

EDUC 231D

Advanced Quantitative Methods: Multilevel Analysis

Winter 2025

Introduction to Multilevel Models and Random Effects

Lecture 3 Presentation Slides

January 14, 2025

Today's Topics

- One-way ANOVA models with random effects – balanced case
- Understanding between-group variance
- Introduction to R for multilevel modeling

One-way ANOVA Model: Balanced Case

Motivating example

- What was the average math score for Grade 8 students in the United States in 2019? How much did math scores differ across schools?
- Use TIMSS data from a sample of 20 U.S. schools
 - For this example, pretend 30 students were sampled from each school
 - So the data include 600 students (20 schools x 30 students each)

Naïve approach: overall average math score

- Option 1: calculate mean across all students

$$\frac{\sum_{i=1}^N Y_{ij}}{N} = 538.98, \text{ where } N = 600$$

- Option 2: calculate mean math score in each school ($\bar{Y}_{.j}$), then calculate average of the $\bar{Y}_{.j}$'s across the 20 schools:

$$\frac{\sum_{j=1}^J \bar{Y}_{.j}}{J} = 538.98, \text{ where } J = 20$$

idschool	j	schnb	$\bar{Y}_{.j}$ schmath_j
5007	1	30	593.42
5010	2	30	611.54
5022	3	30	593.80
5036	4	30	481.27
5059	5	30	556.81
5066	6	30	535.31
5067	7	30	555.26
5080	8	30	619.64
5101	9	30	395.52
5110	10	30	623.74
5112	11	30	370.68
5133	12	30	467.30
5139	13	30	587.14
5144	14	30	508.29
5151	15	30	515.44
5164	16	30	620.04
5205	17	30	524.77
5236	18	30	503.73
5237	19	30	498.66
5277	20	30	617.22

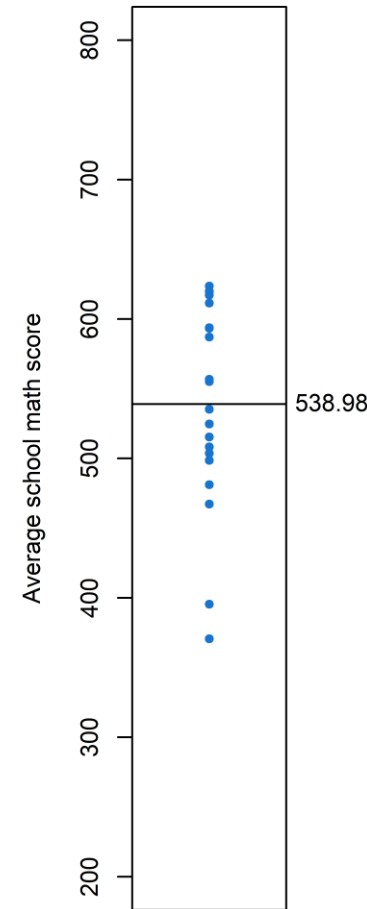
Naïve approach: variation across schools

- Calculate variance of the $\bar{Y}_{.j}$'s across the 20 schools:

$$\frac{\sum_{j=1}^J (\bar{Y}_{.j} - \bar{Y}_{..})^2}{J - 1} = 5,393.40,$$

where $J = 20$ and $\bar{Y}_{..} = 538.98$

- This variance calculation will overestimate the variance if the $\bar{Y}_{.j}$'s are measured with error → More on this later!



Model-based approach: Level 1

- Model the data at two levels: students nested within schools
- Level-1 (within-school) model:

$$Y_{ij} = \beta_{0j} + r_{ij}, r_{ij} \sim N(0, \sigma^2)$$

- j is used to index schools: $j = 1, \dots, J$ ($J = 20$ in our case)
- i is used to index students within each school: $i = 1, \dots, n_j$ (where n is the number of students in school j)

Model-based approach: Level 1

- Level-1 (within-school) model:

Y_{ij} is the 8th grade math score for student i in school j

σ^2 is the within-school variance component

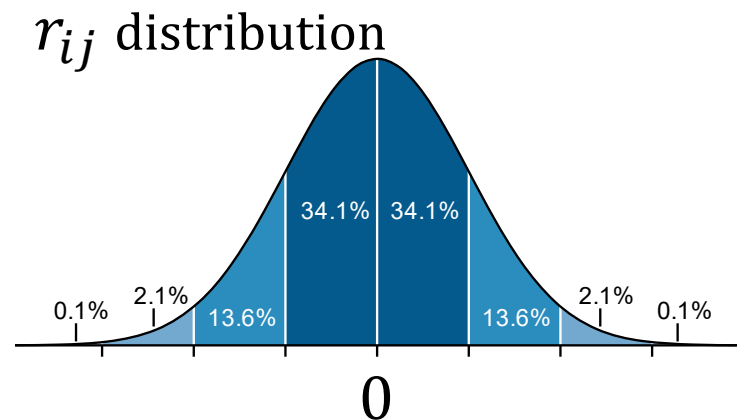
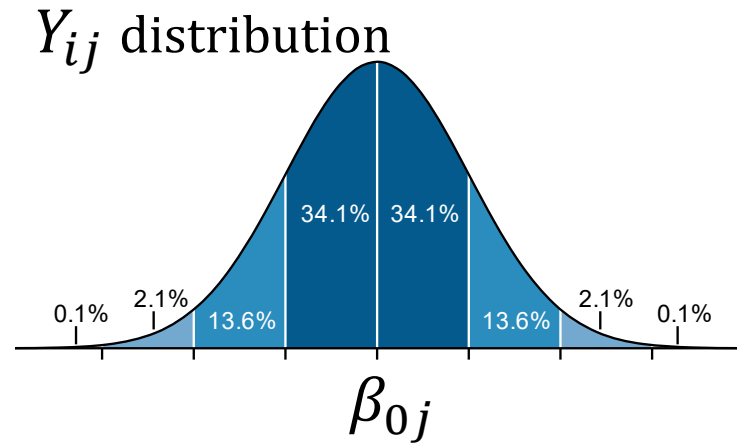
$$Y_{ij} = \beta_{0j} + r_{ij}, r_{ij} \sim N(0, \sigma^2)$$

β_{0j} represents the true mean math score for school j

r_{ij} is the deviation of student i 's math score from β_{0j}

Model-based approach: Level 1

- Within each school, the deviations of student math scores from the school's true mean are assumed normally distributed with a mean of 0 and variance σ^2
- One-way ANOVA assumption: the within-school variance is the same in all schools (homogeneity of variance assumption)



Source of histogram graphic: Ainali - Own work, CC BY-SA 3.0,
<https://commons.wikimedia.org/w/index.php?curid=3141713>

Model-based approach: Level 2

- Hallmark of hierarchical models: view level-1 parameters as outcomes in a level-2 model
- Level-2 (between-school) model:

τ_{00} is the between-school variance component

$$\beta_{0j} = \gamma_{00} + u_{0j}, u_{0j} \sim N(0, \tau_{00})$$

γ_{00} represents the grand mean math score for our population of interest (grade 8 students in U.S. schools in 2019)

u_{0j} is the deviation of the true mean for school j from the grand mean

Model-based approach: Level 2

- Level-2 (between-school) model:

$$\beta_{0j} = \gamma_{00} + u_{0j}, u_{0j} \sim N(0, \tau_{00})$$

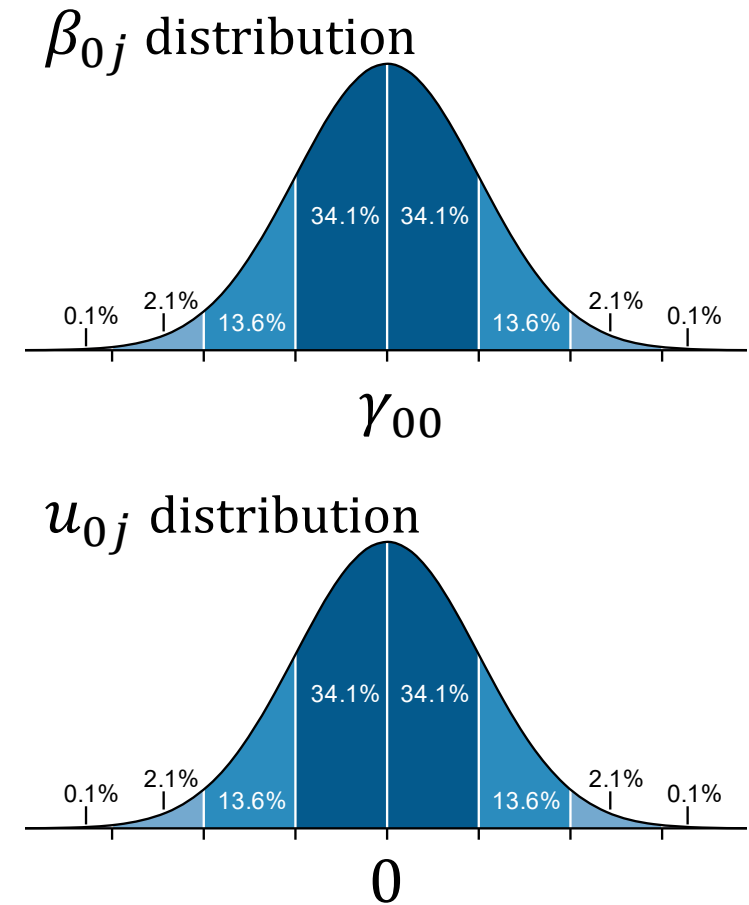
γ_{00} is sometimes called a
“fixed effect”

u_{0j} is sometimes called a
“random effect”

τ_{00} is sometimes called the
“random effect variance
component” or the
“parameter variance”

Model-based approach: Level 2

- Across schools, the deviations of schools' true mean math score from the grand mean are assumed normally distributed with a mean of 0 and variance τ_{00}



Source of histogram graphic: Ainali - Own work, CC BY-SA 3.0,
<https://commons.wikimedia.org/w/index.php?curid=3141713>

Model-based approach: Hierarchical model

- Level-1 (within-school) model: $Y_{ij} = \beta_{0j} + r_{ij}, r_{ij} \sim N(0, \sigma^2)$
- Level-2 (between-school) model: $\beta_{0j} = \gamma_{00} + u_{0j}, u_{0j} \sim N(0, \tau_{00})$
- Combined model: $Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$
- Same as a one-way ANOVA model with random effects, where:

$$\text{Var}(Y_{ij}) = \text{Var}(u_{0j}) + \text{Var}(r_{ij}) = \tau_{00} + \sigma^2$$

Estimate the hierarchical model in R

- Combined model:

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$$

```
m1 <- lmer(bsmmatxx ~ 1 + (1 | idschool), data=td.bx)

print(as_flextable(m1), preview = "pptx")

summary(m1)
```

Estimate the hierarchical model in R

- Combined model:

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$$

$$u_{0j} \sim N(0, \tau_{00})$$

$$r_{ij} \sim N(0, \sigma^2)$$

- $\hat{\gamma}_{00} = 538.979$

- $\hat{\tau}_{00} = (72.298)^2 = 5,227$

- $\hat{\sigma}^2 = (70.666)^2 = 4,994$

	Estimate	Standard Error	df	statistic	p-value	
<u>Fixed effects</u>						
(Intercept)	538.979	16.422	19	32.821	0.0000	***
<u>Random effects</u>						
idschool	sd__(Intercept)	72.298				
Residual	sd__Observation	70.666				

Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05

square root of the estimated residual variance: 70.7

data's log-likelihood under the model: -3,436.7

Akaike Information Criterion: 6,879.4

Bayesian Information Criterion: 6,892.6

Understanding Between-Group Variance

Estimating the between-school variance

- For the overall average math score (γ_{00}), the naïve estimate is the same as the model-based estimate.
- For the between-school variance (τ_{00}), the naïve estimate is greater than the model-based estimate:

- $\frac{\sum_{j=1}^J (\bar{Y}_{.j} - \bar{Y}_{..})^2}{J-1} = 5,393$

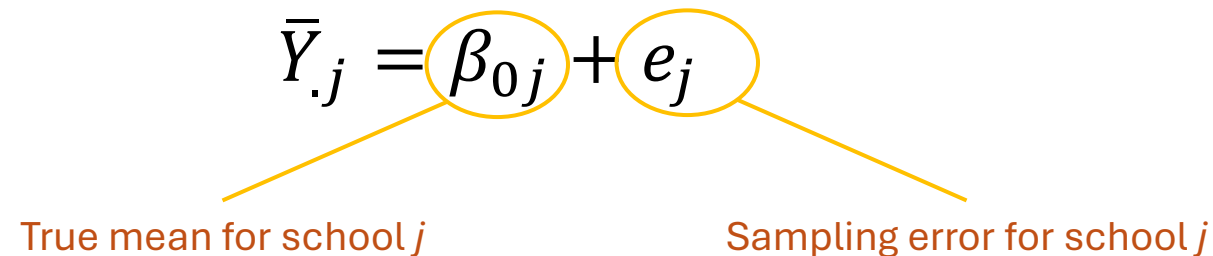
- $\hat{\tau}_{00} = (72.298)^2 = 5,227$

- Why?

Estimating the between-school variance

- The naïve estimate is based on the observed school means ($\bar{Y}_{.j}$), which are comprised of two components:

$$\bar{Y}_{.j} = \beta_{0j} + e_j$$



True mean for school j Sampling error for school j

- So the variance calculation based on the $\bar{Y}_{.j}$ values includes two sources of variance: parameter variance + error variance

$$\text{Var}(\bar{Y}_{.j}) = \Delta_j = \tau_{00} + V_j, \text{ where } V_j = \frac{\sigma^2}{n_j}$$

Estimating the between-school variance

- The naïve estimate: $\bar{Y}_{.j} = \beta_{0j} + e_j$
- We don't know β_{0j} or e_j but we can use the estimated within-school variance estimate ($\hat{\sigma}^2$) to approximate e_j based on the standard error (SE) of $\bar{Y}_{.j}$:

$$\text{SE}(\bar{Y}_{.j}) = \sqrt{V_j} = \sqrt{\frac{\hat{\sigma}^2}{n_j}}$$

Estimating the between-school variance

- Estimate of the error variance:

$$\hat{V}_j = \frac{\hat{\sigma}^2}{n_j} = \frac{4,994}{30} = 166.5$$

- Observed between-school variance:

$$\text{Var}(\bar{Y}_{.j}) = 5,393$$

- We can use these variance numbers to calculate the parameter variance ($\hat{\tau}_{00}$):

$$\hat{\tau}_{00} = \text{Var}(\bar{Y}_{.j}) - \hat{V}_j$$

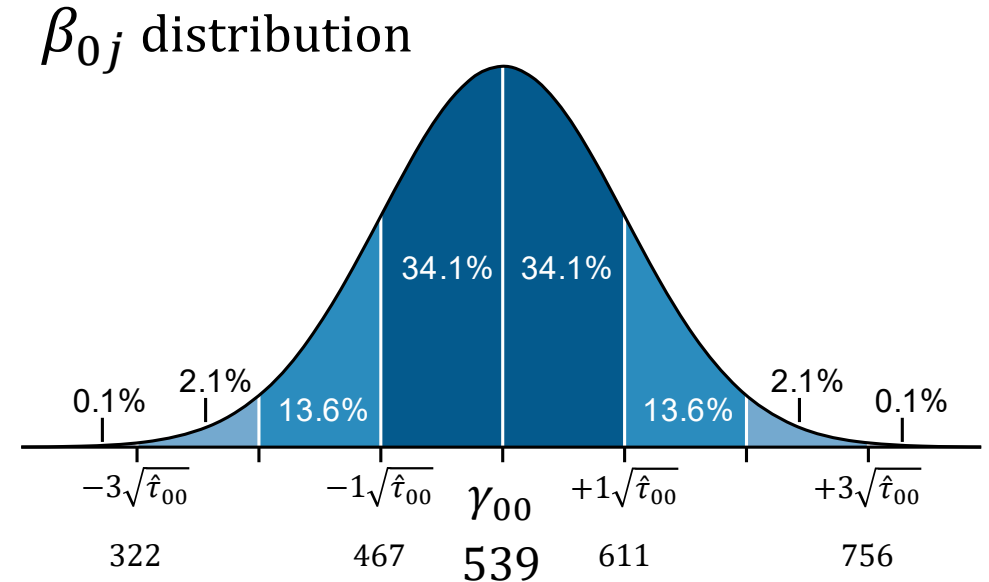
$$\hat{\tau}_{00} = 5,393 - 166.5 = 5,226.5$$

Interpreting the between-school variance

- Level-2 (between-school) model: $\beta_{0j} = \gamma_{00} + u_{0j}$, $u_{0j} \sim N(0, \tau_{00})$
- β_{0j} 's : the true mean math scores for the schools in our population of interest; normally distributed around a grand mean (γ_{00})
- τ_{00} : variance in the true means for the schools in our population around the grand mean (between-school variance)

Interpreting the between-school variance

- The model estimates for γ_{00} and τ_{00} provide the basis for a “best guess” of the distribution of the β_{0j} ’s in the population of interest
 - $\hat{\gamma}_{00} = 538.979$
 - $\hat{\tau}_{00} = (72.298)^2 = 5,227$
- Note: usually easier to interpret standard deviations than variances, so let’s consider $\sqrt{\hat{\tau}_{00}} = 72.298$

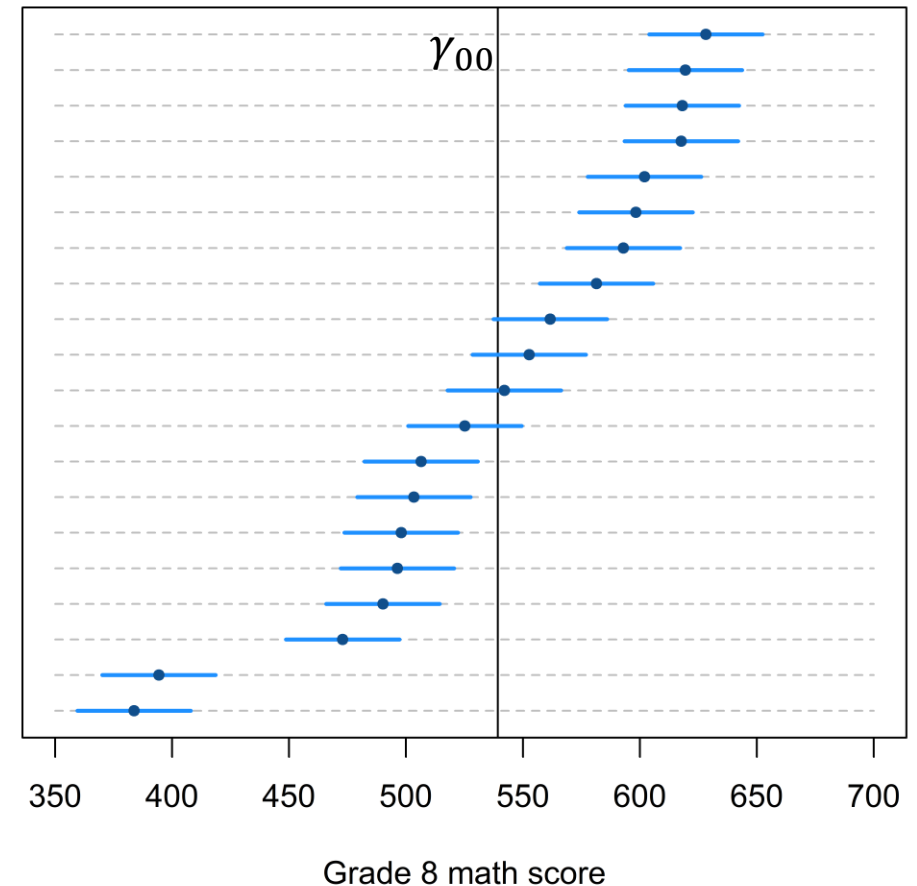


Source of histogram graphic: Ainali - Own work, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=3141713>

Interpreting the between-school variance

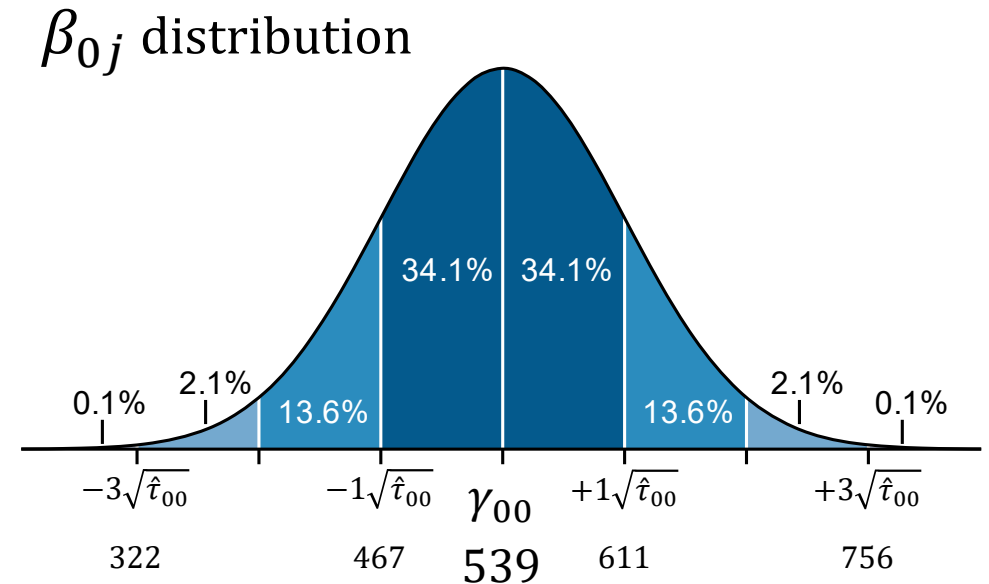
- It can also be useful to visualize the estimated school means and their respective confidence intervals

Estimated school mean math score (β_{0j}^*)
and 95% confidence interval



Interpreting the between-school variance

- What's the expected math score difference between a school 1 SD below average and a school 1 SD above average?
- Is that a substantial difference?



Source of histogram graphic: Ainali - Own work, CC BY-SA 3.0,
<https://commons.wikimedia.org/w/index.php?curid=3141713>

Interpreting the between-school variance

- What proportion of the observed between-school variation, $\text{Var}(\bar{Y}_{.j})$, is due to true parameter variance vs. error variance?

- $\text{Var}(\bar{Y}_{.j}) = \tau_{00} + V_j$
- $\text{Var}(\bar{Y}_{.j}) = \hat{\tau}_{00} + \frac{\hat{\sigma}^2}{n_j}$

97% of the observed between-school variation reflects true differences in school mean math scores

$$\frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \frac{\hat{\sigma}^2}{n_j}} = \frac{5,227}{5,227 + 166.5} = 0.97$$

Reliability of $\bar{Y}_{.j}$ as an estimate of β_{0j}

Interpreting the between-school variance

- What proportion of the total variability in student math scores is due to between-school differences vs. within-school differences?

- Combined model: $Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$

- $u_{0j} \sim N(0, \tau_{00})$

- $r_{ij} \sim N(0, \sigma^2)$

- $\hat{\gamma}_{00} = 538.979$

- $\hat{\tau}_{00} = (72.298)^2 = 5,227$

- $\hat{\sigma}^2 = (70.666)^2 = 4,994$

		Estimate	Standard Error	df	statistic	p-value	
<u>Fixed effects</u>							
	(Intercept)	538.979	16.422	19	32.821	0.0000	***
<u>Random effects</u>							
idschool	sd__(Intercept)	72.298					
Residual	sd__Observation	70.666					

Interpreting the between-school variance

- What proportion of the total variability in student math scores is due to between-school differences vs. within-school differences?
- The intraclass correlation (ICC):

$$\hat{\rho} = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}^2} = \frac{5,227}{5,227 + 4,994} = 0.51$$

51% of the total variation in students' math scores lies between school

- A measure of how similar students are within schools. The higher the ICC value, the more similar students are within schools (more of the variance is between schools)

Interpreting the between-school variance

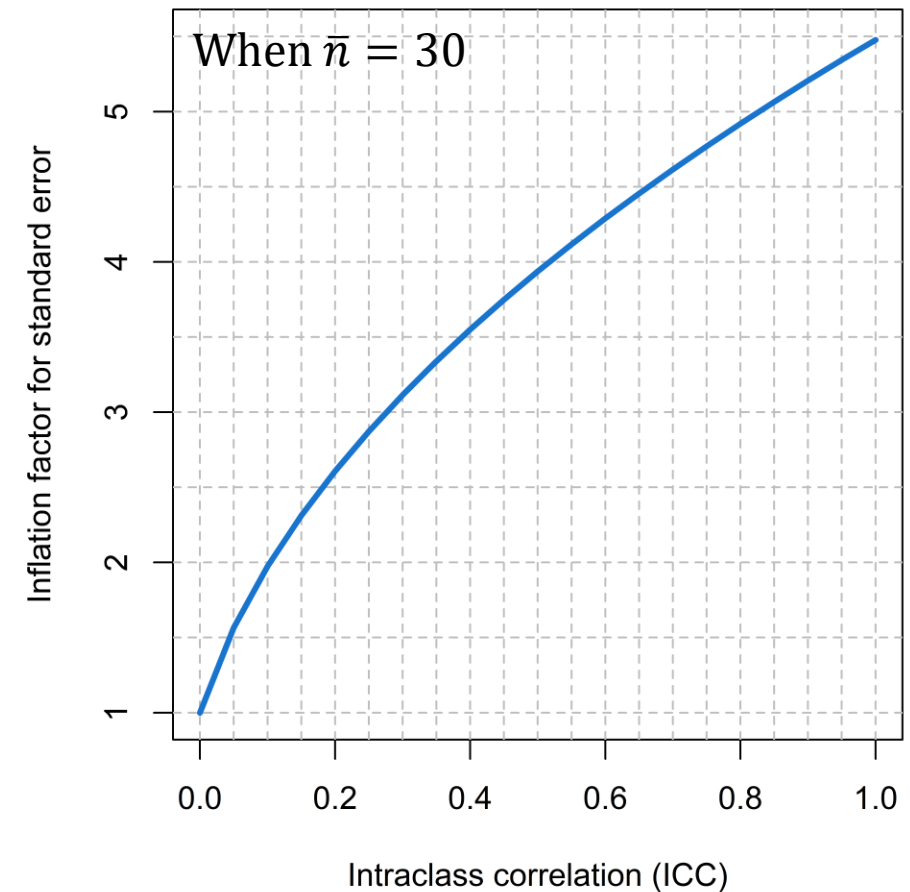
- The ICC is a useful measure of how “important” it is to account for the multilevel/clustered/nested structure of the data
- Higher ICC values indicate that a more significant violation of the independence assumption if using more traditional methods (e.g., OLS regression)
- The *design effect* (DEFF) is another way to express the relationship between the ICC and the independence assumption:

$$DEFF = 1 + \rho(\bar{n} - 1)$$

where \bar{n} is the average sample size within each cluster

Interpreting the between-school variance

- The larger the design effect, the larger the standard errors will be (for a given sample size)
- The square root of the design effect indicates how much a standard error that does not account for clustering should be “inflated” to reflect the actual amount of “independent” information in the data



Introduction to R for Multilevel Modeling

See R Walkthrough #1 Handout