

# 증거기반연구 6차 세미나

Week 3: HLM (위계적 선형 모형)

주상원

서울대학교 행정대학원 석사과정

24 Jan 2023

# 열심히 해봅시다... 영자영자...



# 목차

## i. 이론

- 위계적 선형모형 (Hierarchical Linear Modelling = Mixed Effect Model = Multilevel Model)의 개념
  - 핵심어: Gauss-Markov Theorem, Nested Data, Ecological fallacy
- HLM의 유형 및 분석단계별 유의사항
  - 핵심어: Intercept and Slopes, MLE, ICC, AIC(BIC), Deviance, Fixed and Random Effects, Within vs. Between
- Inter-Level Interaction, Centering and Contextual Effects

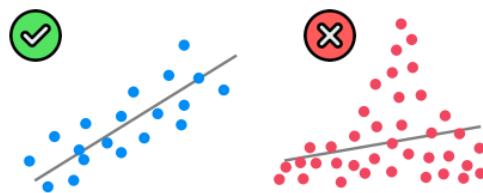
## ii. 실습: R과 Stata로 실제 데이터 분석

# Module I: 위계적 선형모형의 개념

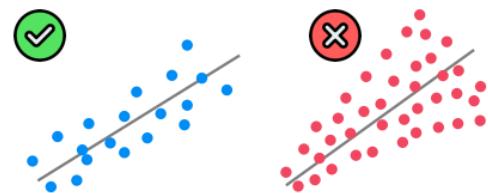
# 회귀분석의 기본 가정s

- **가우스-마르코프 정리:** 선형회귀분석에서 (1) 선형이고, (2, 3) 오차( $\epsilon$ )가  $\epsilon \sim N(0, \sigma^2)$ , (4) 오차가 상관관계가 없고, (5) 설명변수가 외생변수일 때 최소제곱 추정량(OLS)은 BLUE(Best Linear Unbiased Estimator)이다. [출처](#)

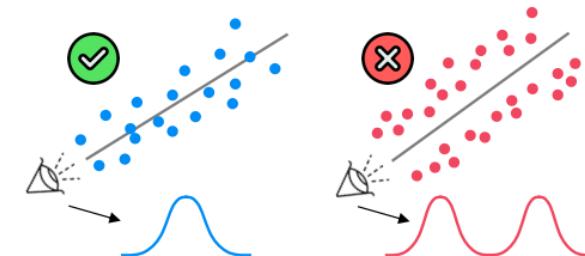
1. Linearity  
(Linear relationship between Y and each X)



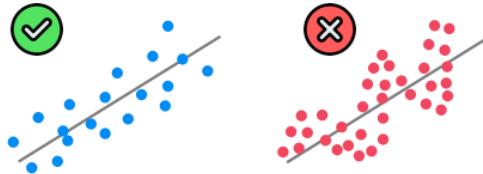
2. Homoscedasticity  
(Equal variance)



3. Multivariate Normality  
(Normality of error distribution)



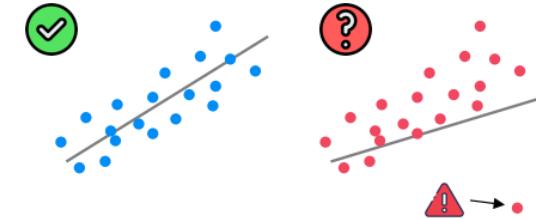
4. Independence  
(of observations. Includes "no autocorrelation")



5. Lack of Multicollinearity  
(Predictors are not correlated with each other)



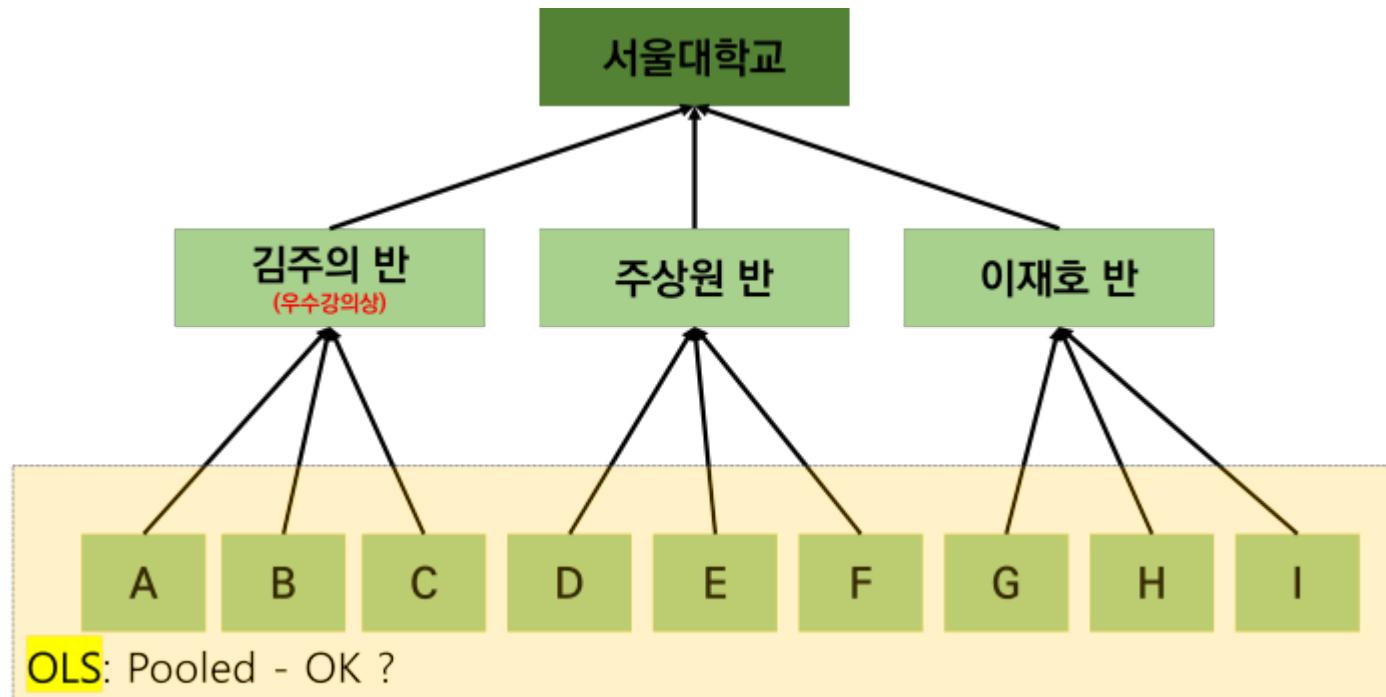
6. The Outlier Check  
(This is not an assumption, but an "extra")



Assumptions

# 그러나 현실은...? 특히 조직연구에서는..?

- Nested Data: 학교 > 학년 > 반 > 개인
- 조직연구의 경우: e.g. 공무원 조직: 공무원 개인 → 과 → 국 → 처



# 위계적 다층자료: Nested Data의 주요 형태

- 위계적 구조: Long Form vs. Wide Form

Table 1: 학생이 학교에 내재한 2-수준 다층자료의 예시

(a) Wide Form Nested

이름	1반	2반	3반	통계성적
A	0	0	0	95
B	1	0	1	65
C	0	1	0	85
D	0	1	0	45
E	0	0	1	75

(b) Long Form Nested

이름	소속반	통계성적
A	1반	95
B	3반	65
C	2반	85
D	2반	45
E	3반	75

- if 소속반을 시간으로 바꾸어 반복측정?: 종단(Longitudinal) Panel Data

# 위계적 다층자료 (Cont'd)

- 학생들은 각 학교에 내재한 구조 동일 반에 속한 학생들은 학교의 문화, 학습환경, 친구관계, 교사 등 수많은 요인들을 공유하므로 동일학교 학생들의 행동 (e.g. 성취도, 학습동기)에 영향을 미침.
- **상호의존성**  
관찰단위인 학생들은 동일 학교, 동일 반 내에서는 **상호의존성을 가지게 되고**, 소속이 다르면 독립성을 갖는다. → 독립성 가정이 결과적으로 위배되게 됨.
- **복수의 Unit of Analysis**  
학생수준에서의 변수 (Level 1): 성취도, 가정배경, 친구관계 등  
학교수준에서의 변수 (Level 2): 교장경력, 교사 수, 학생 수, 소재지, 설립유형 등
- **불균형 자료**  
각 학교별로 학생이 다르기에 관측치는 일반적으로 같을 수 없음.
- 이 외에도 교차적 다층자료, 반복측정 다층자료 등의 경우 각각 고유의 특성 존재.

# 독립성 (Independence) 가정의 위배

## i. Error와 독립변수간에 상관관계가 없어야 함 (Endogeneity issue)

- 중요한 변수가 모형에서 생략되어지거나, 비체계적 오류로 인한 측정오차, 독립변수와 종속변수간 동시상관시 발생

## ii. Error 간에 상관관계가 없어야 함

- $COV(e_i, e_j) = E(e_i e_j) - E(e_i)E(e_j) = 0$

상위수준(Level 2) 군집화 → Error간에 상관관계가 발생한다면?  
회귀분석의 추정치의 분산이 과도하게 커짐 → 정확 X

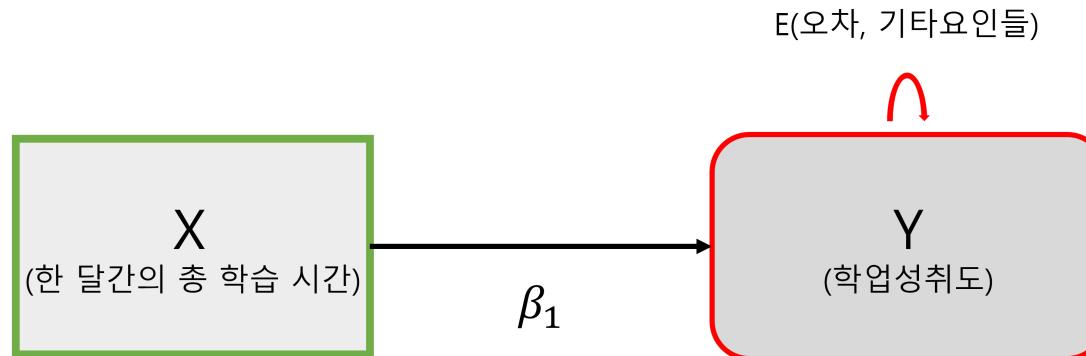
# 독립성 (Independence) 가정의 위배 (Cont'd)

학습시간과 통계과목 성적간의 관계 분석

- 데이터

기술통계량		Raw Data												
skim_type	skim_variable	n_missing	complete_rate	factor.ordered	factor.n_unique	factor.top_counts	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100	numeric.hist
factor	group	0	1	FALSE	30	1: 100, 2: 100, 3: 100, 4: 100	NA	NA	NA	NA	NA	NA	NA	NA
numeric	x	0	1	NA	NA	NA	40.62740	6.759487	22.89716	35.13338	40.44313	45.90419	64.73969	
numeric	y	0	1	NA	NA	NA	66.30165	14.964766	36.70912	53.70978	66.29439	79.20288	97.40393	

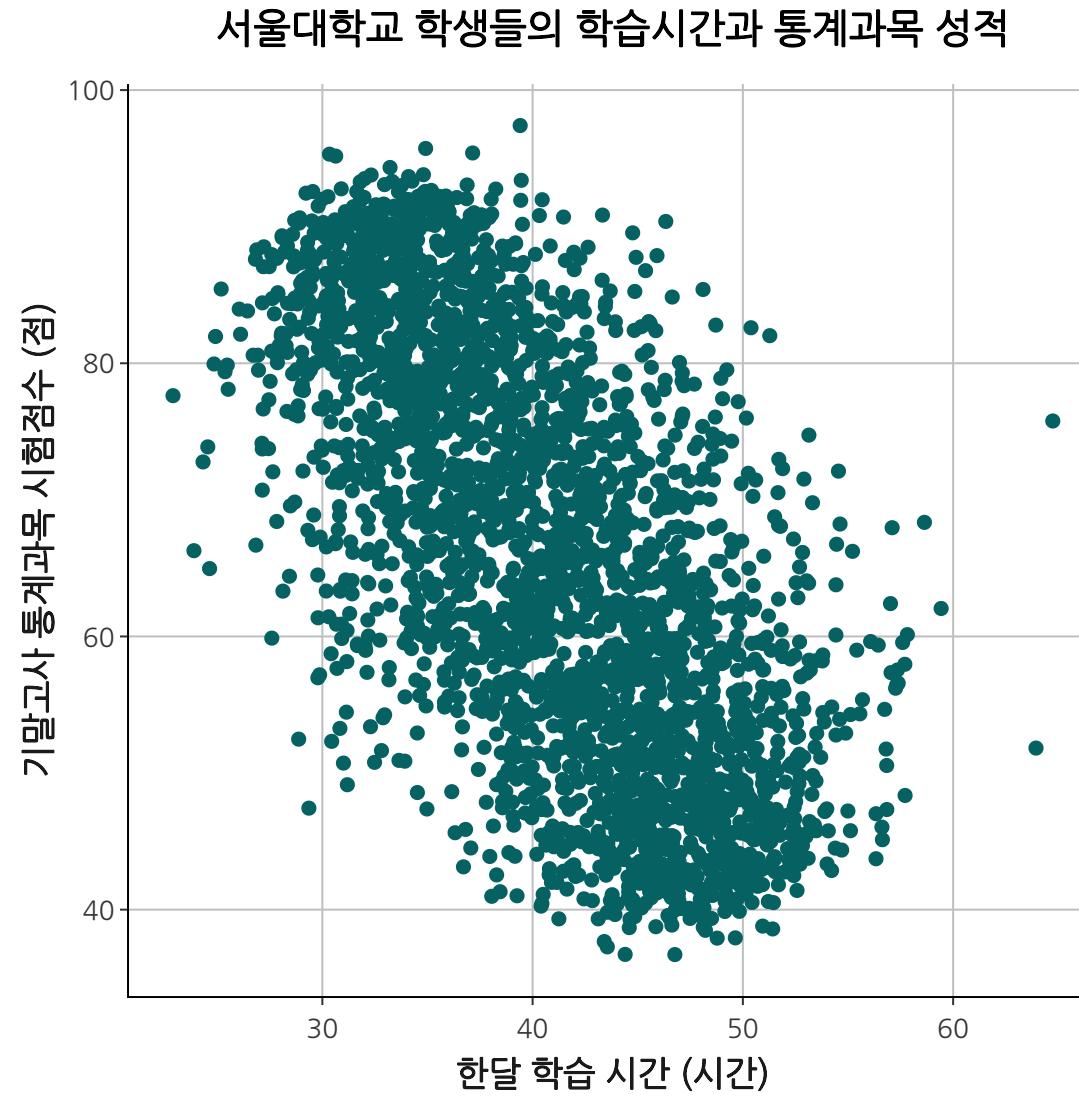
- 모형



$$Y = \beta_0 + \beta_1 X + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2)$$

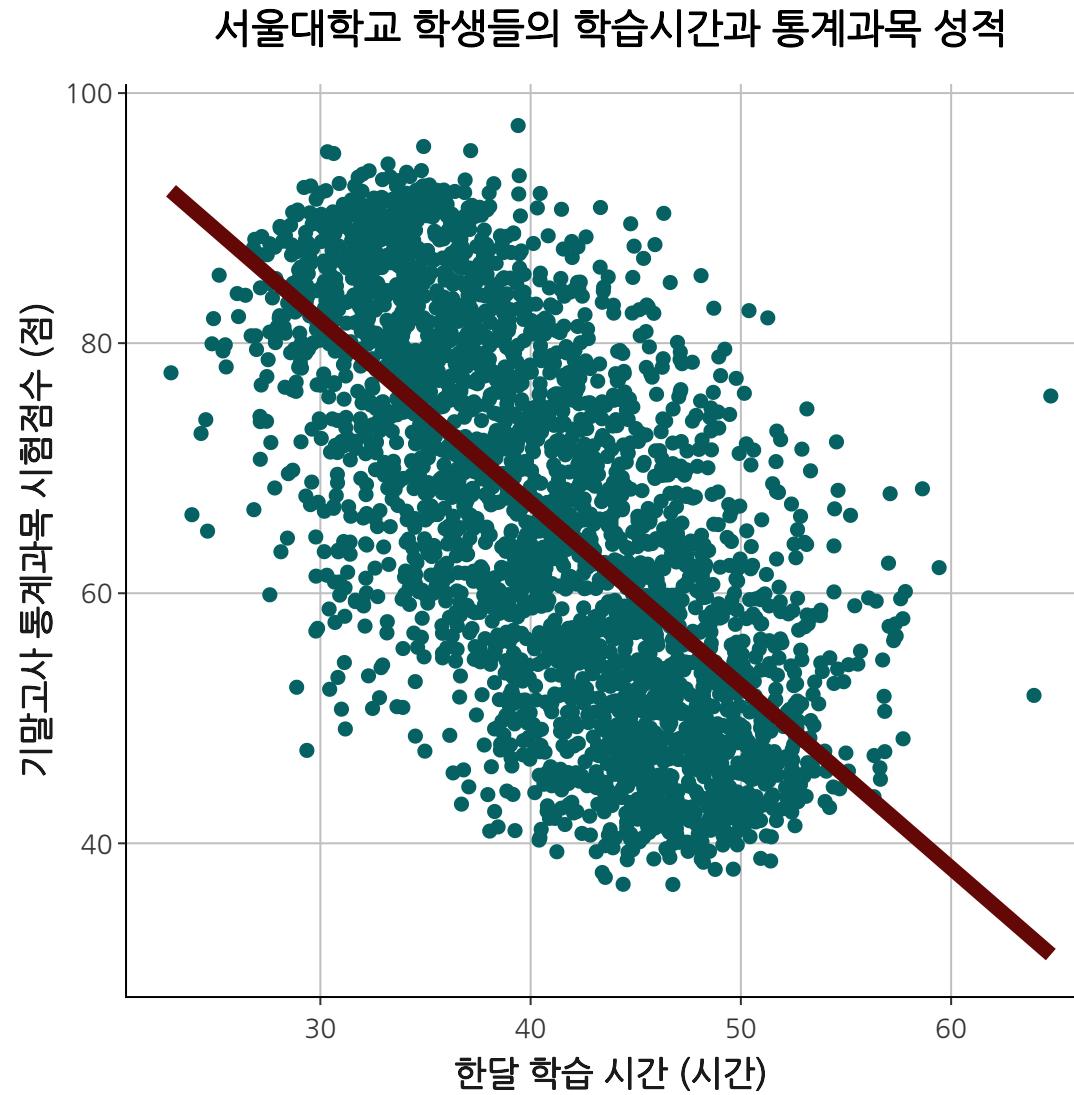
# 독립성 (Independence) 가정의 위배 (Cont'd)

- 시각화



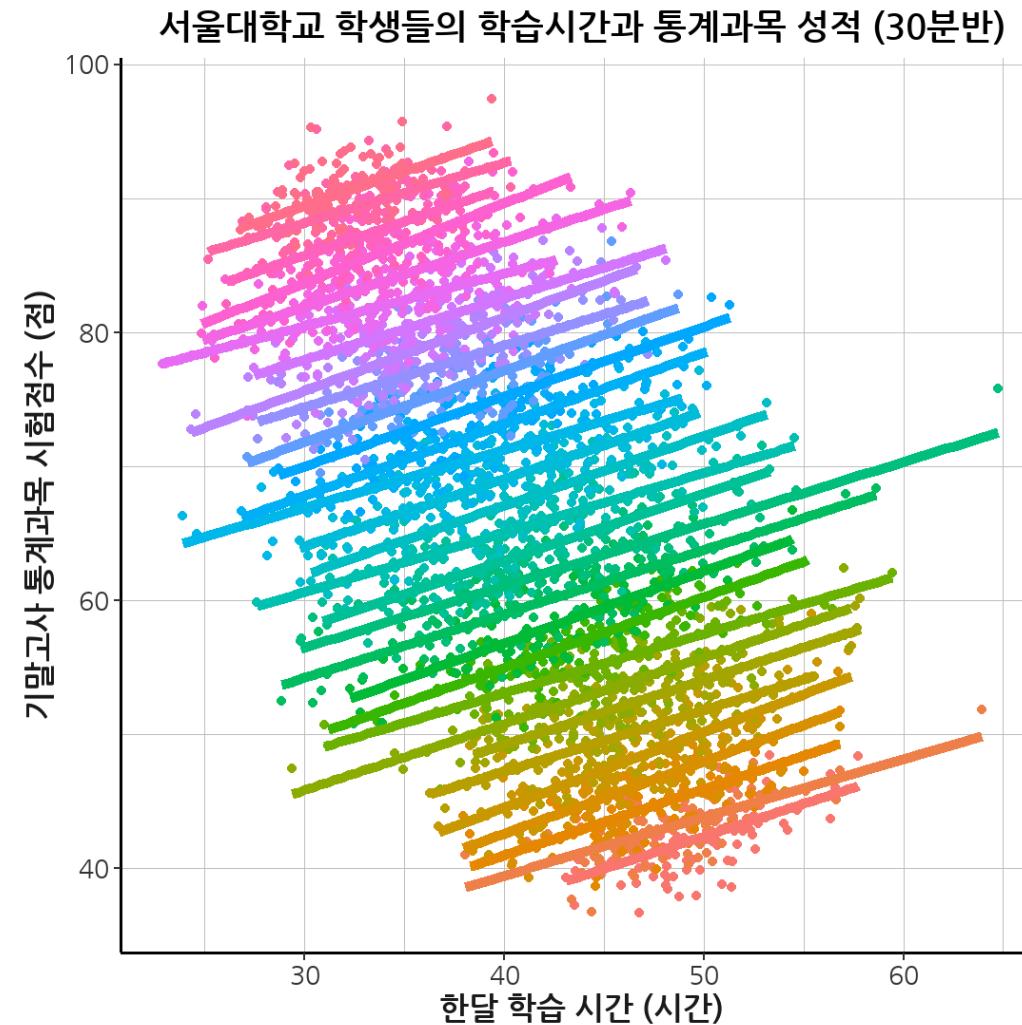
# 독립성 (Independence) 가정의 위배 (Cont'd)

- 시각화



# 독립성 (Independence) 가정의 위배 (Cont'd)

- 시각화



# 독립성 (Independence) 가정의 위배 (Cont'd)

- Pooled OLS

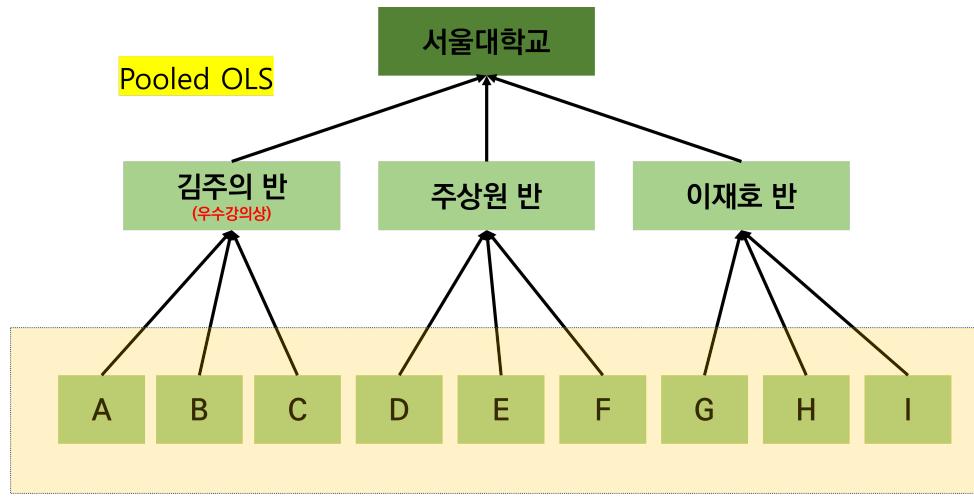
term	estimate	std.error	statistic	p.value
(Intercept)	125.613	1.252	100.337	0
x	-1.460	0.030	-48.028	0

- Grouped regression 30개 그룹의 estimates들의 평균

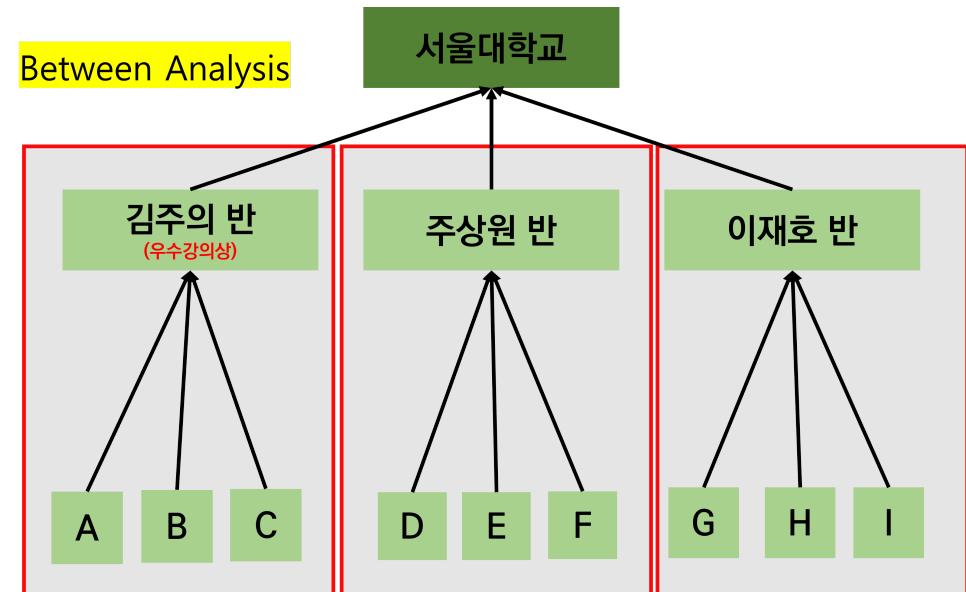
term	estimate	std.error	statistic
(Intercept)	46.263	1.941	25.361
x	0.493	0.048	10.775

# 기준 접근의 한계

- How to investigate relationships between variables that reside at different hierarchical levels (Bryk & Raudenbush, 2002)  
**Disaggregate?**: POLS → independence of obs assumption violated  
**Aggregate?**: Between → waste information from indvs, sample ↓



POLS



Between

# 위계적 선형 모형 (HLM)

대안: Hierarchical Linear Models

(Hierarchical Linear Model = Mixed Effect Model = Multilevel Model)

- designed to overcome the weakness of the disaggregated and aggregated approaches
1. explicitly model both **individual and group level residuals**, therefore, recognizing the partial interdependence of individuals within the same group (compared to OLS)
  2. investigate both **lower level unit and higher level unit variance** in the outcome measure
- ∴ Model both within and between group variance (i.e., able to preserve potentially meaningful within group variance) + Investigate the influence of higher level units on lower level outcomes

# Module 1: Sum-up

- 현대 사회에서 인간의 다양한 위계에 nested 되어있는 존재이다.
- 이러한 조직 내에서의 군집성으로 인해 서로서로에게 영향을 주게 되고, 결과적으로 개체들 간의 독립성을 가정하는 가우스-마르코프 가정을 충족하는 것이 까다로움
- ∴ 개인간의 독립성이 존재하기 어렵고, 자료 자체가 nested되어 있다는 것을 고려하지 못함 → Individual 대상 연구에서 (특히, 조직맥락) OLS는 더이상 만능 X, 생태학적 오류를 범할 수 있게 됨
- HLM의 필요성: within과 between을 가중평균하고, level 1과 2를 동시에 고려

# Module II: HLM의 유형 및 분석

# HLM의 장점

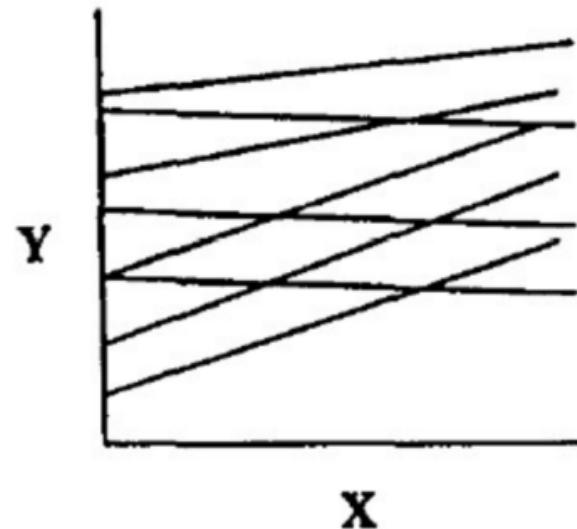
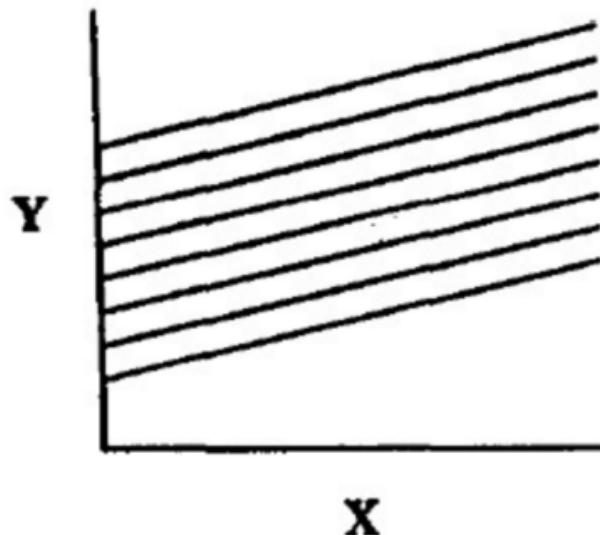
- Improves estimation of individual effects
- Models cross-level effects: an interaction
- Better partitioning of variance and covariance you have variance and covariance of data set, and you can think about how much is due to Level 2 and how much is due to Level 1 etc. E.g. How much is school, how much is student?
- No assumption of homogeneity of slopes i.e., that each data entry can have a different slope
- No assumption of independence because in this model they are correlated
- Missing data OK the structure of the data - you don't need data for people at every time point (e.g., repeated measures), or every group has to have a score for every person (e.g., nested).

# HLM의 어려움

- hard to conceptualize phenomena at more than one level
- requires conceptual and theoretical understanding at various levels
- easy to use the code (R - lmer4, Stata - mixed), difficult to understand what its doing
- most important: interpreting what the estimated parameters actually mean

# HLM의 이론

- Identify whether outcomes vary among persons with different attributes
- Model differences in average outcomes among groups
- Model differences in outcomes associated with individual attributes among groups.



# HLM의 이론 (Cont'd)

- OLS

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad e_i \sim N(0, \sigma^2)$$

- 다층모형은 집단내(개인:  $i$ )과 집단간(집단:  $j$ ) 모형을 별개로 model specification 한다

Level 1:

$$Y_{0j} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \quad r_{ij} \sim N(0, \sigma^2)$$

Level 2:

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

$$\beta_{1j} = \gamma_{10} + u_{1j}, \quad u_{1j} \sim N(0, \tau_{11})$$

- Fixed effects:

$\gamma_{00}$  = average outcome for sample of groups

$\gamma_{10}$  = average individual effect (slope) on outcome

- Random effects:

$u_{0j}$  = unique effect of group  $j$  on average outcome

$u_{1j}$  = unique effect of group  $j$  on average slope

# HLM을 통해 해결가능한 연구문제 examples

1. 우리나라 중학생 수학성취도의 학교간 교육격차는 어느정도인가?
2. 중학교 학생들의 수학성취도는 동일 학교 내에서 어느정도 차이가 있는가?
3. 학생들 사이의 수학성취도 차이에서 몇 %가 소속학교의 영향인가?
4. 학생들의 수학성취도는 개인차 요인의 영향을 더 많이 받는가 아니면 학교차 요인의 영향을 더 많이 받는가?
5. 저소득층 학생비율이 높은 학교의 학생들은 수학성취도에서 어느정도의 불이익을 받는가?
6. 학생 가정의 SES는 수학성취도와 어느정도 관련이 있는가?
7. 학생가정의 SES를 통제한 이후에도, 학교별 고정평균 수학성취도는 여전히 학교간에 차이가 있는가?
8. 가정의 SES가 수학성취도에 미치는 효과는 모든 학교에서 유사한가? 만일 학교에 따라 다르다면 그 크기는 어느 정도인가?
9. 학생의 가정배경을 통제한 이후에 어떠한 특성의 중학교에서 평균 수학성취도가 높은가?
10. 학생들의 수학성취도가 가정환경에 영향을 받는 정도는 어떤 특성의 학교에서 더 커지는가?

# Fixed Effect and Random Effect

In Panel, only on intercepts

- Fixed Effect: Time invariant Observation's Effect vs
- Random Effect: Time invariant Observation's Effect + *Random Part*

In HLM, similar but also on slopes

- **Fixed Effects:** parameter estimates that do not vary across groups (group invariant)  
The  $\gamma$ 's in equations represent fixed effects
- **Random coefficients:** parameter estimates that are allowed to vary across groups such as the level-1 regression coefficients (e.g.,  $\beta_{0j}$  and  $\beta_{1j}$ ).

# Fixed Effect and Random Effect (cont'd)

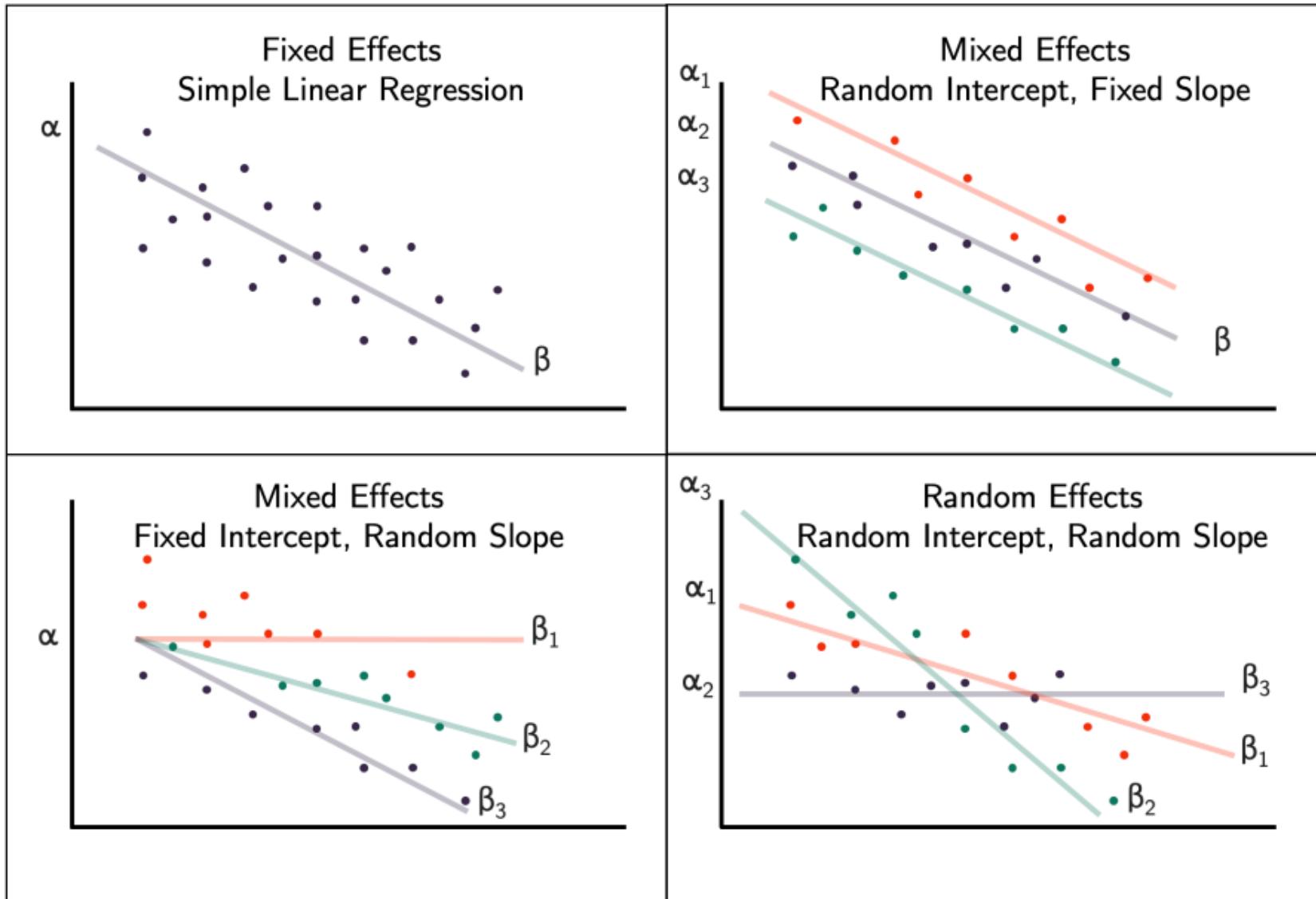
- **Fixed Effects**: parameter estimates that do not vary across groups (group invariant)  
The  $\gamma$ 's in equations represent fixed effects
- **Random coefficients**: parameter estimates that are allowed to vary across groups such as the level-1 regression coefficients (e.g.,  $\beta_{0j}$  and  $\beta_{1j}$ ).
- Equation

$$Y_{0j} = \gamma_{00} + u_{0j} + (\gamma_{10} + u_{1j})X_{ij} + r_{ij}$$
$$= \underline{\gamma_{00} + \gamma_{10}X_{ij}} + \underline{u_{0j} + u_{1j}X_{ij}} + r_{ij}$$

- Error Term이 복잡함: Group들 사이의 분산( $u_{0j}, u_{1j}$ )과 Individual들 사이의 Group내 분산( $r_{ij}$ )이 동시에 존재
- OLS로는 계산이 어렵기에 Maximum Likelihood Estimation을 활용

# Fixed Effect and Random Effect (cont'd)

출처



# Model Specification with R

- R 설치과정 참고: 설치
- 기초 R 연습 참고 사이트: 슬기로운 통계생활
- 기초 R 문서 버전: R for Data Science
- 사용하는 패키지: lmerTest, lme4, bruceR

# 데이터 불러오기

```
1 # install.packages("pacman")
2 pacman::p_load("tidyverse", "broom", "lme4", "lmerTest",
3                 "bruceR", "readstata13", "magrittr",
4                 "ggplot2", "skimr", "psych", "merTools", "bruceR")
5 ## 자신이 설정하고 싶은 곳으로 설정
6 setwd("E:/OneDrive - SNU/(B) 대학원/세미나/HLM/hlm")
7 getwd()
8
9 # Read data
10 data_lv1 <- read.dta13("./HSB1.dta")
11 data_lv2 <- read.dta13("./HSB2.dta")
```

- 정상적으로 로드 되었는지 확인

```
1 # Size
2 dim(data_lv1)
```

```
[1] 7185      5
```

```
1 dim(data_lv2)
```

```
[1] 160      4
```

# Data 설명: High School and Beyond (HS&B)

High School and Beyond (HS&B) is a national *longitudinal* study originally funded by the United States Department of Education's National Center for Education Statistics (NCES) as a part of their longitudinal studies program.

Purpose was to document the educational, vocational, and personal development of young people following them over time as they begin to take on adult roles and responsibilities

**Level-1 file:** HSB1.dta, 7,185 observations with 4 variables

- MINORITY: an indicator for student ethnicity (1 = minority, 0 = other)
- FEMALE: an indicator for student gender (1 = female, 0 = male)
- SES: a standardized scale constructed from variables measuring parental education, occupation, and income
- MATHACH: a measure of mathematics achievement

**Level-2 file:** HSB2.dta, 160 schools with 3 variables

- SIZE: school enrollment
- SECTOR (1 = Catholic, 0 = public)
- HIMNTY (1 = more than 40% minority enrollment, 0 = less than 40%)

# Data Glimpse

- data\_lv1 glimpse

```
1 glimpse(data_lv1)
```

```
Rows: 7,185
Columns: 5
$ id      <chr> "1224", "1224", "1224", "1224", "1224", "1224", "1224...
$ minority <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0...
$ female    <dbl> 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1...
$ ses       <dbl> -1.528, -0.588, -0.528, -0.668, -0.158, 0.022, -0.618, -0.998...
$ mathach   <dbl> 5.876, 19.708, 20.349, 8.781, 17.898, 4.583, -2.832, 0.523, 1...
```

- data\_lv2 glimpse

```
1 glimpse(data_lv2)
```

```
Rows: 160
Columns: 4
$ id      <chr> "1224", "1288", "1296", "1308", "1317", "1358", "1374", "1433...
$ size     <dbl> 842, 1855, 1719, 716, 455, 1430, 2400, 899, 185, 1672, 530, 53...
$ sector   <dbl> 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, ...
$ himinty <dbl> 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, ...
```

# Data Summarise

- data\_lv1 기술통계

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
id*	1	7185	79.96	45.44	79.00	79.71	57.82	1.00	160.00	159.00	0.04	-1.17	0.54
minority	2	7185	0.27	0.45	0.00	0.22	0.00	0.00	1.00	1.00	1.01	-0.98	0.01
female	3	7185	0.53	0.50	1.00	0.54	0.00	0.00	1.00	1.00	-0.11	-1.99	0.01
ses	4	7185	0.00	0.78	0.00	0.02	0.85	-3.76	2.69	6.45	-0.23	-0.38	0.01
mathach	5	7185	12.75	6.88	13.13	12.91	8.12	-2.83	24.99	27.82	-0.18	-0.92	0.08

- data\_lv2 기술통계

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
id*	1	160	80.50	46.33	80.5	80.50	59.30	1	160	159	0.00	-1.22	3.66
size	2	160	1097.83	629.51	1061.0	1058.24	695.34	100	2713	2613	0.46	-0.61	49.77
sector	3	160	0.44	0.50	0.0	0.42	0.00	0	1	1	0.25	-1.95	0.04
himinty	4	160	0.28	0.45	0.0	0.22	0.00	0	1	1	1.00	-1.01	0.04

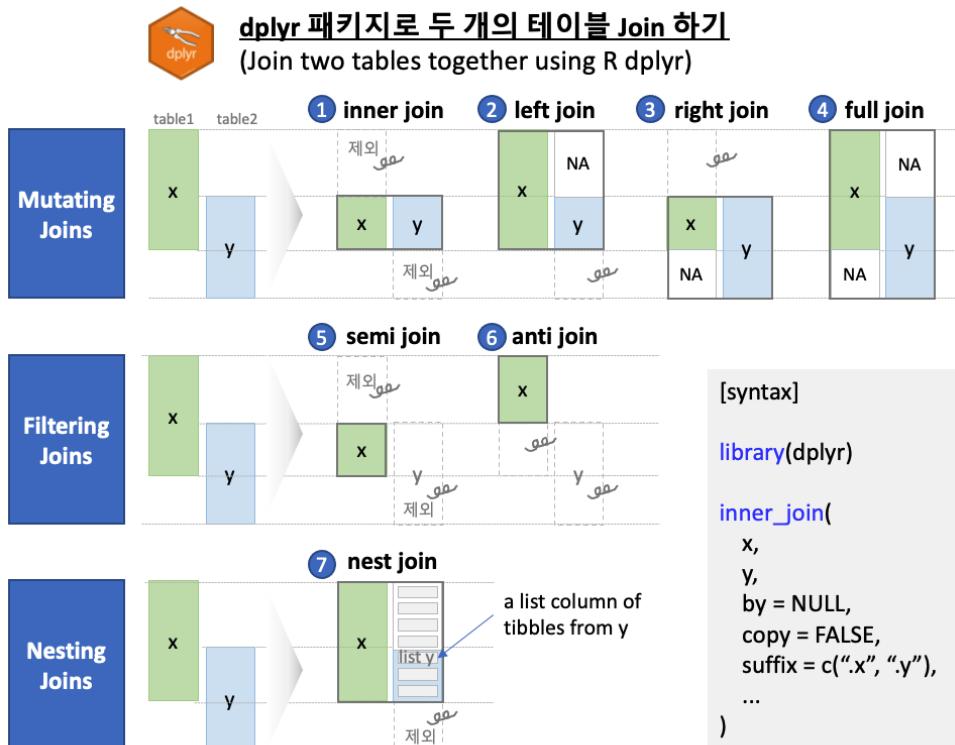
# Data Merge: dplyr package - joins

dplyr 패키지는 다양한 데이터 merge 함수들을 제공하고 있음. 이중 HLM에서 자주 활용되는 `left_join`, `right_join`, `full_join`, `inner_join`, `semi_join`, `anti_join`등에 대해 간단히 다루고 넘어감.

- `left_join(right_join)`: Join matching rows from y to x (x to y)
- `full_join`: Join data. Retain all values, all rows
- `inner_join`: Join data. Retain only rows in both sets
- `semi_join`: All rows in a that have a match in b
- `anti_join`: All rows in a that do not have a match in b
- Join 함수의 구조

```
1 left_join(  
2   x, # Level 1 data-set name  
3   y, # Level 2 data-set name  
4   by = c("id_x" = "id_y"), # 각각의 데이터 셋에서 어떤 변수를 기준으로 merge되는지 설정  
5   copy = FALSE, # 가만히 두기  
6   suffix = c(".x", ".y"), # 만약 id 이외에 서로 겹치는 변수가 있을때 어느 데이터셋인지  
7   ...  
8   keep = FALSE # 가만히 두기  
9 )
```

# Data Merge: dplyr package - joins (Cont'd)

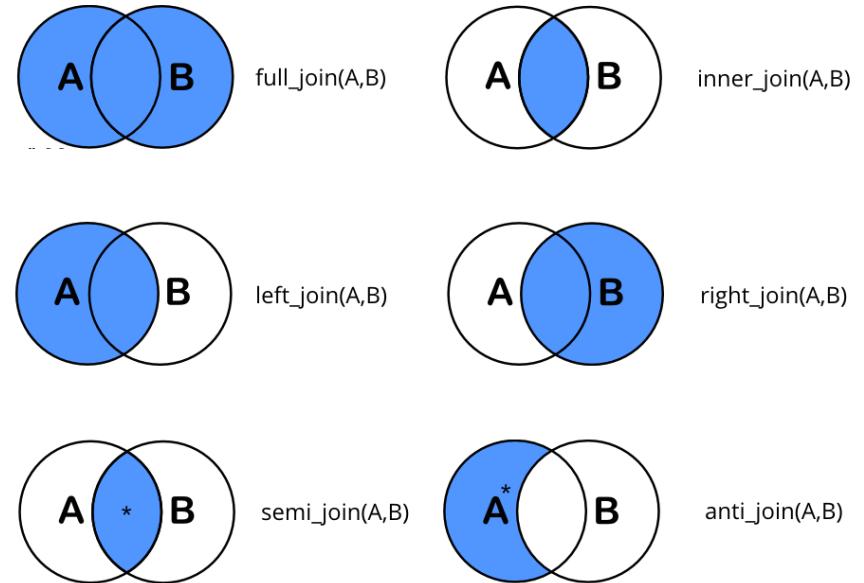


[syntax]  
`library(dplyr)`  
`inner_join(  
 x,  
 y,  
 by = NULL,  
 copy = FALSE,  
 suffix = c(".x", ".y"),  
 ...  
)`

[R, Python 분석과 프로그래밍의 친구] <https://rfriend.tistory.com>

Join 함수들

출처



벤다이어그램

출처

# Data Merge: data\_lv1 and data\_lv2

Pipe Operator: `%>%`

- lhs `%>%` rhs는 lhs의 결과를 rhs의 첫번째 변수로 넘겨주는 역할. '.'은 앞의 값의 위치를 구체적으로 지정하기 위해 사용
- $x \%>\% f$  is equivalent to  $f(x)$
- $x \%>\% f(y)$  is equivalent to  $f(x, y)$
- $x \%>\% f(y, .)$  is equivalent to  $f(y, x)$

data\_lv1와 data\_lv2 데이터 병합

```
1 data_merged <- data_lv1 %>%
2   left_join(data_lv2, by = "id")
3 head(data_merged, 5)
```

id	minority	female	ses	mathach	size	sector	himinty
1224	0	1	-1.53	5.88	842	0	0
1224	0	1	-0.59	19.71	842	0	0
1224	0	0	-0.53	20.35	842	0	0
1224	0	0	-0.67	8.78	842	0	0
1224	0	0	-0.16	17.90	842	0	0

# Five models in HLM

Overview of HLM Two-Level Models

	(1) One-way ANOVA	(2) Means-as-Outcomes (Between)	(3) One-way ANCOVA	(4) Random Coefficient	(5) Intercept-and-Slopes-as- Outcomes
<b>Level-1- Models:</b>			(Different Intcpt, Same Slope)	(Different Intcpt, Different Slope)	〈우리의 목표〉 (Different Intcpt, Different Slope)
For Level-1 Intercept	$Y_{ij} = \beta_{0j} + r_{ij}$	$Y_{ij} = \beta_{0j} + r_{ij}$	$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}$	$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}$	$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}$
Level 1 Independent Variable	NO	NO	YES	YES	YES
<b>Level-2- Models:</b>					
For Level-1 Intercept:	$\beta_{0j} = \gamma_{00} + u_{0j}$	$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$	$\beta_{0j} = \gamma_{00} + u_{0j}$	$\beta_{0j} = \gamma_{00} + u_{0j}$	$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$
Level 2 Independent Variable	NO	YES	NO	NO	YES
For Level 1 Slopes	NO	NO	$\beta_{1j} = \gamma_{10}$	$\beta_{1j} = \gamma_{10} + u_{ij}$	$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{ij}$
Level 2 Independent Variables			Fixed	Random	Yes (Sometimes)

# 변수들에 대한 설명

- 분석 목적: Level-1 변수인 SES (Social Economic Status)와 Level-2 변수인 Sector (Public or Private)가 학생들의 수학성취도에 미치는 영향
- 독립변수: ses, sector
- 종속변수: mathach
- $i$ 는 개인수준 (Level-1)의 첨자,  $j$ 는 집단수준 (Level-2)의 첨자
- $\therefore Y_{ij}$ :  $j$  번째 학교에 다니는  $i$ 번째 학생의 수학성취도 점수,  $\beta_{0j}$ 는  $j$ 번째 학교의 평균 수학성취도 점수
- 평균 학급당 인원: 44.91명 (sd: 11.85)

# Model 0. Preliminary Analysis

One-Way ANOVA 집단수준의 Mean Squares들의 값과 개인수준의 Mean Squares들의 값을 비교함을 통해서, 개인수준의 변량대비 집단 수준의 변량 비교

→ 쉽게 말해 집단간 모평균(여기서는 수학성취도)이 서로 다른가? 수준효과가 존재하는가?

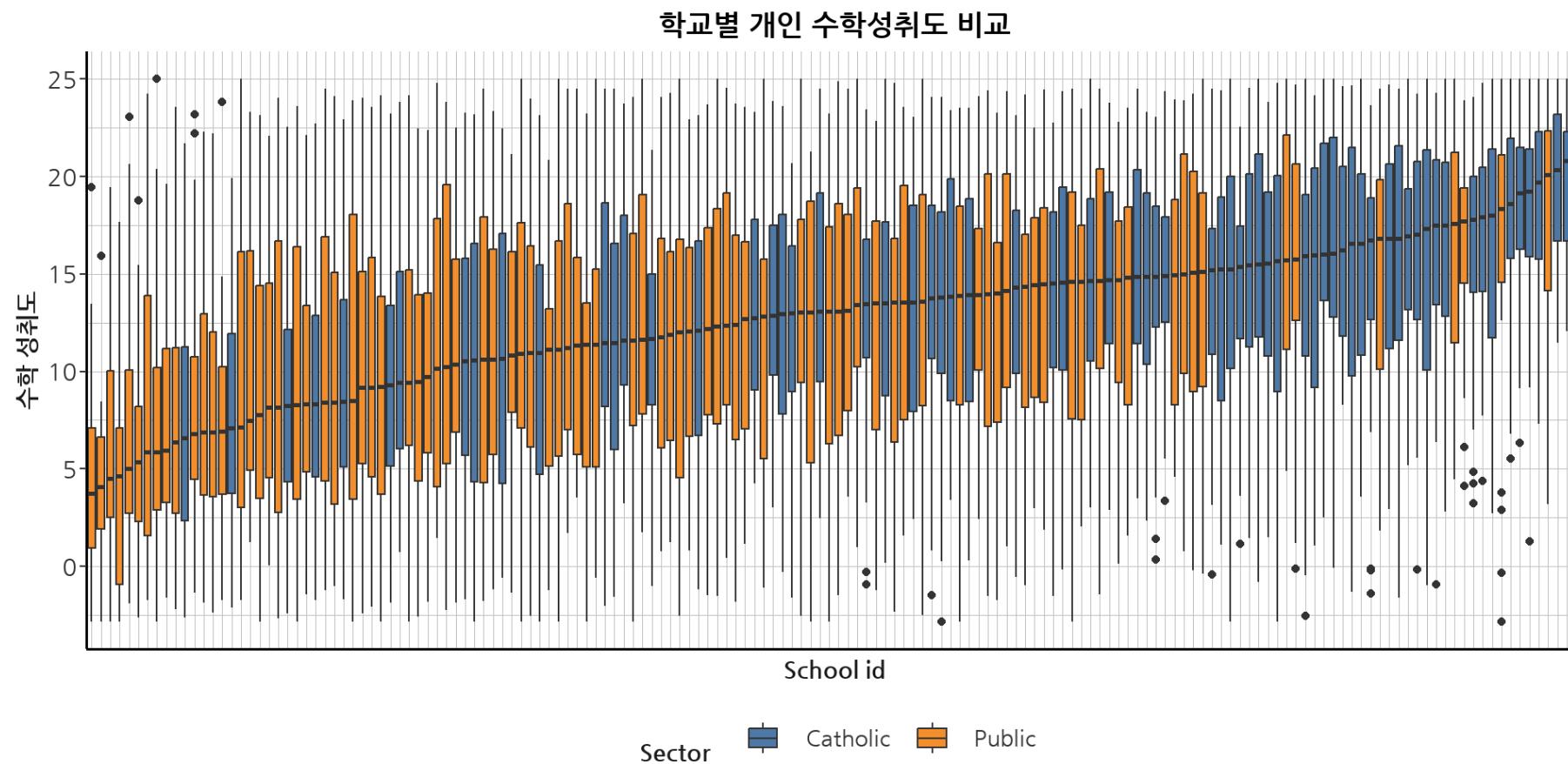
$$SST = SSA + SSE \quad \sum_{i=1}^a \sum_{j=1}^r (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^a r(\bar{Y}_{i\cdot} - \bar{Y})^2 + \sum_{i=1}^a \sum_{j=1}^r (Y_{ij} - \bar{Y}_{i\cdot})^2$$

- SSA (sum of squares of treatment): 집단변량 제곱합 - 집단수준 변동
- SSE (sum of squares of error): 오차제곱합 - 집단 내 변동
- 집단간 평균 차이가 존재할 때 HLM의 가장 최소한의 근거가 됨

```
1 model1 <- aov(mathach~id, data=data_merged)
2 anova(model1)
```

term	df	sumsq	meansq	statistic	p.value
id	159	64906.96	408.220	10.429	0
Residuals	7025	274969.98	39.142		

# Model 0. Preliminary Analysis: Visualization



# Model 1. One-way ANOVA

$$Level 1 : Y_{ij} = \beta_{0j} + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

$$Level 2 : u_{0j} = \gamma_{00} + u_{0j}, \quad u_{0j} \sim N(0, \tau_{00})$$

- 모형에 아무런 predictor variable들이 투입되지 않음.

```
1 modell1 <- lmer(mathach ~ 1 + (1 | id), data=data_merged)
2 summary(modell1)
```

Linear mixed model fit by REML. t-tests use Satterthwaite's method [  
lmerModLmerTest]

Formula: mathach ~ 1 + (1 | id)

Data: data\_merged

REML criterion at convergence: 47116.8

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.0631	-0.7539	0.0267	0.7606	2.7426

Random effects:

Groups	Name	Variance	Std.Dev.
id	(Intercept)	8.614	2.935
Residual		39.148	6.257

Number of obs: 7185, groups: id, 160

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t )
(Intercept)	12.6370	0.2444	156.6473	51.71	<2e-16 ***

---

Signif. codes: 0 '\*\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
1 icc(modell1)
```

# Intraclass Correlation Coefficient

Adjusted ICC: 0.180  
Unadjusted ICC: 0.180

# Model 1. One-way ANOVA: Model Fit

```
1 HLM_summary(model1,test.rand = T, digits = 3)
```

```
Model Information:  
Formula: mathach ~ 1 + (1 | id)  
Level-1 Observations: N = 7185  
Level-2 Groups/Clusters: id, 160  
  
Model Fit:  
AIC = 47122.793  
BIC = 47143.433  
R_(m)^2 = 0.00000 (Marginal R^2: fixed effects)  
R_(c)^2 = 0.18035 (Conditional R^2: fixed + random effects)  
Omega^2 = 0.18903 (= 1 - proportion of unexplained variance)
```

통계 모델간의 적합성을 비교하고 확인하기 위해서는 다음과 같은 기준들이 사용됨

- $k$ =투입되는 변수의 갯수,  $n$ =데이터의 갯수

다음 세가지 모형은 0과 가까워질 수록 해당 모형의 Model Fit이 좋아짐

- $Deviance = -2 * \ln(\text{Likelihood})$
- $AIC = -2 * \ln(\text{Likelihood}) + 2 * k$
- $BIC = -2 * \ln(\text{Likelihood}) + k * \log(n)$

변수가 많은 모형일수록 우도는 자연스럽게 0과 가까워지기에, AIC와 BIC는 Overfitting 문제해결과 모형 Parsimony를 위해 독립변수가 증가하는 것에 대한 패널티를 부여하여 모형의 품질을 평가. 이후, LRtest를 활용하여 모형간 차이를 검정하여 변수의 투입으로 인한 model fit 개선효과를 확인하기도 함.

- $R^2(m) = \text{Pooled OLS의 } R^2 \text{ 과 동일}$
- $R^2(c) = \text{Pooled OLS의 } R^2 \text{ 에 random effect의 효과}$
- $\Omega^2 = \frac{SS_{\text{effect}} - (df_{\text{effect}})(MS_{\text{error}})}{MS_{\text{error}} + SS_{\text{total}}}$

how much variance in the response variables are accounted for by the explanatory variables

# Model 1. One-way ANOVA: 계수해석

```
1 HLM_summary(model1,test.rand = T, digits = 3)
```

Fixed Effects:  
Unstandardized Coefficients (b or  $\gamma$ ):  
Outcome Variable: mathach

b/ $\gamma$	S.E.	t	df	p	[95% CI of $b/\gamma$ ]
(Intercept)	12.637	(0.244)	51.71	156.6	<.001 *** [12.154, 13.120]

'df' is estimated by Satterthwaite approximation.

Random Effects:

Cluster	K	Parameter	Variance	ICC
id	160	(Intercept)	8.61402	0.18035
		Residual	39.14832	

ANOVA-like table for random-effects: Single term deletions

Model:  
mathach ~ (1 | id)  
npar logLik AIC LRT Df Pr(>Chisq)  
<none> 3 -23558 47123  
(1 | id) 2 -24052 48107 986.12 1 < 2.2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' '

추정된 고정효과 모수: 평균 수학성취도

- $\gamma_{00} = 12.637$

추정된 임의효과 모수:

- Level 1 (개인수준): 평균적으로 동일 학교 내에서 학생들의 수학성취도가 어느정도 차이가 있는가?

$$\hat{var}(e_{ij}) = \hat{\sigma}^2 = 39.14832$$

- Level 2 (집단수준): 학교 간 수학성취도의 평균이 서로 얼마나 다른지?

$$\hat{var}(u_{0j}) = \hat{var}(\beta_{0j}|\gamma_{00}) = \hat{\tau} = 8.61402$$

집단수준 분산 ( $\tau$ ; 학교간 차이) 검정 (Significance test for the intercept variance):

- Absolute Null Model ( $\tau_{00}$ )만 투입된 모형인 Pooled-OLS모형과 HLM 모형의 비교
- $\chi^2$  검정 (Likelihood-Ratio Test)을 통해 Null Model 대비 p-value 값이 2.2e-16로 매우 작게 통계적으로 유의한 것으로 나타남

# Model 1. One-way ANOVA: ICC

```
1 HLM_ICC_rWG(data_merged, group="id", icc.var="mathach")
```

----- Sample Size Information -----

Level 1: N = 7185 observations ("mathach")  
Level 2: K = 160 groups ("id")

n (group sizes)  
Min. 14.00000  
Median 47.00000  
Mean 44.90625  
Max. 67.00000

----- ICC(1), ICC(2), and rWG -----

ICC variable: "mathach"

ICC(1) = 0.180 (non-independence of data)  
ICC(2) = 0.901 (reliability of group means)

rWG variable: "mathach"

rWG (within-group agreement for single-item measures)

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
rWG	0.000	0.270	0.381	0.386	0.515	0.806

# Model 1. One-way ANOVA: ICC (Cont'd)

Intraclass correlation (ICC):

- ICC(1)은 전체 관찰 분산에서 집단 간 분산이 차지하는 비율 (강상진, 2016)
- 전체 분산(Level-2 분산과 Level-1 또는 잔차 분산의 합)에 대한 Level-2 분산(집단 평균의 분산)의 비율이 클수록, 집단 간(between)의 유사성보다 집단 내부(within)간의 유사성이 큼을 의미

$$ICC(1) = \frac{\text{학교 간 분산}}{\text{전체 관찰분산}} = \frac{Var(\beta_{0j})}{Var(Y_{ij})} = \frac{\tau}{\sigma^2 + \tau} = \frac{8.614}{39.148 + 8.614} = 0.18035$$

- 만약 ICC(1)의 값이 0이라면 한 집단에 속한 응답치들 간의 유사성이 다른집단에 속한 응답치들과 보이는 유사성과 다르지 않음 (일반적으로 0.05~0.25 정도)
- ICC(1)=0.5 represents a small to medium effect (LeBreton and Senter, 2008)
- 그러나 통일된 기준은 존재하지 않으며, 이론적으로 집단을 고려함을 통해 설명되는 정도가 어느정도인지를 이해하는 것이 중요함. ICC 값이 매우 작아 0에 가깝더라도 측정값과 다른 측정값 사이의 관계가 모든 집단에서 동일하다는 것을 의미하지 않음 (Nezlek, 2008)
- Simulated situation only 1% of the variance is attributed to group membership ICC(1)=.01) and, still, strong group-level relationships were detected (Bilese, 1998)
- ICC(2)는 집단간 평균의 신뢰도를 측정하기 위함이며, 집단별 평균(학교별 수학성취도 평균)은  $\beta_{0j}$  표본에 따라 그 값이 달라지기 때문에 통계적 추정의 차원에서는 의미가 없으나, 잔차분석에서 제공하는  $\beta_{0j}$ 가 어느정도 신뢰로운 값인지 알려줌.  $\beta_{0j}$ 가 높으면 학교 정보로서의 가치가 높고, 낮으면  $\beta_{0j}$ 에 의한 평가가 위험함
- ICC(2) <0.40 are poor, those from 0.40 to 0.75 are fair to good, and those >0.75 are excellent (Fleiss, 1986)

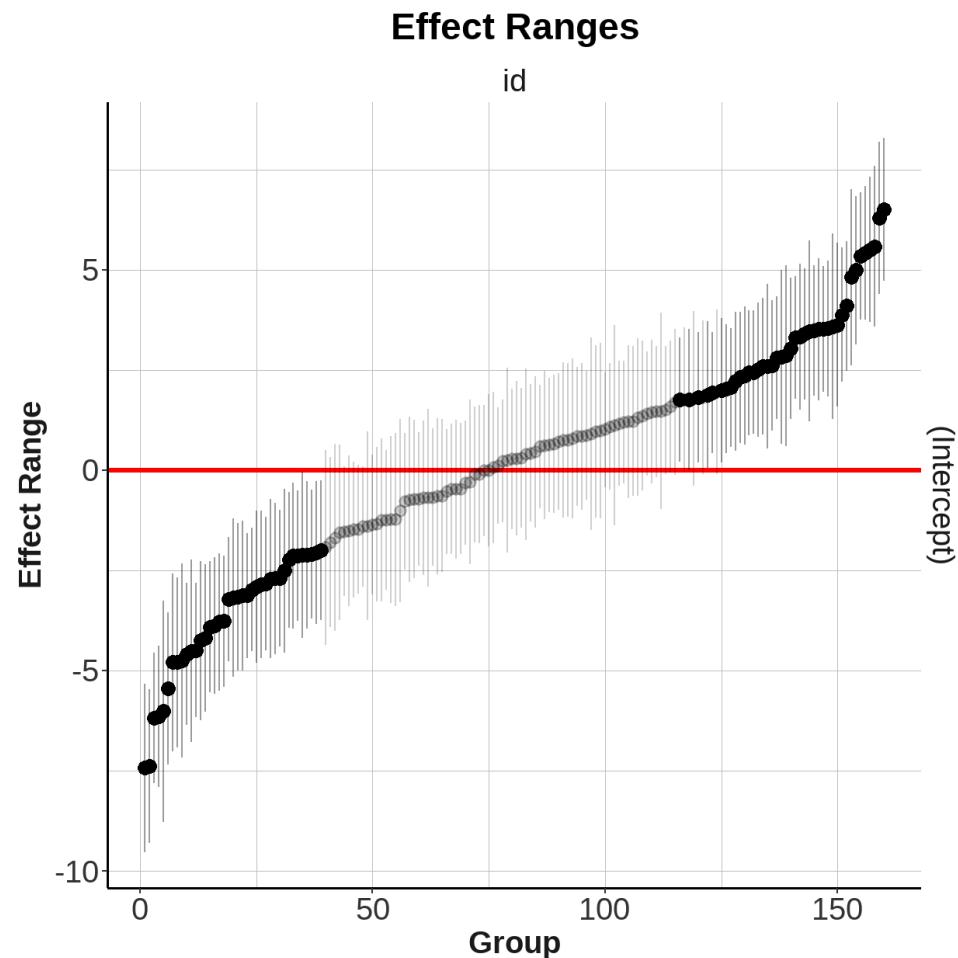
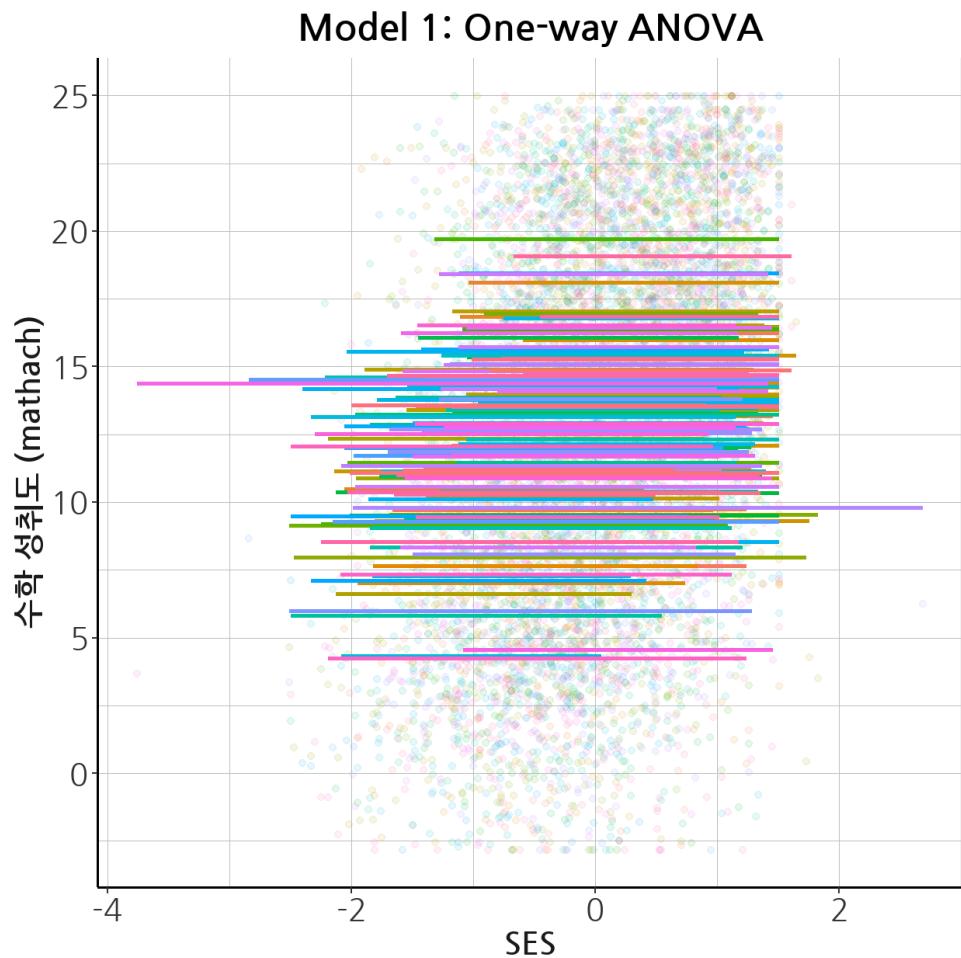
# Model 1. One-way ANOVA: ICC (Cont'd)

Intraclass correlation (ICC) (참고):

$$ICC(2) = \frac{Var(\text{진 점수})}{Var(Y)} = \frac{\tau_{00}}{\tau_{00} + \sigma^2/n_j}$$

- This number indicates whether estimated differences across schools are reliable indicators of real differences among schools' population means
- $\tau_{00}$ 이 크거나 각 학교의 표본이 크면 이 값은 커지게 됨
- 일반적으로 무선효과의 추정은 Random Level-1 coefficients에 대해 이 Level-1 모형의  $Y_{.j}$ 와 Level-2 모형의  $\hat{\gamma}_{00}$ 를 동시에 고려하는 추정치 (a weighted combination (WLS), known as a Bayes estimator)
- HLM 은  $Y_{.j}$  의 신뢰도가 높으면  $Y_{.j}$  가 더 많이 가중되고 그 신뢰도가 낮으면 Level-2 모형에서 얻어지는  $\hat{\gamma}_{00}$  값에 더 많은 가중치를 주는 방식으로  $\beta_{0j}^*$  를 추정한다. 이러한 이유로  $\beta_{0j}^*$ 은 전체 평균 (grand mean)  $\gamma_{00}$ 로 집약되는 모습을 보여서 shrinkage estimator 라고 불린다.

# Model 1. One-way ANOVA: Visualization



# Model 3. One-Way ANCOVA

```
1 g
```

```
Error in eval(expr, envir, enclos): object 'g' not found
```

# What about tables?

## knitr::kable()

```
1 tab <- starwars |>
2   tidyr::drop_na(species) |>
3   group_by(species) |>
4   summarise(
5     n = n(),
6     mean_heighth = round(mean(height, na.rm = TRUE)),
7     mean_mass = round(mean(mass, na.rm = TRUE))
8   ) |>
9   slice_max(order_by = n, n = 4)
10
11 knitr::kable(tab)
```

species	n	mean_height	mean_mass
Human	35	177	83
Droid	6	131	70
Gungan	3	209	74
Kaminoan	2	221	88
Mirialan	2	168	53
Twi'lek	2	179	55
Wookiee	2	231	124
Zabrak	2	173	80

# DT::datatable()

With the `smaller` class in the slide! Ex: `## slide name {.smaller}`

Show 5 ▾ entries

Search:

	species	n	mean_height	mean_mass
1	Human	35	177	83
2	Droid	6	131	70
3	Gungan	3	209	74
4	Kaminoan	2	221	88
5	Mirialan	2	168	53

Showing 1 to 5 of 8 entries

Previous

1

2

Next

# gt::gt()

species	n	mean_heigth	mean_mass
Human	35	177	83
Droid	6	131	70
Gungan	3	209	74
Kaminoan	2	221	88
Mirialan	2	168	53
Twi'lek	2	179	55
Wookiee	2	231	124
Zabrak	2	173	80

# reactable::reactable()

species	n	mean_heighth	mean_mass
Human	35	177	83
Droid	6	131	70
Gungan	3	209	74
Kaminoan	2	221	88
Mirialan	2	168	53
Twi'lek	2	179	55
Wookiee	2	231	124
Zabrak	2	173	80

# Diagrams with Mermaid!

Read about how to create a diagram in this post by Mine Çetinkaya-Rundel.

# Exporting into PDF

You can use the function `pagedown::chrome_print()` to print the HTML version into a PDF!

```
1 pagedown::chrome_print("path-to-file.html")
```

