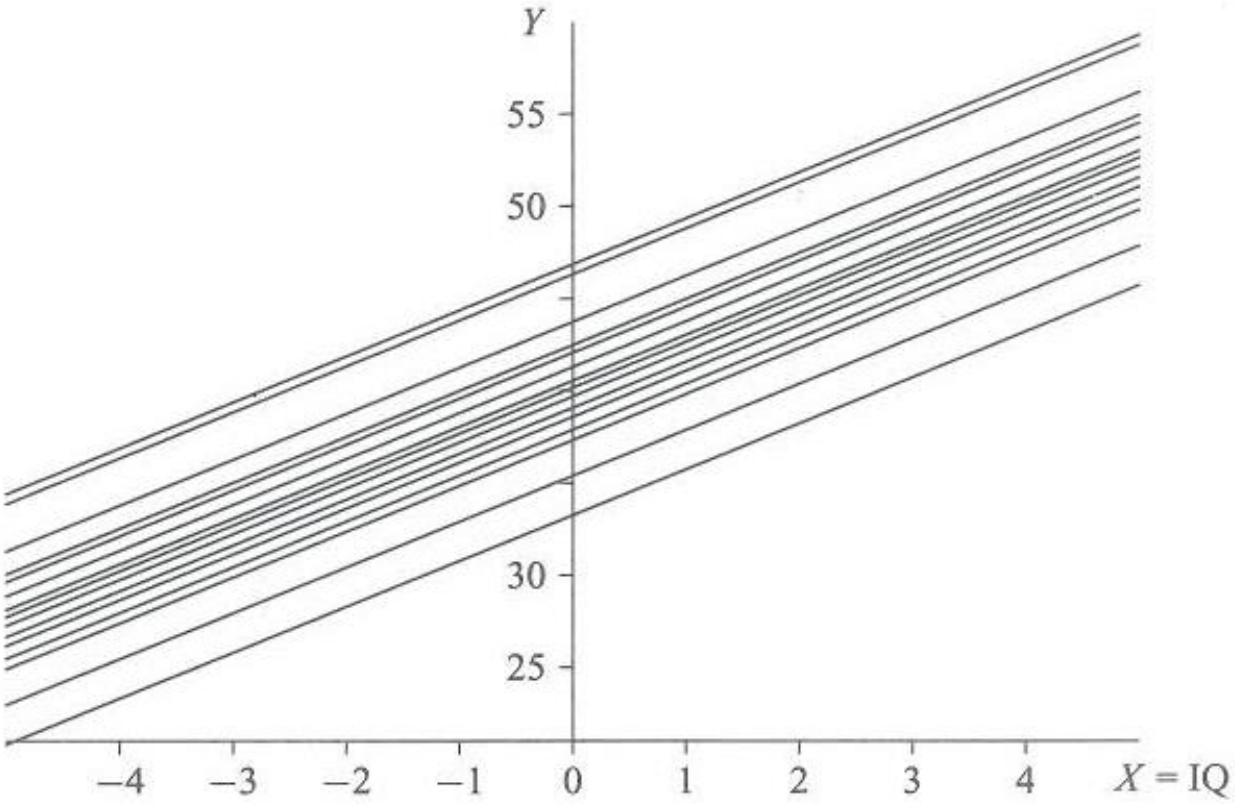

Figure 4.2: Fifteen randomly chosen regression lines according to the random intercept model of Table 4.2.

- A school with a typical low average achievement (bottom 2.5%) will have a value of U_{oj} of about two standard deviations below the expected value of U_{0j}, so that it will have a regression line
- **Y = 41.06 – 2 x 3.14 + 2.507IQ = 34.78 + 2.507IQ**

- Whereas a school with a typical high achievement (top 2.5%) will have a regression line
- **Y = 41.06 + 2 x 3.14 + 2.507IQ**

  = 47.34 + 2.507IQ
- There appears to be a strong effect of IQ.
- Each additional measurement unit of IQ leads, on average, to 2.507 additional measurement units of the language score

- To obtain a scale for effect that is independent of the measurement units, one can calculate standardized coefficients, that is, coefficients expressed in standard deviations as scale units.
- These are the coefficients that would be obtained if all variables were rescaled to unit variances.

- They are given by
- $\frac{S.D.(X)}{S.D.(Y)}\gamma$
- In this case estimated by $\frac{2.04}{9.00}(2.507) = 0.57$
- In other words, each additional standard deviation on IQ leads, on average, to an increase in language score 0.57 standard deviations.

- The residual intraclass correlation is estimated by
- $\hat{\rho}_I(Y|X) = \frac{9.85}{40.47+9.85} = 0.20$
- slightly smaller than the raw intraclass correlation of 0.22 (see Table 4.1)

Within-and between-group regressions

- Group means are an especially important type of explanatory variable.
- A group mean for a given level-one explanatory variable is defined as the mean over all individuals, or level-one units, within the given group, or level-two unit.
- This can be an important contextual variable.

Example 4.3: Within- and between-group regressions for IQ

- We continue Example 4.2 by allowing differences between the within-group and between-group regressions of the language score on IQ.
- The results are displayed in Table 4.4 in the next slide.
- IQ here is the variable with overall centering but no group centering.
- Thus, the results refer to model (4.9)

Table 4.4: Estimates for random intercept model with different within- and between-group regressions.

Fixed effect	Coefficient	S.E.
γ_{00} = Intercept	41.11	0.23
γ_{10} = Coefficient of IQ	2.454	0.055
γ_{01} = Coefficient of $\overline{\text{IQ}}$ (group mean)	1.312	0.262
Random part	Variance component	S.E.
Level-two variance:		
$\tau_0^2 = \text{var}(U_{0j})$	8.68	1.10
Level-one variance:		
$\sigma^2 = \text{var}(R_{ij})$	40.43	0.96
Deviance	24,888.0	

- The within-group regression coefficient is 2.454 and the between-group regression coefficient is 2.454+1.312=3.766.
- A pupil with a given IQ obtains, on average, a higher language test score if he or she is in a class with a higher average IQ.

- Table 4.4 represents within each class, denoted j, a linear regression equation
- $Y = 41.11 + U_{oj} + 2.454IQ + 1.312\overline{IQ}$
- Where U_{oj} is a class-dependent deviation with mean 0 and variance 8.68 (standard deviation 2.95).
- The within-class deviations about this regression equation, R_{ij}, have a variance of 40.43 (standard deviation 6.36).

- Within each class, the effect (regression coefficient) of IQ is 2.454
- The within-group and between-group regression coefficients would be equal if, in formula (4.9), the coefficient of average IQ were 0 (i.e. $\gamma_{01} = 0$).

- This null hypothesis can be tested by the t-ratio defined as
- $t = \frac{estimate}{standard\ error}$
- Given here by $\frac{1.312}{0.262} = 5.01$, a highly significant result.
- In other words, we may conclude that the within- and between-group regression coefficients are indeed different.

Summary of Main Teaching Points

- It is conceive of the unexplained variation within groups and that between groups as random variability.
- Additional effects of the nesting structure can be represented by letting the regression coefficients vary from group to group

Question and Answer Session

Q & A

What we will cover next

- **Hierarchical Linear Model**