

# 증거기반연구 6차 세미나

Week 3: HLM (위계적 선형 모형)

주상원

서울대학교 행정대학원 석사과정

02 Feb 2023

# 열심히 해봅시다... 영자영자...

## Going to Grad School

What my friends think I do



What my mom thinks I do



What society thinks I do



What my advisor thinks I do



What I think I do



What I really do



# 목차

## i. 이론

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

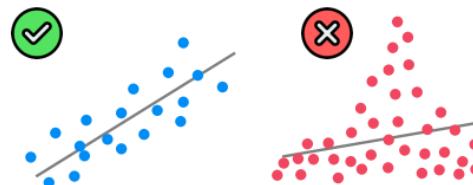
## ii. 실습: R과 Stata로 실제 데이터 분석

# Module I: 위계적 선형모형의 배경

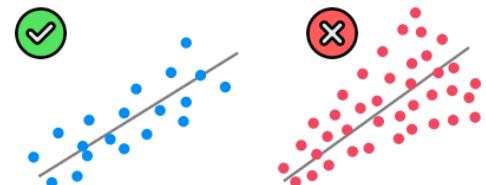
# 회귀분석의 기본 가정s

- Gauss-Markov Theorem: 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](#)

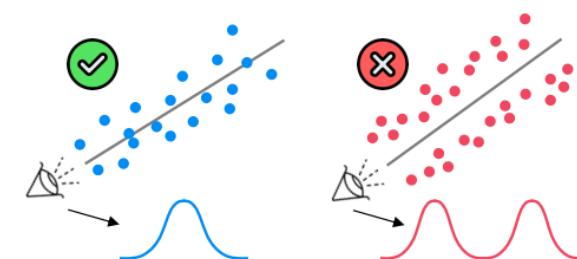
1. Linearity  
(Linear relationship between Y and each X)



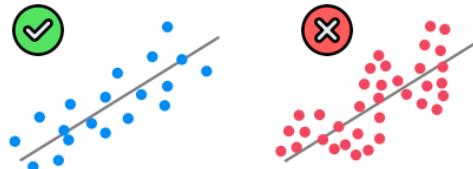
2. Homoscedasticity  
(Equal variance)



3. Multivariate Normality  
(Normality of error distribution)



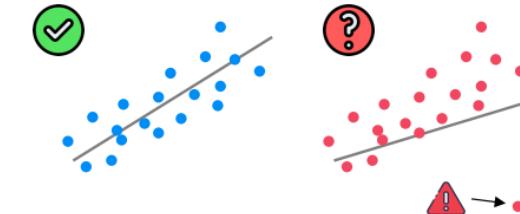
4. Independence  
(of observations. Includes "no autocorrelation")



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

$X_1 \not\sim X_2$        $X_1 \sim X_2$

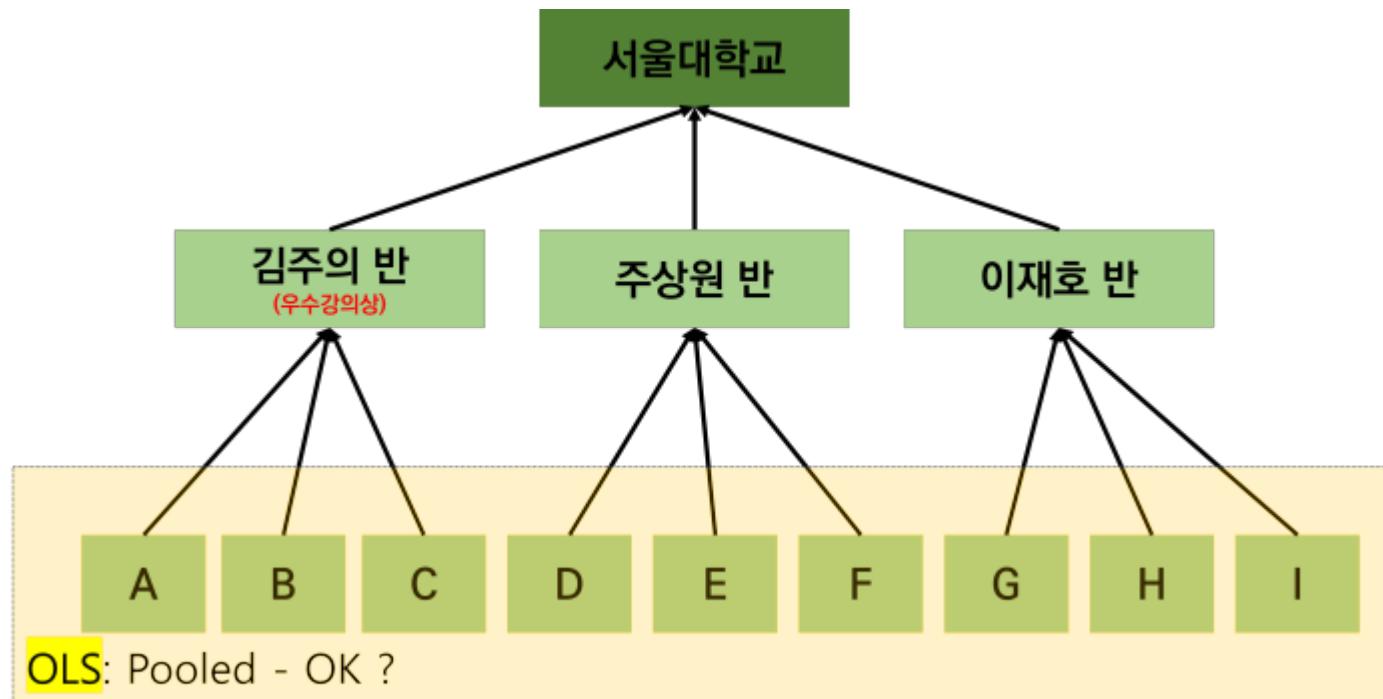
6. The Outlier Check  
(This is not an assumption, but an "extra")



Assumptions

# 그러나 현실은...? 특히 조직연구에서는..?

- Nested Data: 학교 > 학년 > 반 > 개인
- 조직수준의 변수가 조직 구성원의 개인 변수에 영향을 미치는 것이 일반적



# 위계적 대상자료

- 위계구조(hierarchical structure)
  - 하나의 단위가 그보다 상위 수준(level)에 속해 있는 구조
  - 데이터의 위계는 연구자의 목적에 따라 여러 단계로 계층화 할 수도 있고, 여러 수준으로 구 분하여 구축할 수 있음



# 위계적 대상자료 (Cont'd)

- 학생들은 각 학교에 내재한 구조
  - 동일 반에 속한 학생들은 학교의 문화, 학습환경, 친구관계, 교사 등 수많은 요인들을 공유
- 상호의존성
  - 개인을 둘러싼 지역의 경제, 사회, 문화, 제도 및 물리적 특성이 개개인의 행동, 선호도, 가치관 등에 영향을 미치기 때문
  - 관찰단위인 학생들은 동일 학교, 동일 반 내에서는 상호의존성을 가지게 되고, 소속이 다르면 독립성을 갖음 → 독립성 가정이 결과적으로 위배되게 됨
- 복수의 Unit of Analysis
  - 학생수준에서의 변수 (Level 1), 학교수준에서의 변수 (Level 2)
- 불균형 자료
  - 각 학교별로 학생이 다르기에 관측치는 일반적으로 같을 수 없음

# 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

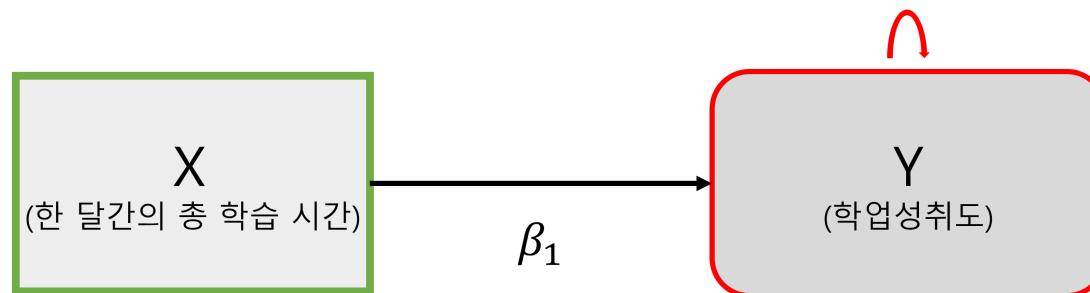
# 독립성 (Independence) 가정의 위배 (Cont'd)

- 사례 분석: 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

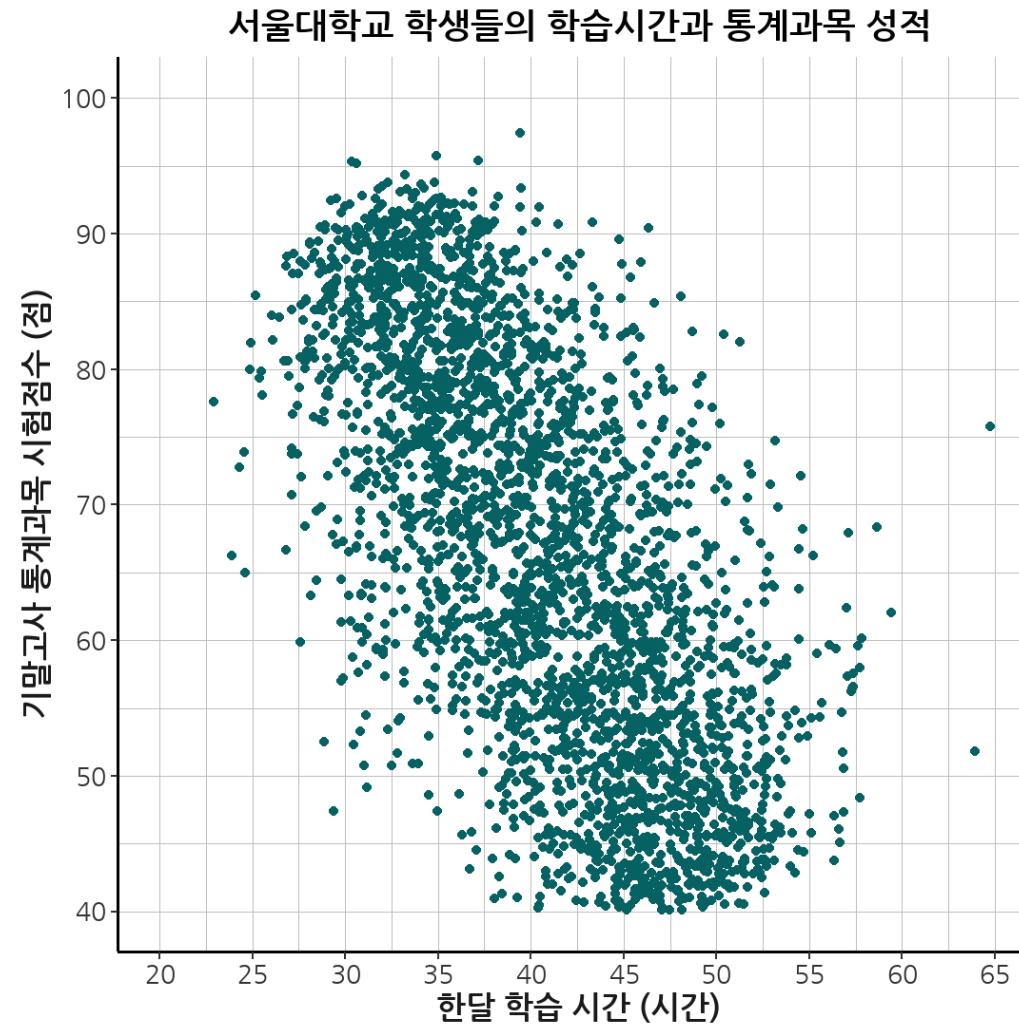
E(오차, 기타요인들)



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

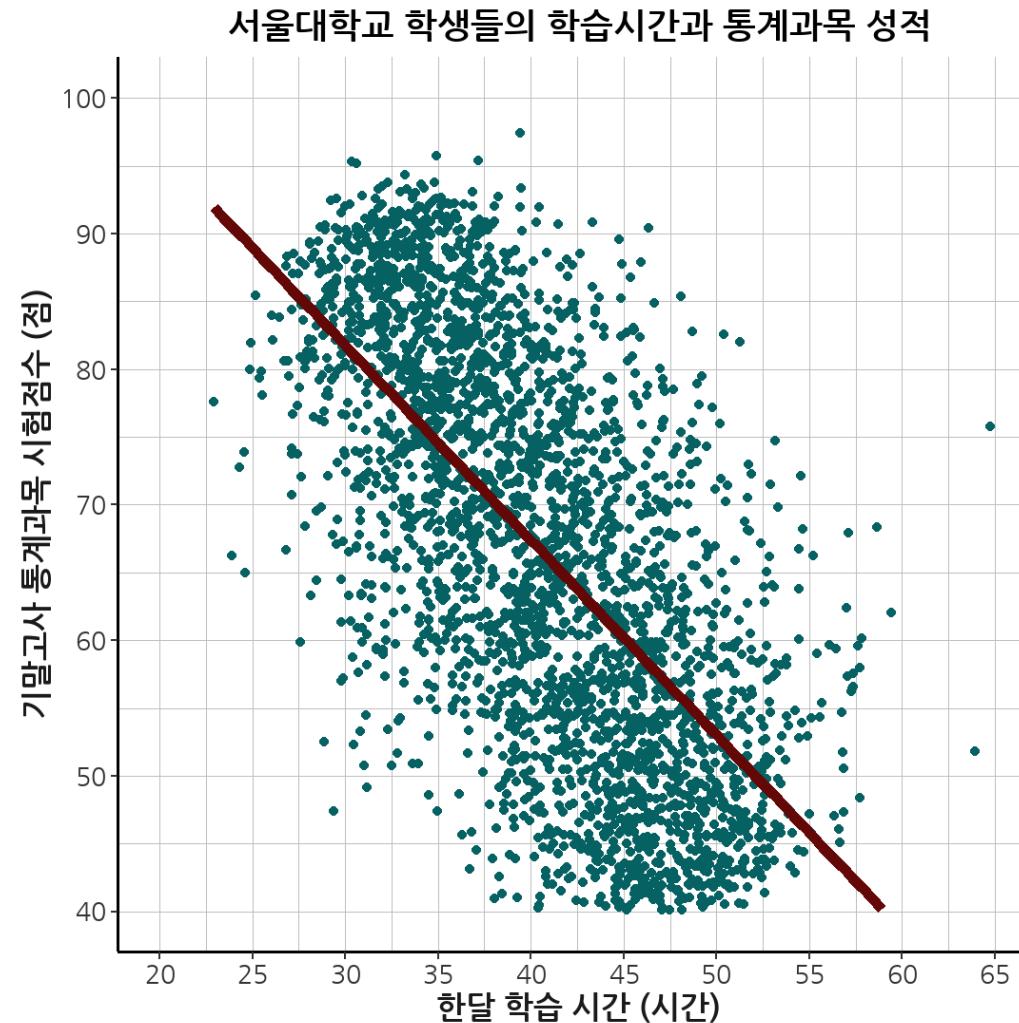
# 독립성 (Independence) 가정의 위배 (Cont'd)

- 시각화



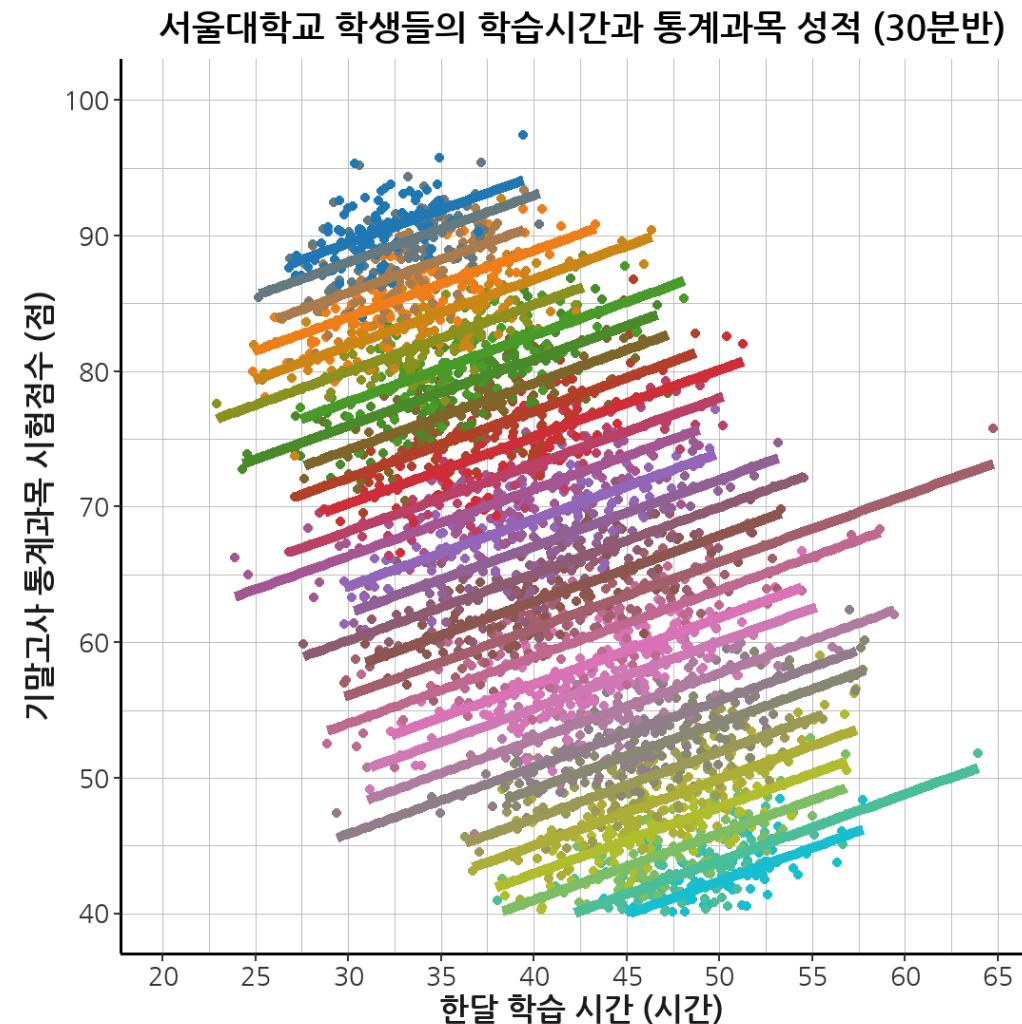
# 독립성 (Independence) 가정의 위배 (Cont'd)

- 시각화



# 독립성 (Independence) 가정의 위배 (Cont'd)

- 시각화



# 독립성 (Independence) 가정의 위배 (Cont'd)

- Pooled OLS

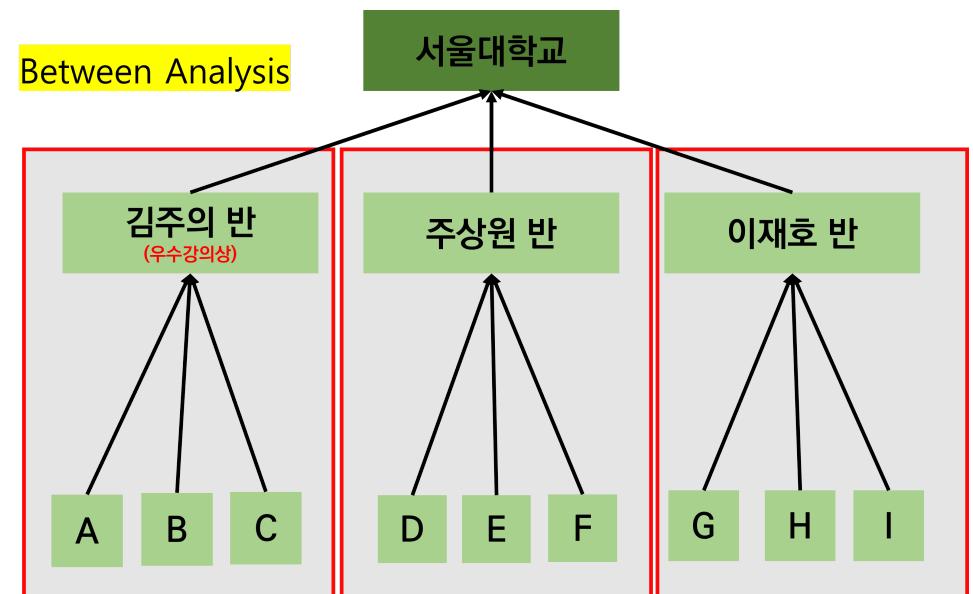
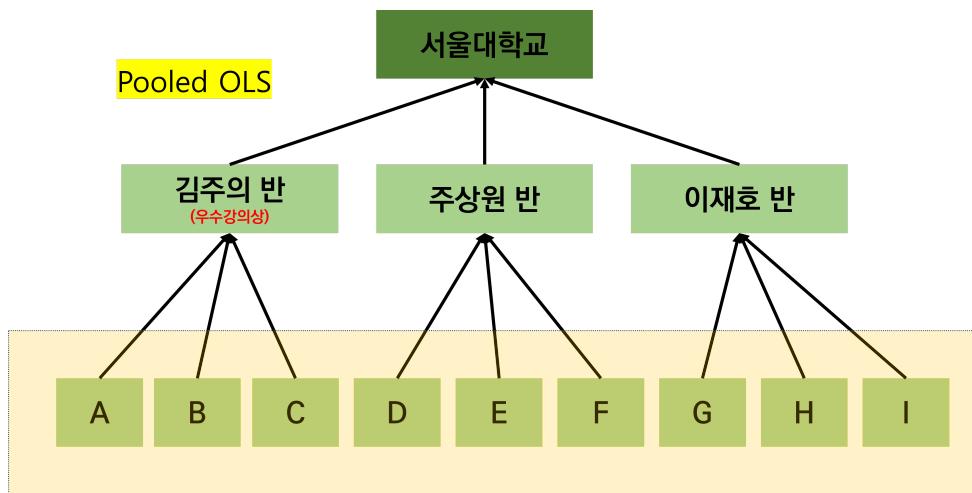
term	estimate	std.error	statistic	p.value
(Intercept)	125.613	1.252	100.337	0
x	-1.460	0.030	-48.028	0

- 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, 개인주의적 오류(individualistic fallacy)
- **Aggregate?**: Between → waste information from indvs, sample ↓, 생태학적 오류(Ecological Fallacy)



# Module I: 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
  - 기존방법은 (1) 개인 간의 완벽한 독립성이 존재하기 어렵고, (2) 자료 자체가 nested되어 있다는 것을 고려하지 못함 → Individual 대상 연구에서 (특히, 조직맥락) OLS는 더이상 만능 X, 개인주의적, 생태학적 오류를 범할 수 있게 됨
- Pooled-OLS는 집단간 군집성 무시, 가우스 마르코브 가정 위배, 심슨의 패러독스, 개인주의적 오류, Between 모형은 개개인들의 특성이 고려되지 않는 문제, 생태학적 오류
- HLM의 필요성: within과 between을 동시에 다루고, level 1과 2를 동시에 고려

# Module II: 위계적 선형모형의 개념

# 위계적 선형모형 (HLM)

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

- 위계적 선형모형은 비군집화되거나 군집화된 접근의 한계를 극복하기 위해서 개발됨
1. 개인 수준과 집단 수준의 잔차를 모델링함을 통해서 결과적으로 OLS대비 동일한 그룹 내의 개인들 간의 부분적인 독립성을 확보함
  2. 낮은 수준의 단위와 높은 수준의 단위 내에 존재하는 분산을 동시에 고려하는 접근
- ∴ 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

# 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 이론

- OLS는 집단에 대한 고려없이 모든 개인들을 동일하게 고려

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

- HLM은 집단 내(개인:  $i$ )과 집단 간(집단:  $j$ ) 모형을 별개로 model specification 한다
  - Level 1:

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

- Level 2:

$$(Intercept) \beta_{0j} = \gamma_{00} + u_{0j}, \quad (Slope) \beta_{1j} = \gamma_{10} + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix}\right)$$

- **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 (강상진, 2016)

1. 우리나라 중학생 수학성취도의 학교 간 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

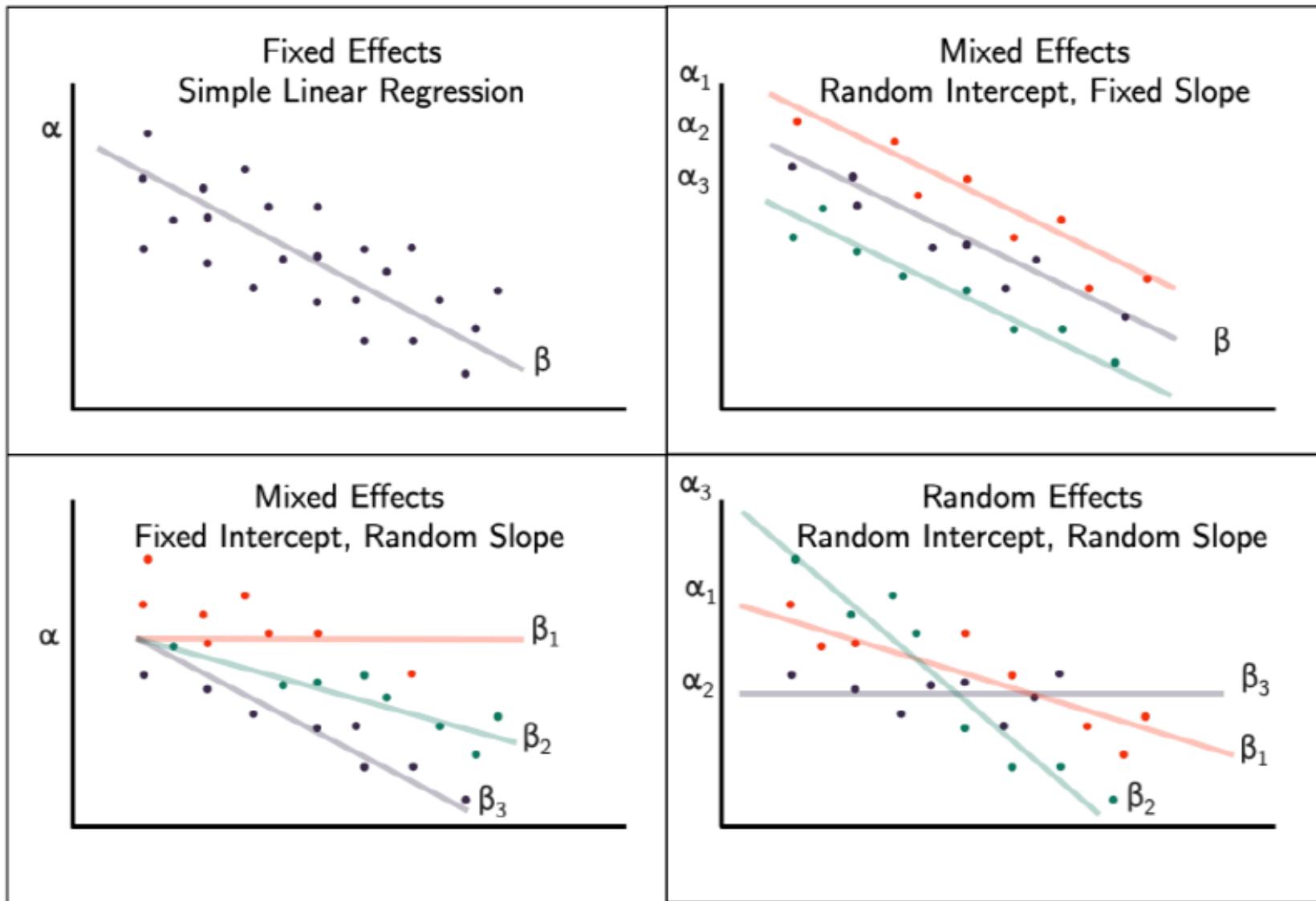
- **Fixed Effects:** 선형 회귀모델과 같이 절편 및 기울기의 추정계수가 집단에 따라 변화하지 않고 단 하나의 값을 가짐
- **Random Effect:** 절편 및 기울기의 추정계수가 하나의 값이 아니라 상위 수준인 집단의 특성에 따라 여러 개의 값을 가짐 (e.g.,  $\beta_{0j}$  and  $\beta_{1j}$ ).
- Equation

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

- 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



## Module II: Sum-up

- HLM은 서로 다른 수준의 분석단위를 하나의 모델에 포함시켜 하위 수준과 상위 수준의 모수를 동시에 추정 가능하도록 하는 통계 방법
- 각 개인은 그가 속한 지역이나 집단의 특성으로 영향을 받고 있으며, 특정 조직 또는 집단에 속한 개인들은 그와 다른 집단이나 지역에 속한 개인들과 구별되는 공통 특성 가짐
  - 종속변수: 개인 수준에서 측정
  - 독립변수: 하위수준(개인) & 상위수준(집합단위)
  - 선형 모형이기에 변수들 사이의 선형관계를 가정
- 어떤 변수들이 어떤 수준에 속하는지 지정 가능, 수준 간 교차 및 상호작용효과 고려 가능
- 기존 회귀분석 방법과 달리 잔차의 독립성에서 자유롭고, 분산을 수준별로 산출 가능

# Module III: 위계적 선형모형의 구체화

# Model Specification with R

- R 설치과정 참고: <https://www.youtube.com/watch?v=LVkhk4MXQAg&t>
- 기초 R 연습 참고 사이트: <https://www.youtube.com/watch?v=jLcDVcgQpPI&list=PLKtLBdGReMmww86INhCWxJNfwBUcnxgZ6>
- 기초 R 문서 버전: [R for Data Science](#)
- 사용하는 패키지: lmerTest, lme4, bruceR

# 데이터 불러오기

```
1 # install.packages("pacman")
2 pacman::p_load("tidyverse", "magrittr", # data분석 필수 패키지
3                  "broom", # 분석결과 정리해주는 패키지
4                  "lmerTest", "lme4", "merTools", "nlme", # HLM 패키지
5                  "bruceR", # HLM 결과 정리 표 생성
6                  "readstata13", # statafile (.dta) 열어주는 패키지
7                  "skimr", "psych", # 요약표, 심리학 분석
8                  "plm") # 패널회귀분석 (within, random)
9 ## 자신이 설정하고 싶은 곳으로 설정
10 setwd("E:/OneDrive - SNU/(B) 대학원/세미나/HLM"); getwd()
11
12 # Read data
13 data_lv1 <- read.dta13("./HSB1.dta"); data_lv2 <- read.dta13("./HSB2.dta")
```

- 정상적으로 로드 되었는지 확인

```
1 # Size
2 dim(data_lv1); dim(data_lv2)
```

```
[1] 7185      5
```

```
[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", "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, 1, 0, 0, 1, 0, 0, 0, 0, 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\_1v1 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\_1v2 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

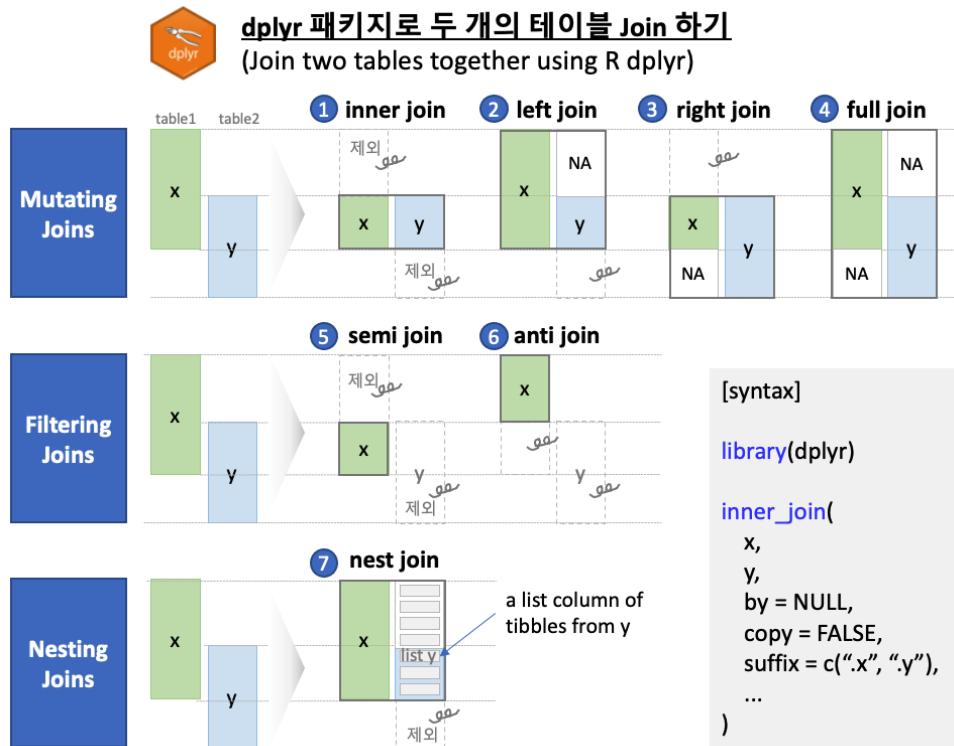
# Data Merge: dplyr package - joins

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.

- `dplyr` join 함수들의 유형
  - `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

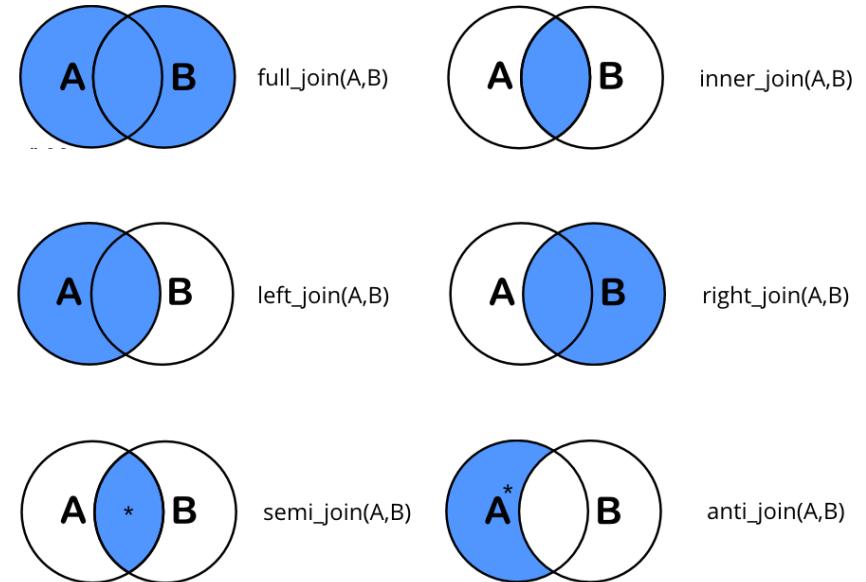
```
1 left_join(  
2   x, # Level 1 data-set name  
3   y, # Level 2 data-set name  
4   by = c("id_x" = "id_y"), # 각각의 데이터 셋에서 어떤 변수를 기준으로 merge되는지 설정  
5   copy = FALSE, # 가만히 두기  
6   suffix = c(".x", ".y"), # 만약 id 이외에 서로 겹치는 변수가 있을때 어느 데이터셋인지  
7   ...  
8   keep = FALSE # 가만히 두기  
9 )
```

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



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

출처



# Data Merge: dplyr package - binds

- bind\_rows()
  - 다수의 데이터 프레임(티블)을 행 기준으로 합치기 (binding multiple data frames by row)
- bind\_cols()
  - 다수의 데이터 프레임(티블)을 열 기준으로 합치기 (binding multiple data frames by columns)
- bind 함수의 구조

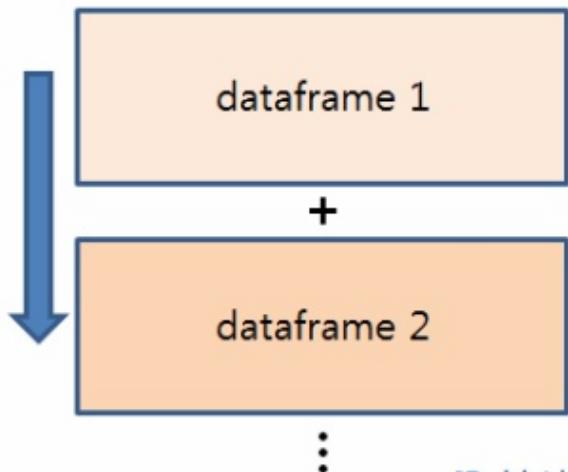
```
1 # row 병합
2 bind_rows(list(data*, data*, ... ))
3
4 ## id의 경우 어떤 data frame, tibble에서 결합되었는지 확인 용도
5 bind_rows(list(a = one, b = two), .id = "id")
6
7 # column 병합
8 bind_cols(list(data*, data*, ... ), .name_repair = c("unique", "universal"))
```

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

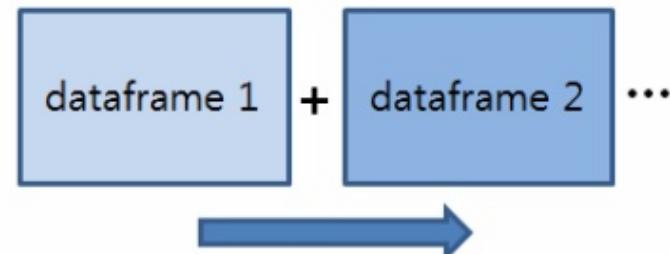
**bind** in *{dplyr}* package

: Efficiently bind multiple data frames by row and column

**bind\_rows(df1, df2)**



**bind\_cols(df1, df2)**



[R 분석과 프로그래밍] <http://rfriend.tistory.com>

출처

# Data Merge: data\_lv1 and data\_lv2

Pipe Operator: `%>%`

- (`lhs %>% rhs`)는 lhs의 결과를 rhs의 첫번째 변수로 넘겨주는 역할. '.'은 앞의 값의 위치를 구체적으로 지정하기 위해 사용 (단축키: `ctrl + shift + m` (mac: `cmd + shift + m`))
  - `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)
```

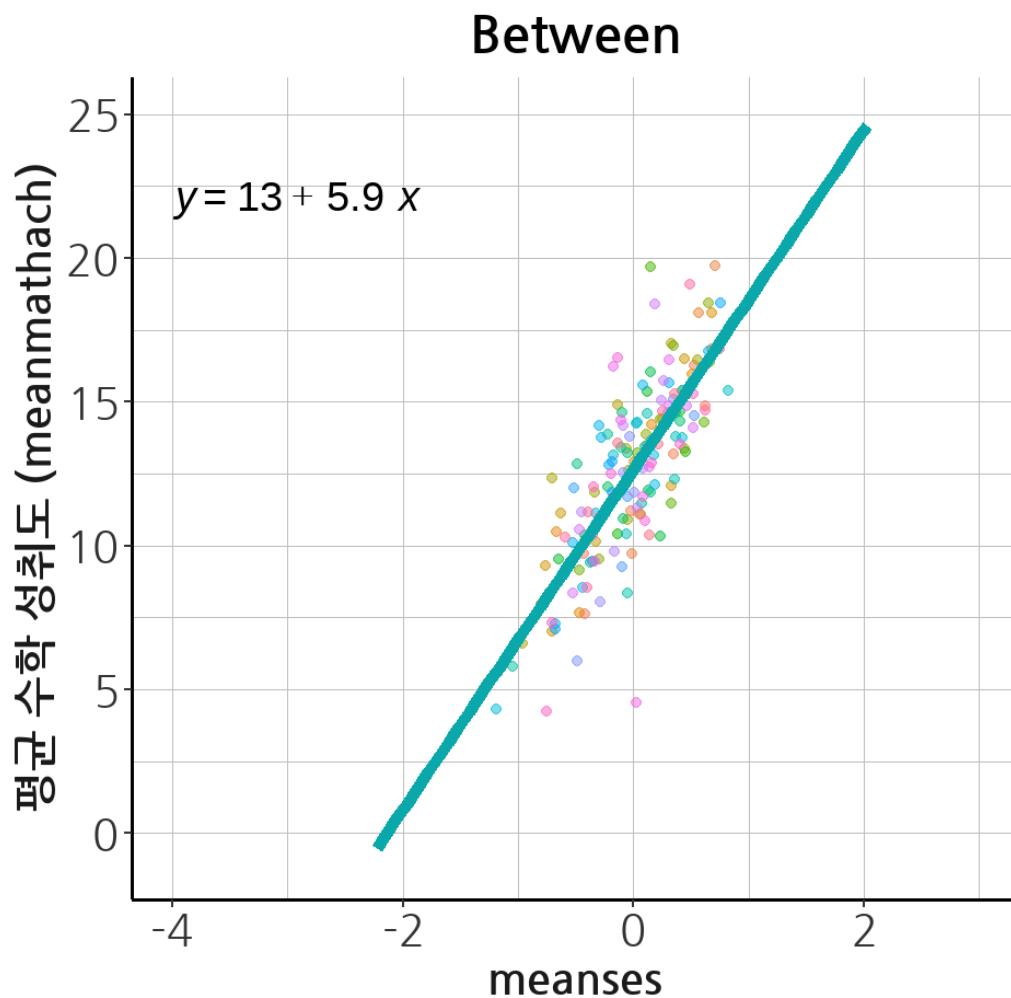
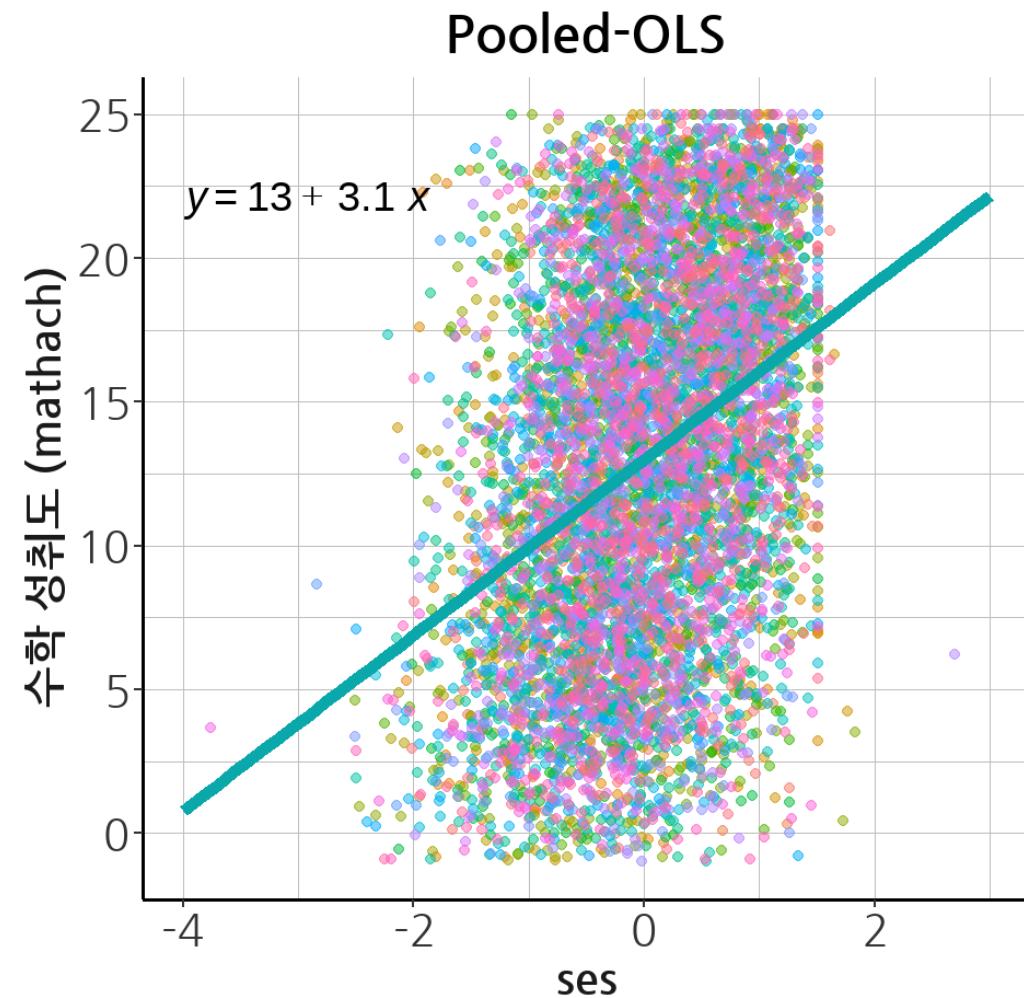
<b>id</b>	<b>minority</b>	<b>female</b>	<b>ses</b>	<b>mathach</b>	<b>size</b>	<b>sector</b>	<b>himinty</b>
1224	0	1	-1.53	5.88	842	0	0
1224	0	1	-0.59	19.71	842	0	0
1224	0	0	-0.53	20.35	842	0	0
1224	0	0	-0.67	8.78	842	0	0
1224	0	0	-0.16	17.90	842	0	0

# Data Pivot: pivot\_longer and pivot\_wider

- **Long Form:** 긴 형태의 데이터는 열로 표현된 변수들을 행 방향으로 풀어 넣음으로써 열의 개수는 줄고 행의 개수는 늘어나는 형태의 데이터
  - 긴 형태의 데이터는 데이터 시각화, 패널회귀분석에 매우 적합한 형태나 직관성이 떨어짐
- **Wide Form:** 넓은 형태의 데이터는 행로 표현된 변수들을 열 방향으로 풀어 넣음으로써 행의 개수는 줄고 열의 개수는 늘어나는 형태의 데이터
  - 사용자가 직관적으로 데이터의 전반적 분포를 살펴보기가 좋지만, 활용이 어려움

```
1 # pivot_longer
2 data("billboard", package = "tidyverse")
3 billboard %>%
4   pivot_longer(cols = wk1:wk76, # starts_with('wk')
5                 names_to = "week",
6                 values_to = "rank",
7                 values_drop_na = T)
8 billboard %>%
9   pivot_longer(cols = wk1:wk76, # starts_with('wk')
10                names_to = "week",
11                values_to = "rank",
12                values_drop_na = T,
13                names_prefix = "wk",
14                names_transform = as.integer)
15
16 # pivot_wider
17 data("fish_encounters", package = "tidyverse")
18 fish_encounters %>%
19   pivot_wider(names_from = "station", values_from = "seen" )
```

# Pooled-OLS and Between Model



# Five models in HLM (Bryk & Raudenbush, 1992)

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
  - $Y_{ij}$ :  $j$  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, 개인 수준의 변량대비 집단 수준의 변량 비교

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

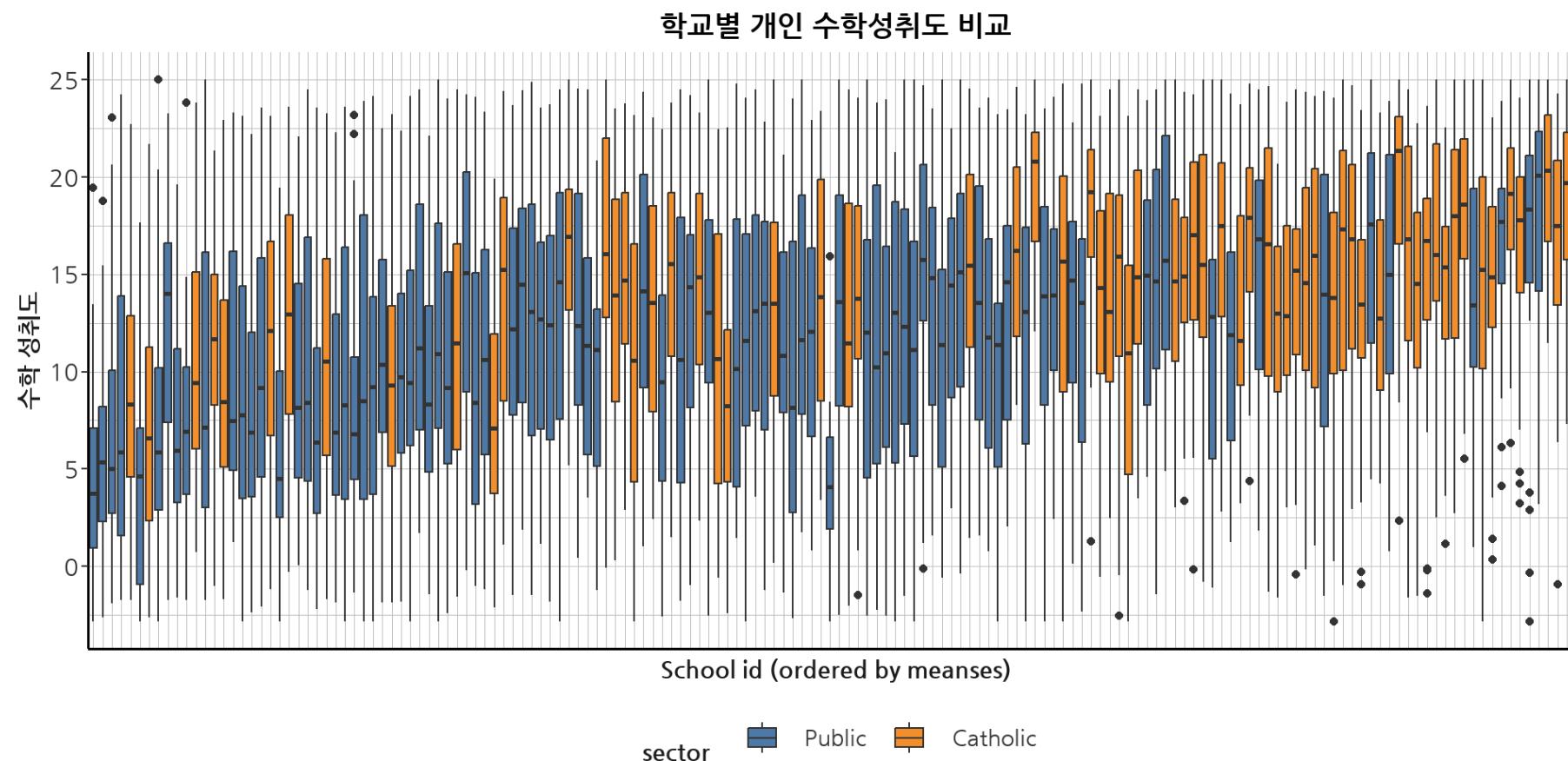
$$SST = SSA + SSE \quad \sum_{i=1}^a \sum_{j=1}^r (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^a r(\bar{Y}_{i\cdot} - \bar{Y})^2 + \sum_{i=1}^a \sum_{j=1}^r (Y_{ij} - \bar{Y}_{i\cdot})^2$$

- 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:  $Y_{ij} = \beta_{0j} + \epsilon_{ij}$     $\epsilon_{ij} \sim N(0, \sigma^2)$
- Level 2:  $\beta_{0j} = \gamma_{00} + u_{0j}$ ,    $u_{0j} \sim N(0, \tau_{00})$
- 급간상관계수(ICC)와 신뢰도 계산이 주 목적으로 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
```

# 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 R2: fixed effects)  
 $R_c^2 = 0.18035$  (Conditional R2: fixed + random effects)  
 $\Omega^2 = 0.18903$  (= 1 - proportion of unexplained variance)
```

## 통계 모델간의 적합성 비교 기준

- $k$ =투입되는 변수의 갯수,  $n$ =데이터의 갯수
- 다음 세가지 기준은 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 우도가 0과 가까워지기에, AIC 와 BIC는 Overfitting 문제해결과 모형 Parsimony를 위해 독립변수 증가에 대한 패널티 부여. LRtest로 모형 간 차이를 검정하여 변수 투입으로 인한 model fit 개선효과를 확인.

- 다음 기준들은 클 수록 좋음
  - $R^2(m) = \text{Pooled OLS의 } R^2$  과 동일
  - $R^2(c) = \text{Pooled OLS의 } R^2$  에 random effect의 효과

# Model 1. One-way ANOVA: 계수해석

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

```
Fixed Effects:  
Unstandardized Coefficients (b or γ):  
Outcome Variable: mathach  
  
b/γ S.E. t df p [95% CI of b/γ]  
  
(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 ' ' 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:

- Null Model과 HLM 모형의 비교
- $\chi^2$  검정 (Likelihood-Ratio Test)을 통해 Null Model 대비  $p\text{-value}$  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)(1):

- ICC(1)은 전체 관찰 분산에서 집단 간 분산이 차지하는 비율 (강상진, 2016)
- ICC(1) is typically interpreted as a measure of effect size (Bliese, 2000; Bryk & Raudenbush, 1992)
- 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)
- The value of .01 might be considered a small effect, a value of .10 might be considered a medium effect, and a value of .25 might be considered a large effect (Murphy & Myors, 1998).
- 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)

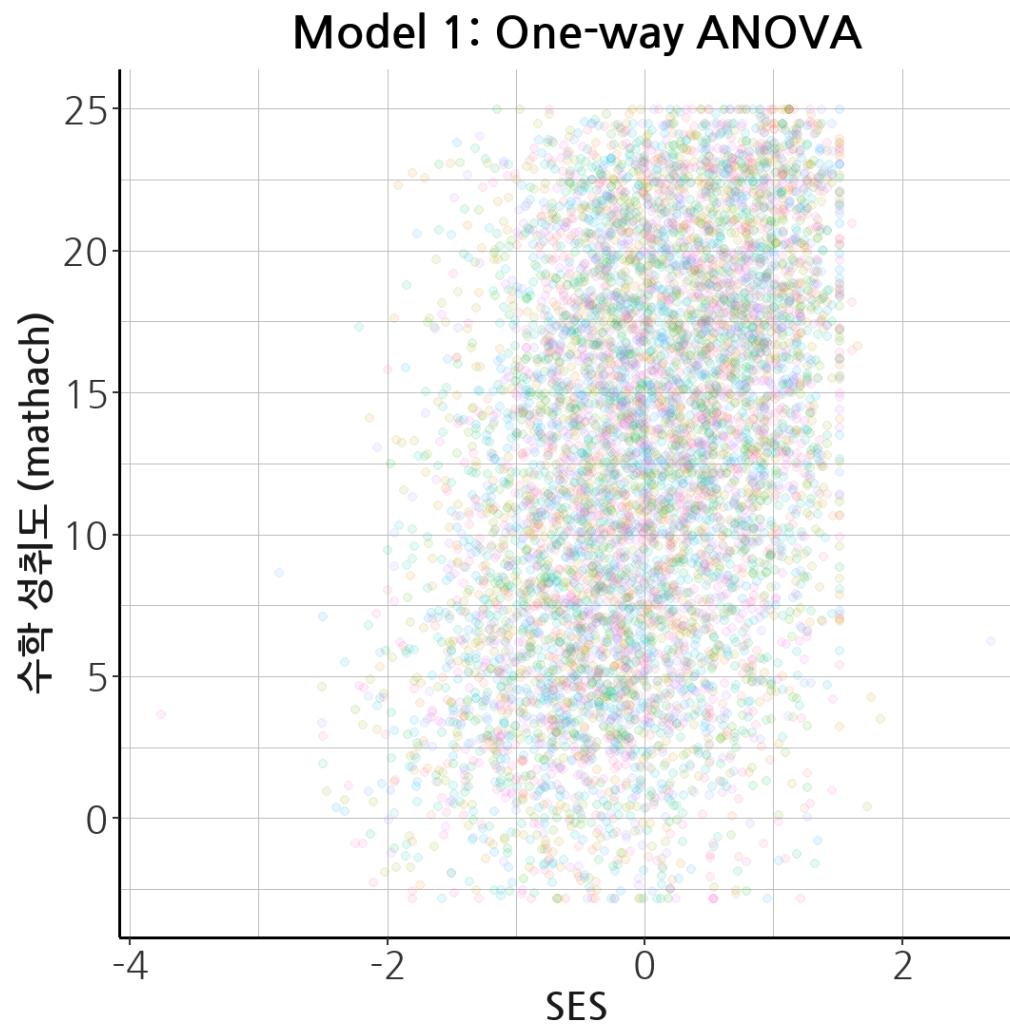
# Model 1. One-way ANOVA: ICC (Cont'd)

Intraclass correlation (ICC)(2) :

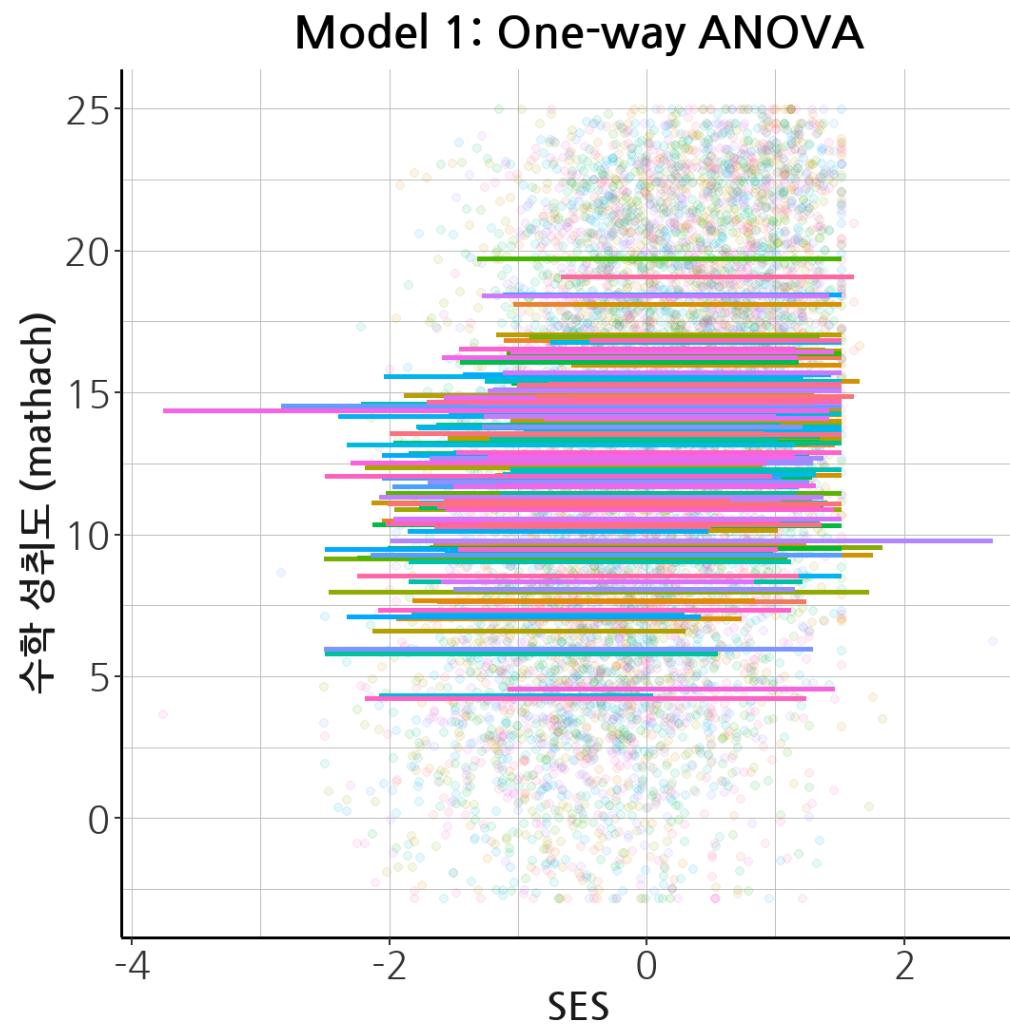
$$ICC(2) = \frac{Var(\text{진 점수})}{Var(Y)} = \frac{\tau_{00}}{\tau_{00} + \sigma^2/n_j}$$

- 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](#))
- $\tau_{00}$ 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.

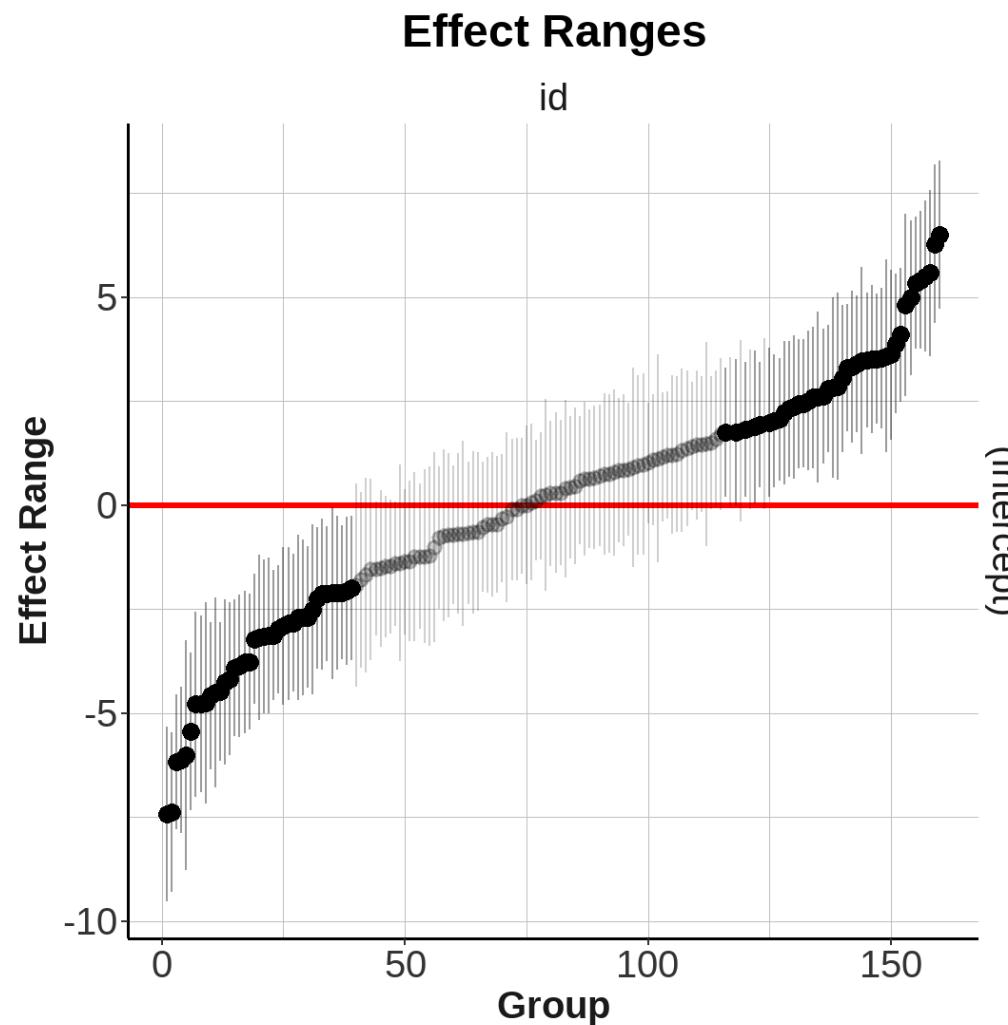
# Model 1. One-way ANOVA: Visualization



# Model 1. One-way ANOVA: Visualization



# Model 1. One-way ANOVA: Visualization



# Model 3. One-Way ANCOVA

- Model 3. One-Way ANCOVA는 Random intercept model로 불리기도 함.
- ANOVA의 경우, average or expected outcomes among groups (level-2 units) 이외에 관심이 없을 때는 유용함. 그러나, Level 1 독립변수들이 종속변수와 Level 2에 상관되어 있는데 Level 1의 독립변수를 모형에 포함하지 않으면 L-2 의 효과를 편파적으로 추정 → ANCOVA 필요
- 예를 들어, 교육연한이나 일 경험은 임금수준에 긍정적 효과를 주는 것으로 알려져 있는데 특정 업무를 하는 사람들 중에서 남자들이 교육연한이나 일경험에 대한 변수 값이 더 높을 때 그 업무에 종사하는 여자와 남자의 평균임금수준에 대한 비교분석 결과는 교육연한이나 일 경험을 통제했는지 여부에 달려있게 될 것
- One-Way ANCOVA는 level 1 독립변수를 투입하고, 기울기를 fixed하게 고려
  - Level 1:  $Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij}$     $\epsilon_{ij} \sim N(0, \sigma^2)$
  - Level 2 Intercept:  $\beta_{0j} = \gamma_{00} + u_{0j}$     $u_{0j} \sim N(0, \tau_{00})$
  - Level 2 Slope:  $\beta_{1j} = \gamma_{10}$
- Fixed Effect:  $\gamma_{00}$  = average outcome for sample of groups,  $\gamma_{10}$  = average individual effect (slope) on outcome
- Random Effect:  $\epsilon_{ij}$  = residuals,  $u_{0j}$  = unique effect of group j on average outcome

# Model 3. One-Way ANCOVA (Cont'd)

- level-1 predictor(s) 도입하려면, 아래와 같은 문제에 대해 결정해야 함.
  1. Whether to introduce random coefficients
  2. Whether to center or transform the level-1 predictor (Module III 참고)
- 일반적으로 ANCOVA에서는 Grand Mean Centering을 적용 Covariate 는 Level-1 종속변수와 영향을 주지만 Level-2 를 구성하는 집단(학교)에 따라 개인들의 Covariate 값이 다를 수 있기 때문에
- Grand-mean Centering 후 모형 추정

```
1 data_merged <- data_merged %>%
2   mutate(ses_grandmc = ses - mean(ses))
3
4 model3 <- lmer(mathach ~ ses_grandmc + (1 | id), data=data_merged)
5 icc(model3)
6
7 # the individual random effects (level 2 residuals of the intercept, i.e. the u0i)
8 ranef(model3)
9 ranova(model3)
10
11 ## Use BruceR package
12 HLM_summary(model3, test.rand = T, digits = 3)
13
14 # lrtest
15 anova(model1, model3)
```

# Model 3. One-Way ANCOVA (Cont'd)

Hierarchical Linear Model (HLM)  
(also known as) Linear Mixed Model (LMM)  
(also known as) Multilevel Linear Model (MLM)

## Model Information:

Formula: mathach ~ ses\_grandmc + (1 | id)

Level-1 Observations:  $N = 7185$

Level-2 Groups/Clusters: id, 160

## Model Fit:

AIC = 46653.169

BIC = 46680.688

$R_m^2 = 0.07665$  (*Marginal R<sup>2</sup>*: fixed effects)

$R_c^2 = 0.18197$  (*Conditional R<sup>2</sup>*: fixed + random effects)

$\Omega^2 = 0.23192$  (= 1 - proportion of unexplained variance)

## ANOVA Table:

	Sum Sq	Mean Sq	NumDF	DenDF	F	p
ses_grandmc	18930.65	18930.65	1.00	6838.08	511.16	<.001 **

## Fixed Effects:

Unstandardized Coefficients (b or  $\gamma$ ):

Outcome Variable: mathach

	b/ $\gamma$	S.E.	t	df	p	[95% CI of b/ $\gamma$ ]
(Intercept)	12.658	(0.188)	67.33	148.3	<.001 **	[12.286, 13.029]
ses_grandmc	2.390	(0.106)	22.61	6838.1	<.001 **	[ 2.183, 2.597]

'df' is estimated by Satterthwaite approximation.

## Standardized Coefficients ( $\beta$ ):

Outcome Variable: mathach

	$\beta$	S.E.	t	df	p	[95% CI of $\beta$ ]
ses_grandmc	0.271	(0.012)	22.61	6838.1	<.001 **	[0.247, 0.294]

## Random Effects:

Cluster	K	Parameter	Variance	ICC
id	160	(Intercept)	4.76817	0.11406
		Residual	37.03440	

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

### Model:

mathach ~ ses\_grandmc + (1 | id)  
npar logLik AIC LRT Df Pr(>Chisq)  
<none> 4 -23323 46653  
(1 | id) 3 -23552 47110 458.92 1 < 2.2e-16 \*\*

—  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

## Model 3. One-Way ANCOVA (Cont'd)

- 두개의 고정효과 값 추정:  $\gamma_{00} = 12.658$ ,  $\gamma_{10} = 2.390$
- SES = 0인 사람들에 대한 수학성취도가  $\gamma_{00}=12.658$ 으로 추정되는데 평균 SES의 성적
- Level 1 독립변수에 의해서 설명되는 Level 1 분산의 비율

$$R^2_{Level-1} = \frac{\sigma_{\epsilon_1}^2 - \sigma_{\epsilon_2}^2}{\sigma_{\epsilon_1}^2} = \frac{39.148 - 37.034}{39.148} = 0.054$$

- Level 1 독립변수에 의해서 설명되는 전체 분산의 비율

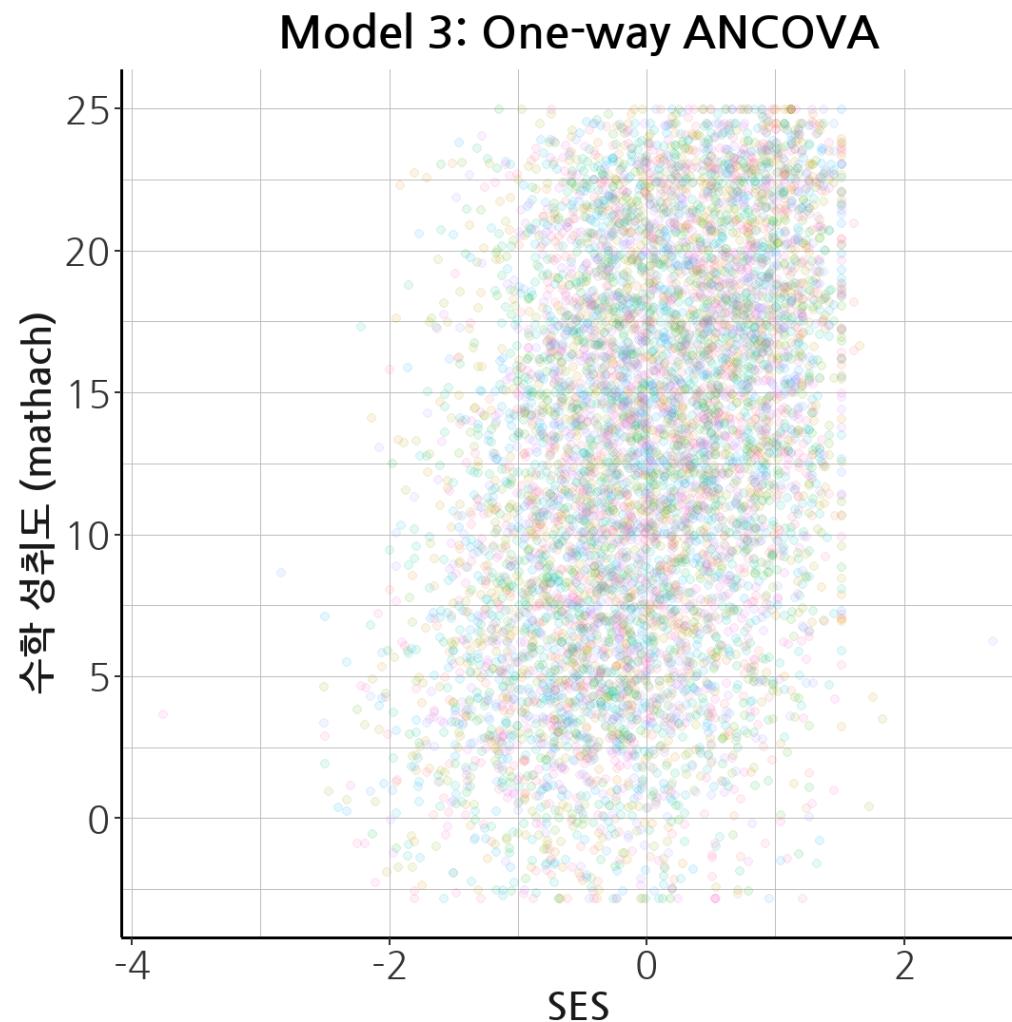
$$R^2_{tot} = \frac{(\sigma_{v_01}^2 + \sigma_{\epsilon_1}^2) - (\sigma_{v_02}^2 + \sigma_{\epsilon_2}^2)}{\sigma_{v_01}^2 + \sigma_{\epsilon_1}^2} = \frac{(8.553 + 39.148) - (4.768 + 37.034)}{(8.553 + 39.148)} = 0.124$$

- Students' SES explained 12% of the total variance in math achievement. (ICC보다 설명된 분산에 해당하는 값을 제시하는 것이 HLM을 정당화하기에 더 적절할 수 있음)
- $\bar{X}_{.j}$ : Group Mean,  $\bar{X}_{..}$ : Grand Mean

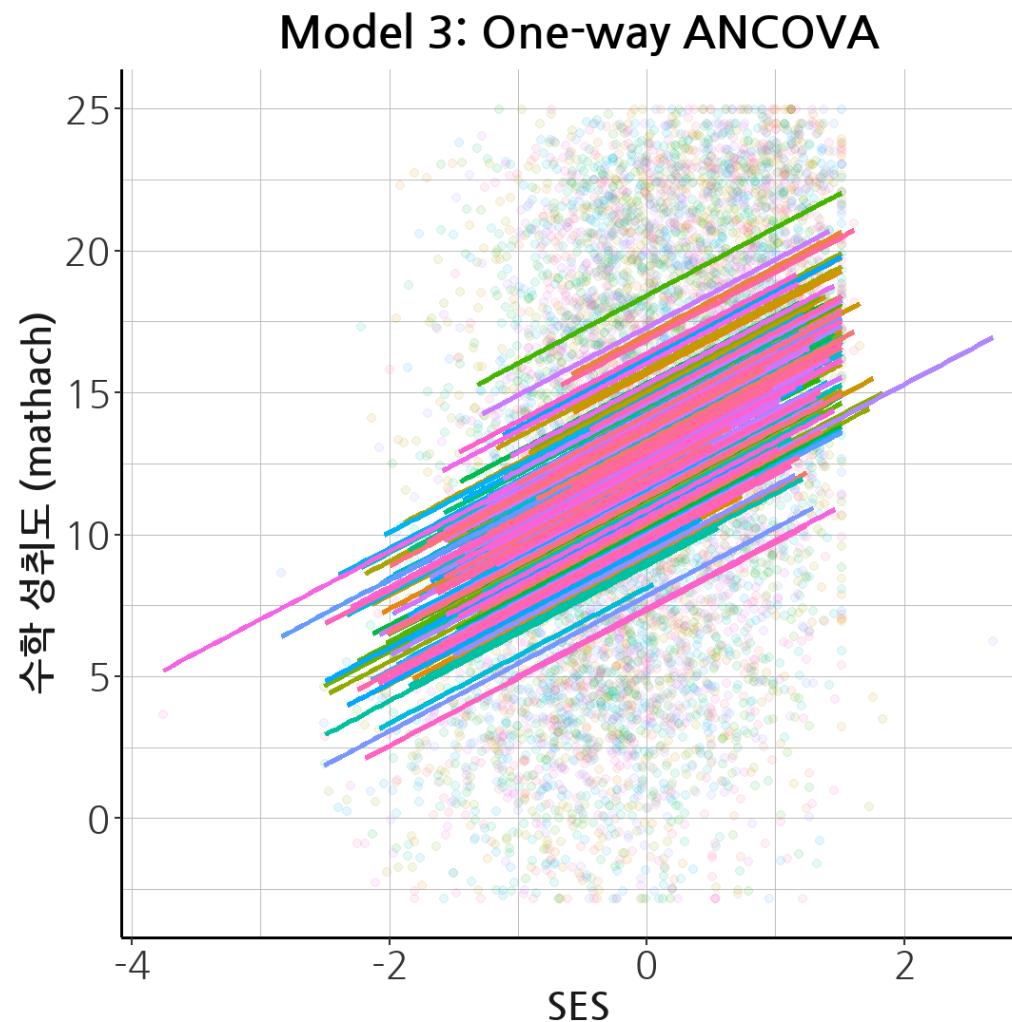
$$\beta_{0j} = u_j - \beta_{1j}(\bar{X}_{.j} - \bar{X}_{..})$$

- Grand-mean centered 모형에서 절편은 각 집단의 평균에서 predictor의 집단평균과 전체평균의 편차를 반영한 adjustment 를 뺀 값

# Model 3. One-Way ANCOVA: Visualization



# Model 3. One-Way ANCOVA: Visualization



# Model 4. Random-Coefficient Model

- Level-1의 절편과 기울기 모두를 random하게 변화하도록 모델링하지만 그러나 절편과 기울기의 random성에서 기인하는 구체적 variation에 대한 예측이나 모델링은 하지 않음 (hence “unconditional” model (NOT fully !))
- 목적: To investigate whether effects of level-1 predictors vary between level-2 units

e.g. SES에 의한 학생들간 수학 성취도 수준 차이가 학교별로 서로 다르게 나타나는가?

- level-1 predictors should be group-mean centered
  - Level 1:  $Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij}$      $\epsilon_{ij} \sim N(0, \sigma^2)$
  - Level 2 intercept:  $\beta_{0i} = \gamma_{00} + u_{0j}$      $u_{0j} \sim N(0, \tau_{00})$
  - Level 2 slope:  $\beta_{1j} = \gamma_{10} + u_{1j}$      $u_{1j} \sim N(0, \tau_{11})$
- $\beta_{0j}$ : 집단 j의 절편,  $\beta_{1j}$ : 집단 j의 기울기
- $\gamma_{00}$ : Level-2 집단들의 평균 절편,  $\gamma_{10}$ : Level-2 집단들의 평균 기울기
- 해석:  $\tau_{00}$  값이 크다면 학교 간 수학성취도 평균에 격차가 큼,  $\tau_{11}$ 의 값이 크다면 학교에 따라 가정배경(SES)으로 학생들의 수학성취도가 차별되는 수준에 큰 차이가 있다는 것,  $\tau_{01} > 0$ 이려면 학교 평균( $\beta_{0j}$ )이 높은 학교에서 SES에 의한 차별효과( $\beta_{1j}$ )가 더 큰 것이고 음수면 그 반대이다.

# Model 4. Random-Coefficient Model (Cont'd)

$$Var\left(\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix}\right) = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix} = T$$

- $Var(u_{0j}) = \tau_{00}$  = unconditional variance in level-1 intercepts
- $Var(u_{1j}) = \tau_{11}$  = unconditional variance in level-1 slopes
- $Cov(u_{0j}, u_{1j}) = \tau_{01} = \rho_{01}$  = unconditional variance between level-1 intercepts and slopes (association between mean school achievement and SES effect on achievement)

```
1 data_merged <- data_merged %>% group_by(id) %>%
2   mutate(meanses = mean(ses),
3         ses_groupmc = ses - meanses)
4 model4 <- lmer(mathach ~ ses_groupmc + (ses_groupmc | id), data=data_merged)
5 model4_alt <- lmer(mathach ~ ses_groupmc + (1 | id), data=data_merged)
6
7 # the individual random effects (level 2 residuals of the intercept, i.e. the u0i)
8 ranef(model4)
9 ranova(model4)
10 bdiag(VarCorr(model4))
11 ## Use BruceR package
12 HLM_summary(model4,test.rand = T, digits = 3)
13 # lrtest for coefficient randomness
14 anova(model4, model4_alt)
```

# Model 4. Random-Coefficient Model (Cont'd)

Model Information:

Formula: mathach ~ ses\_groupmc + (ses\_groupmc | id)

Level-1 Observations: N = 7185

Level-2 Groups/Clusters: id, 160

Model Fit:

AIC = 46726.234

BIC = 46767.513

R<sub>m</sub><sup>2</sup> = 0.04393 (*Marginal R<sup>2</sup>*: fixed effects)

R<sub>c</sub><sup>2</sup> = 0.23194 (*Conditional R<sup>2</sup>*: fixed + random effects)

Omega<sup>2</sup> = 0.24448 (= 1 - proportion of unexplained variance)

ANOVA Table:

	Sum Sq	Mean Sq	NumDF	DenDF	F	p
ses_groupmc	10731.21	10731.21	1.00	155.22	292.40	<.001 ***

Fixed Effects:

Unstandardized Coefficients (b or γ):

Outcome Variable: mathach

	b/γ	S.E.	t	df	p	[95% CI of b/γ]
(Intercept)	12.636	(0.245)	51.68	156.8	<.001 ***	[12.153, 13.119]
ses_groupmc	2.193	(0.128)	17.10	155.2	<.001 ***	[ 1.940, 2.447]

'df' is estimated by Satterthwaite approximation.

Standardized Coefficients (β):

Outcome Variable: mathach

	β	S.E.	t	df	p	[95% CI of β]
ses_groupmc	0.211	(0.012)	17.10	155.2	<.001 ***	[0.186, 0.235]

Random Effects:

Cluster	K	Parameter	Variance	ICC
id	160	(Intercept)	8.68104	0.19129
		ses_groupmc	0.69400	
Residual			36.70019	

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

Model:

mathach ~ ses_groupmc + (ses_groupmc   id)	npar	logLik	AIC	LRT	Df	Pr(>Chisq)
<none>				6	-23357	46726
ses_groupmc in (ses_groupmc   id)	4	-23362	46732	9.7617	2	0.007591 **
---						
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1						

# Model 4. Random-Coefficient Model (Cont'd)

## 1. Fixed effects:

- $\hat{\gamma}_{00} = 12.64$  (the average school mean)
- $\hat{\gamma}_{10} = 2.19$  (the average SES-achievement)

## 2. Random effects

- Level 1:  $(\sigma^2) = 36.70$
- $Var(u_{0j})$ : The estimated variance among the means
- Level 2 intercept:  $(Var(u_{0j}) = \tau_{00}) = 8.681$
- $Var(u_{1j})$ : The estimated variance of slopes
- Level 2 slope:  $(Var(u_{1j}) = \tau_{11}) = 0.69$
- Range of plausible values:

$$Intercept : 12.64 \pm 1.96\sqrt{8.68}, \quad Slope : 2.193 \pm 1.96\sqrt{0.69}$$

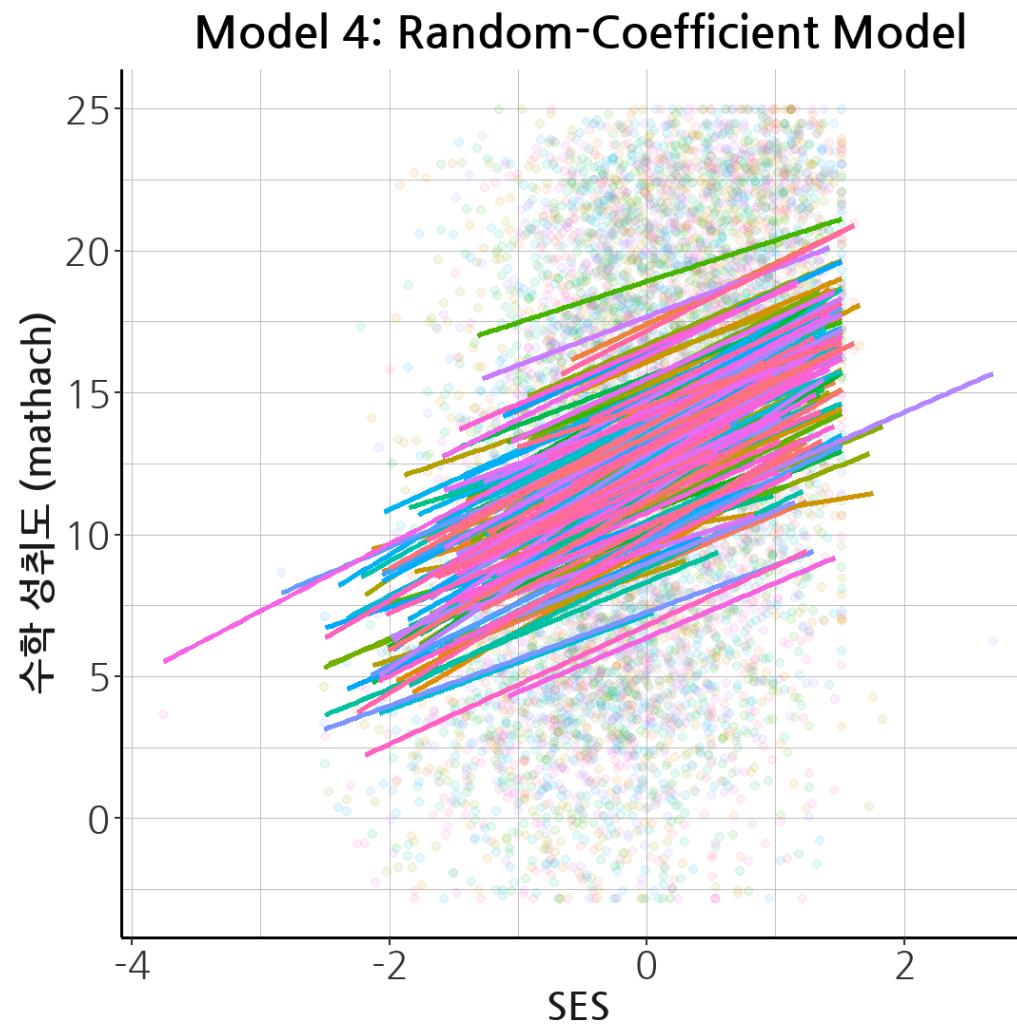
## 3. Variance Explained (at level 1): explains 6.3% of level-1 variance

$$\frac{\hat{\sigma}(Random ANOVA) - \hat{\sigma}(SES)}{\hat{\sigma}(Random ANOVA)} = \frac{39.148 - 36.70019}{39.148} = 0.063$$

# Model 4. Random-Coefficient Model: Visualization



# Model 4. Random-Coefficient Model: Visualization



# Model 5-1. Intercepts-as-Outcomes Model

- Intercepts-as-Outcomes의 경우 Random-effects ANCOVA with Level-2 predictor 혹은 Contextual Effect Model로 볼 수 있음

## 1. Random-effects ANCOVA with Level-2 predictor

- 기본적으로는 Model 3와 동일하며, Level-2 predictor가 추가적으로 투입되어 추정이 이루어짐
- level-1 covariate 는 fixed effects (i.e., assuming that the effects of these covariates are the same for all schools) 일수도 random effects (i.e., assuming that the effects of level 1 variables vary across schools) 일수도 있음
- i.e. **with** a random intercept, and **with or without** a random slope
- Level 1:  $Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij}$     $\epsilon_{ij} \sim N(0, \sigma^2)$
- Level 2:  $\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j}$ ,  $\beta_{1j} = \gamma_{10}$     $u_{0j} \sim N(0, \tau_{00})$
- Fixed Effect:  $\gamma_{00}$  = average outcome for sample of groups,  $\gamma_{01}$  = average group effect (intercept) on outcome,  $\gamma_{10}$  = average individual effect (slope) on outcome
- Random Effect:  $\epsilon_{ij}$  = residuals,  $u_{0j}$  = unique effect of group j on average outcome

# Model 5-1. Intercepts-as-Outcomes Model (Cont'd)

## 1. Random coefficient not included

- Level 1:  $Y_{ij} = \beta_{0j} + \beta_{1j} SES_{ij} + \epsilon_{ij}$     $\epsilon_{ij} \sim N(0, \sigma^2)$
- Level 2 intercept:  $\beta_{0i} = \gamma_{00} + \gamma_{01} SECTOR_j + u_{0j}$     $u_{0j} \sim N(0, \tau_{00})$
- Level 2 slope:  $\beta_{1j} = \gamma_{10}$

## 2. Random coefficient included

- Level 1:  $Y_{ij} = \beta_{0j} + \beta_{1j} SES_{ij} + \epsilon_{ij}$     $\epsilon_{ij} \sim N(0, \sigma^2)$
- Level 2 intercept:  $\beta_{0i} = \gamma_{00} + \gamma_{01} SECTOR_j + u_{0j}$     $u_{0j} \sim N(0, \tau_{00})$
- Level 2 slope:  $\beta_{1j} = \gamma_{10} + u_{1j}$     $u_{1j} \sim N(0, \tau_{11})$

# Model 5-1. Intercepts-as-Outcomes Model (Cont'd)

- Random coefficient not included

```
1 ## Random-effects ANCOVA with Level-2 predictor
2
3 # without random coefficient
4 model5 <- lmer(mathach ~ ses + sector + (1 | id), data=data_merged)
5 icc(model5)
6
7 # the individual random effects (level 2 residuals of the intercept, i.e. the u0i)
8 ranef(model5)
9 ranova(model5)
10
11 ## Use BruceR package
12 HLM_summary(model5, test.rand = T, digits = 3)
```

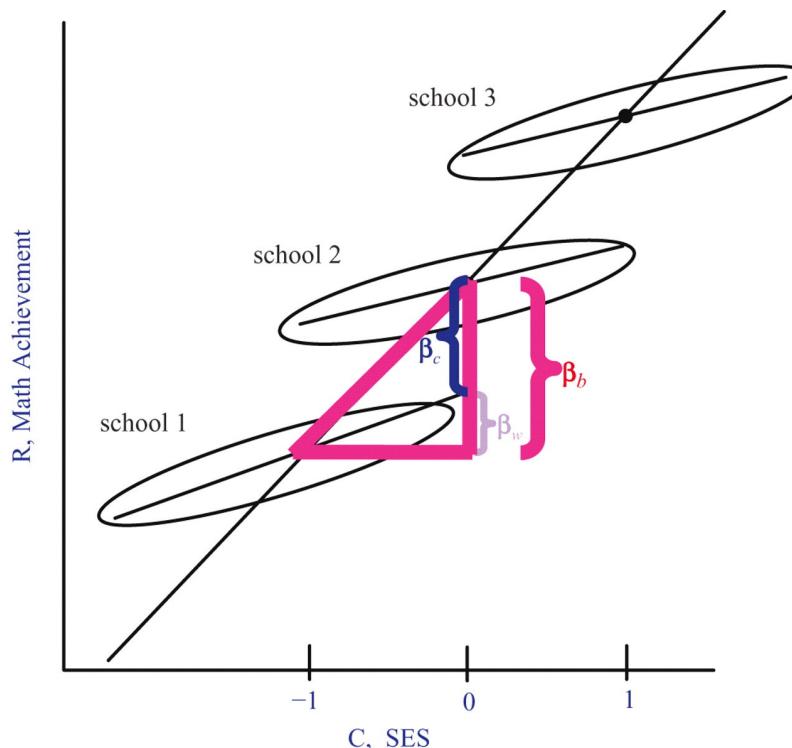
- Random coefficient included

```
1 # with random coefficient
2
3 model6 <- lmer(mathach ~ ses + sector + (ses | id), data=data_merged)
4 icc(model6)
5
6 # the individual random effects (level 2 residuals of the intercept, i.e. the u0i)
7 ranef(model6)
8 ranova(model6)
9
10 ## Use BruceR package
11 HLM_summary(model6, test.rand = T, digits = 3)
```

# Model 5-1. Context effect model

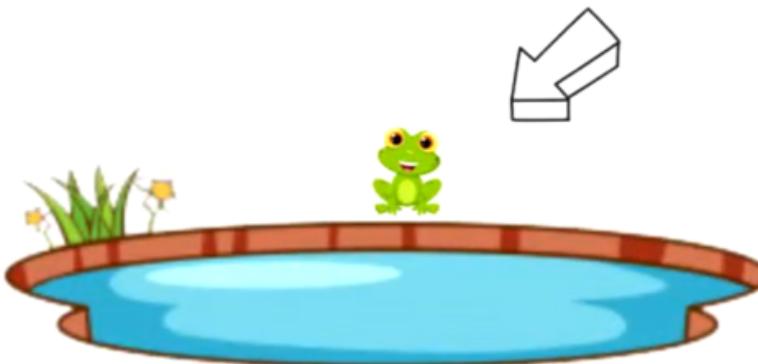
## 2. Context effect model

- Context effect: a group-level effect above and beyond the individual level effect
  - $\beta_w$ : Difference in Y between two student *within* same school
  - $\beta_b$ : Difference in mean of Y ( $\bar{Y}$ ) *between* two schools
  - $\beta_c = \beta_b - \beta_w$ : Difference in between two students who have the same individual SES, but who attend schools that differ by one unit of mean SES



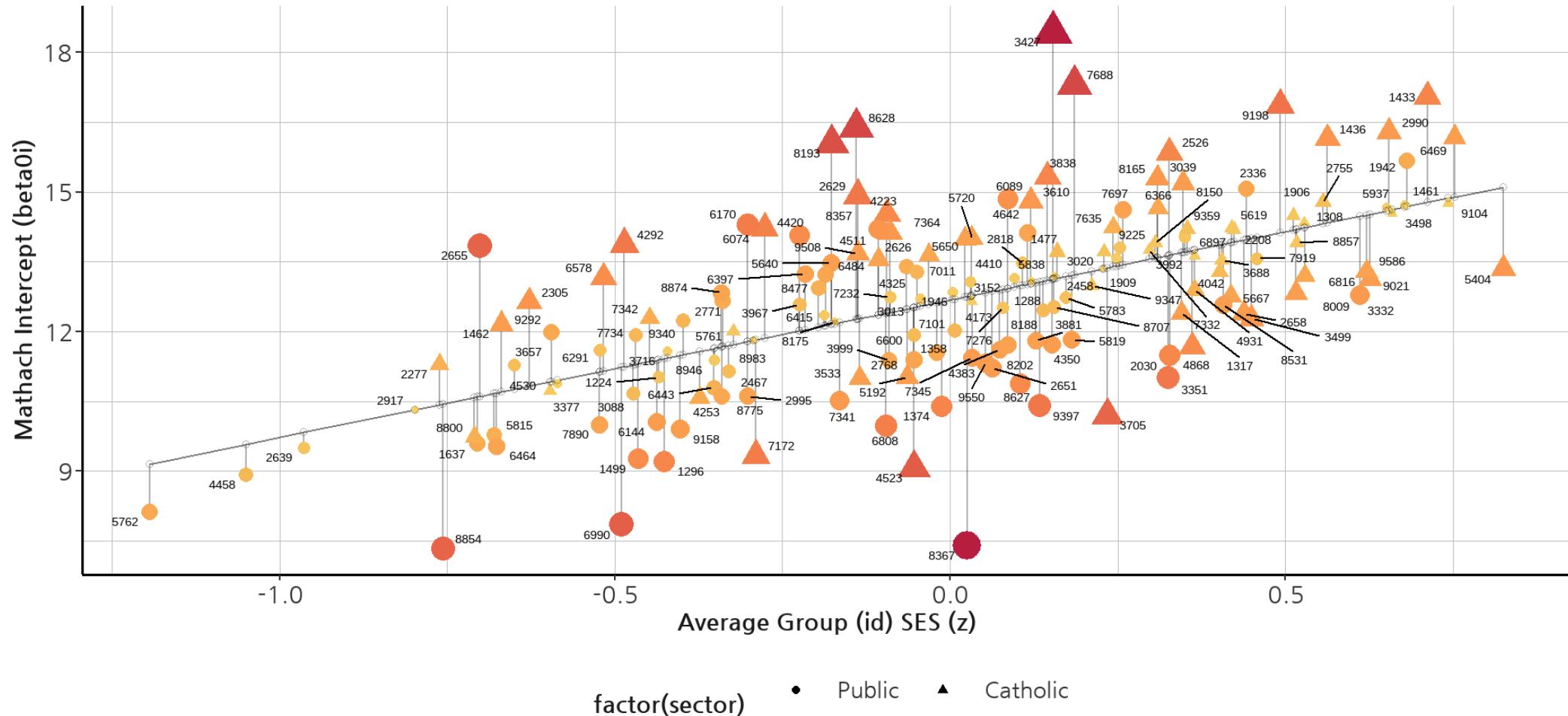
# Model 5-1. Context effect model (Cont'd)

- 개구리-연못 이론 (Davis, 1966): Social Comparison theory에 기반
  - 조직연구에서 개인수준의 변수를 조직 또는 집단의 수준에서 aggregate 했을 때 가중되는 효과
  - 지금까지는 학생들의 SES의 수학성취도에 대한 영향을 Level-1 equation에서 학교의 영향과 상관없이 고려하였으나, 실제는 어떤 학교에서 학습은 영향 (e.g. Peer Effects, Proxy for other variable not in model)



# Model 5-1. Context effect model (Cont'd)

학교별 평균 SES 대비 평균 수학성취도: Level-2 Residual in Intercept



# Model 5-1. Context effect model (Cont'd)

- 일반적으로 Group-mean Centering을 적용함
  - Level 1:  $Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_j) + \epsilon_{ij}$   $\epsilon_{ij} \sim N(0, \sigma^2)$
  - Level 2 intercept:  $\beta_{0i} = \gamma_{00} + \gamma_{01}\bar{X}_j + u_{0j}$   $u_{0j} \sim N(0, \tau_{00})$
  - Level 2 slope:  $\beta_{1j} = \gamma_{10} + u_{1j}$   $u_{1j} \sim N(0, \tau_{11})$
- Group-mean Centering을 선택하면  $X_{ij}$ 와  $Y_{ij}$ 의 관계는 집단 내와 집단 간 영향으로 구분되어지는데, 이를 분리하기 위한 contextual effect를 추가하여 완전한 집단 내 영향으로 추정

```
1 ## Contextual Effect
2 model17 <- lmer(mathach ~ ses_groupmc + meanses + (1 | id), data = data_merged, REML = TRUE)
3 icc(model17)
4 summary(model17)
5
6 # the individual random effects (level 2 residuals of the intercept, i.e. the u0i)
7 ranef(model17)
8 ranova(model17)
9
10 ## Use BruceR package
11 HLM_summary(model17, test.rand = T, digits = 3)
12
13 # LRtest
14 anova(model13, model17)
```

# Model 5-1. Context effect model (Cont'd)

Model Information:

Formula: mathach ~ ses\_groupmc + ses\_grouped + (1 | id)

Level-1 Observations: N = 7185

Level-2 Groups/Clusters: id, 160

Model Fit:

AIC = 46578.584

BIC = 46612.983

R<sub>m</sub><sup>2</sup> = 0.16733 (*Marginal R<sup>2</sup>*: fixed effects)

R<sub>c</sub><sup>2</sup> = 0.22379 (*Conditional R<sup>2</sup>*: fixed + random effects)

Omega<sup>2</sup> = 0.23072 (= 1 - proportion of unexplained variance)

ANOVA Table:

	Sum Sq	Mean Sq	NumDF	DenDF	F	p
ses_groupmc	15051.53	15051.53	1.00	7021.51	406.59	<.001 ***
ses_grouped	9737.33	9737.33	1.00	153.37	263.04	<.001 ***

Fixed Effects:

Unstandardized Coefficients (b or γ):

Outcome Variable: mathach

	b/γ	S.E.	t	df	p	[95% CI of b/γ]
(Intercept)	12.683 (0.149)	84.91	153.7	<.001 ***	[12.388, 12.978]	
ses_groupmc	2.191 (0.109)	20.16	7021.5	<.001 ***	[ 1.978, 2.404]	
ses_grouped	5.866 (0.362)	16.22	153.4	<.001 ***	[ 5.152, 6.581]	

'df' is estimated by Satterthwaite approximation.

Standardized Coefficients (β):

Outcome Variable: mathach

	β	S.E.	t	df	p	[95% CI of β]
ses_groupmc	0.210 (0.010)	20.16	7021.5	<.001 ***	[0.190, 0.231]	
ses_grouped	0.353 (0.022)	16.22	153.4	<.001 ***	[0.310, 0.396]	

Random Effects:

Cluster	K	Parameter	Variance	ICC
id	160	(Intercept)	2.69253	0.06780
Residual			37.01906	

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

Model:

mathach ~ ses\_groupmc + ses\_grouped + (1 | id)

npar logLik AIC LRT Df Pr(>Chisq)

<none> 5 -23284 46579

(1 | id) 4 -23417 46842 265.12 1 < 2.2e-16 \*\*\*

---

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

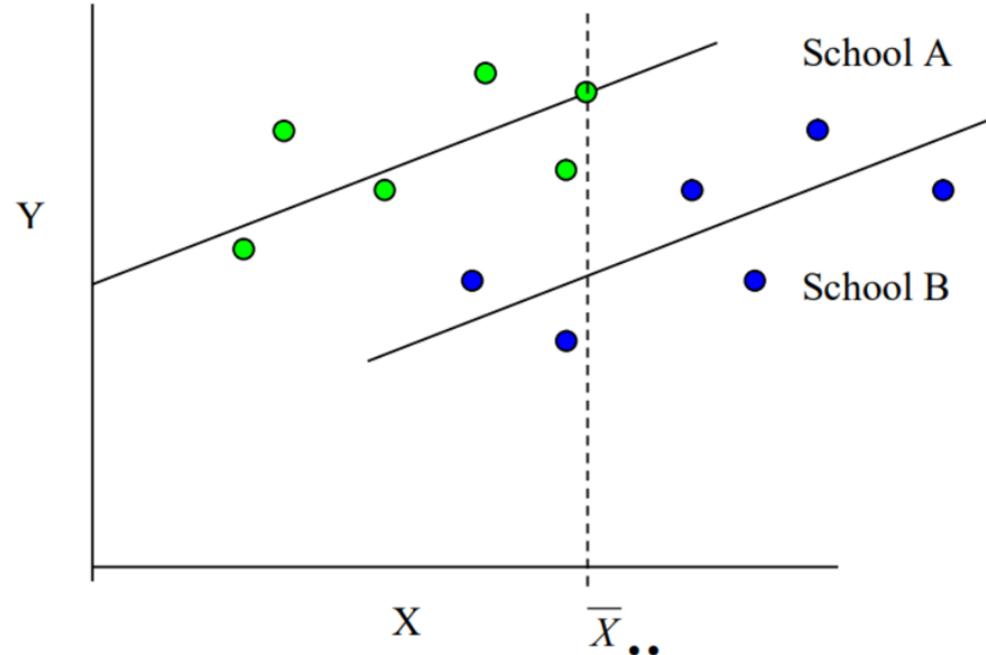
# Module IV: 중심화와 수준 간 상호작용

# Centering

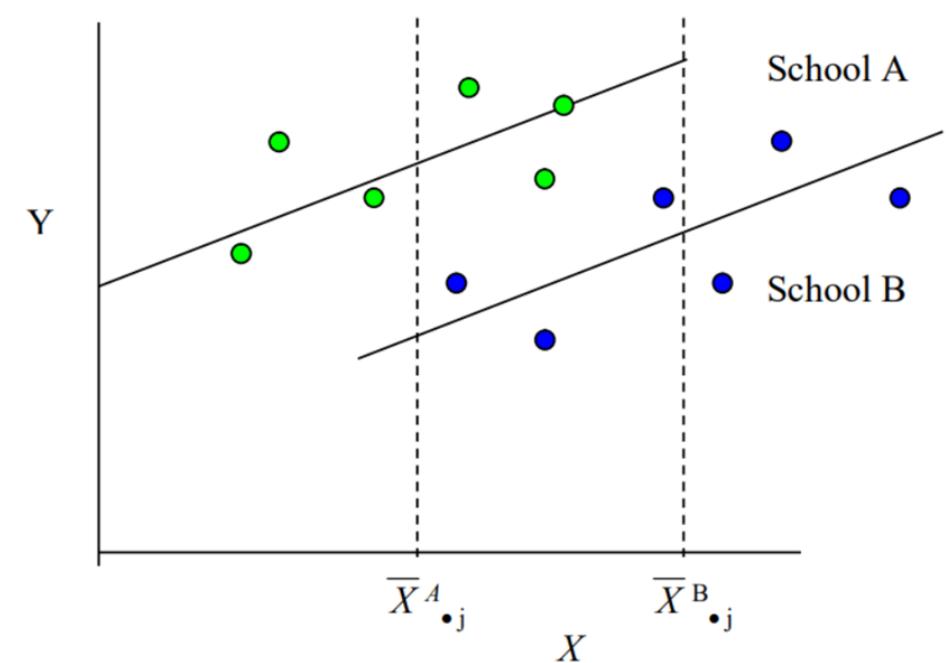
- 일반적인 회귀분석에서 변수를 중심화하는 경우는 대부분 상호작용 효과를 분석할 때이다. 그리고 이 때 변수를 중심화하는 것은 절편(intercept)을 의미있는 값으로 해석하기 위해서, 그리고 다중공선성(multicollinearity)의 문제를 해결하기 위해서 사용됨
- Raudenbush and Bryk(2002): 변수 중심화는 1) 절편의 의미, 2) 독립변수 추정치의 의미, 3) 추정치의 수치적 안정성(numerical stability)에 영향
- Centering을 하지 않은 모형:  $Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}$
- e.g.  $X_{ij}$  = previous math achievement, 절편은 0일때의 기대값, 현실적으로 해당 변수가 0이 될 수 있는가?
- 변수 중심화의 두가지 유형
  - Grand mean centering (전체 평균 중심화):  $Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_{..}) + \epsilon_{ij}$ 
    - when expected value of  $Y_{ij}$  when  $X_{ij} = \bar{X}_{..}$
    - $\beta_{ij}$  (절편) = The mathach of the “average student” in school j
  - Group mean centering (집단 평균 중심화):  $Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_{.j}) + \epsilon_{ij}$ 
    - when expected value of  $Y_{ij}$  when  $X_{ij} = \bar{X}_{.j}$
    - $\beta_{ij}$  (절편) = The mathach of the “average student” in the all sample of schools
    - 어느 학교에서도  $\bar{X}_{..}$ 인 값들은 서로 같음

# Grand-mean Centering & Group-mean Centering

Grand-Mean Centering



Group-Mean Centering



# Grand-mean Centering & Group-mean Centering (Cont'd)

- 그래서 도대체 언제 Centering을 해야 하는가? [출처](#)

## 1. Level-1 변수들 간의 관계에 관심이 있을 때

- 개인-집단의 2수준 자료에서  $X_{ij}$ 에는 집단 내 변량(within-group variation)과 집단 간 변량(between-group variation)이 혼재됨 (Contextual Effect Model 참고)
- 중심화를 하지 않거나, 전체평균 중심화를 하면  $X_{1j}$ 의 계수인  $\gamma_{10}$ 은 집단 내 효과와 집단 간 효과의 가중평균이 되므로, 집단 내 효과도 아니고 집단 간 효과도 아니고 총 효과도 아닌 값으로 추정이 이루어짐
- (만약 집단 평균 중심화를 하지 않는다면  $\beta_w = \beta_b$ 를 가정하므로 이게 현실적인 가정인지 고민이 필요함)
- 따라서, 집단평균 중심화를 통해 집단 간(between) 변량을 제거하여 순수한  $\beta_{Within}$  값을 추정해야 함
- 2 수준에 집단평균 변수  $\bar{X}_{.j}$ 를 투입하여 Contextual Effect를 함께 고려하는 것이 바람직하지만, 그렇지 않은 논문들도 多

## 2. Level-2 변수들과의 관계에 관심이 있을 때

- Level-1 변수를 집단평균 중심화하면 Level-1 절편은 조정되지 않은 집단의 평균값이기 때문에 집단 간 변량이 제거되어 Level-2에서 영향을 미치지 않음
- Level-1 변수를 전체평균 중심화하면 Level-1 절편은 조정된 평균값이기 때문에 Level-2 변수의 영향을 추정할 때 Level-1 변수의 집단 간 평균값 차이에 의해 조정
- 따라서, Level-1 변수를 중심화하지 않거나, 전체평균 중심화를 해야함

# Grand-mean Centering & Group-mean Centering (Cont'd)

- 그래서 도대체 언제 Centering을 해야 하는가? [출처](#)
  - 3. Level-1 & Level-2 수준 간 상호작용에 관심이 있을 때
    - Level-1 변수( $X_{ij}$ )의 영향이 Level-2 변수( $Z_j$ )에 따라 어떻게 달라지는지가 연구문제일 때
    - Level-1 변수가 집단 내 효과( $\beta_w$ 를 의미하도록 집단평균 중심화를 하는 것이 적절 (Hofmann and Gavin, 1998)
    - 전체평균 중심화 하면 집단 간 상호작용과 수준 간 상호작용의 혼재가 발생

# Grand-mean Centering & Group-mean Centering (Cont'd)

```
1 # id는 각 학교들의 식별자이기에, 식별자를 기준으로 집단화
2
3 data_merged <- data_merged %>%
4   # 전체평균 중심화 (ses - grandmean)
5   mutate(ses_grandmc = ses - mean(ses)) %>%
6   # 맥락효과(meanses = groupmean)와 집단평균 중심화 (ses_groupmc = ses - groupmean)
7   # id 기준으로 observations를 group화
8   group_by(id) %>%
9   mutate(meanses = mean(ses),
10         ses_groupmc = ses - meanses) %>%
11   # group화 해제 (반드시 해야함)
12   ungroup()
13
14 # 각 학교별 인원, 평균 ses 확인하는 법
15 data_merged %>%
16   group_by(id) %>%
17   # n = 인원수, meanses = 평균 ses 점수
18   summarise(n = n(),
19             meanses = mean(ses))
```

# Model 5-2. Slopes-as-Outcomes Model

- intercepts and slopes are conditioned by level-2 predictors (따라서, 집단평균 중심화)
  - Estimation techniques similar to previous models except estimates are conditioned
- 목적: What level-2 factors predict differences in slopes between level-2 groups?
  - RQ: (e.g. whether variation in SES effects across schools can be attributed to the type of school - public vs Catholic)
- 모형
  - Level 1 Model:  $Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \epsilon$     $\epsilon \sim N(0, \sigma^2)$
  - Level 2 Model (intercept):  $\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j}$     $u_{0j} \sim N(0, \tau_{00}^2)$
  - Level 2 Model (slope):  $\beta_{1j} = \gamma_{10} + \gamma_{11}Z_j + u_{1j}$     $u_{1j} \sim N(0, \tau_{11}^2)$
- Overall Model Example: (intercept + fixed + random)
  - intercept:  $\beta_{0j} + \gamma_{00}$
  - fixed effect:  $\gamma_{01}(SECTOR_j) + \gamma_{10}(SES_{ij}) + \gamma_{11}(SECTOR_j)(SES_{ij})$
  - random effect:  $u_{1j}(SES_{ij}) + u_{0j} + \epsilon_{ij}$
- 장점: allows us to explain the variation in both intercepts and slopes
- HLM의 최종 목적인 모형이며, 가장 어려운 모형이기도 하다.

# Model 5-2. Slopes-as-Outcomes Model (Cont'd)

- Random coefficient included

```
1 ## with random coefficient
2 model8 <- lmer(mathach ~ ses_groupmc * sector + (ses_groupmc | id), data = data_merged, REML = FALSE)
3
4 summary(model8)
5 icc(model8)
6
7 # the individual random effects (level 2 residuals of the intercept, i.e. the u0i)
8 ranef(model8)
9 ranova(model8)
10
11 ## Use BruceR package
12 HLM_summary(model8, test.rand = T, digits = 3)
```

- Random coefficient not included

```
1 ## without random coefficient
2 model9 <- lmer(mathach ~ ses_groupmc * sector + (1 | id), data = data_merged, REML = FALSE)
3
4 summary(model9)
5 icc(model9)
6
7 # the individual random effects (level 2 residuals of the intercept, i.e. the u0i)
8 ranef(model9)
9 ranova(model9)
10
11 ## Use BruceR package
12 HLM_summary(model9, test.rand = T, digits = 3)
13
14 # LRtest
15 anova(model8, model9) # with randomness is better
```

# Model 5-2. Slopes-as-Outcomes Model (Cont'd)

Model Information:

Formula: mathach ~ ses\_groupmc \* sector + (ses\_groupmc | id)

Level-1 Observations: N = 7185

Level-2 Groups/Clusters: id, 160

Model Fit:

AIC = 46649.881

BIC = 46704.919

$R_m^2 = 0.09015$  (*Marginal R<sup>2</sup>*: fixed effects)

$R_c^2 = 0.23139$  (*Conditional R<sup>2</sup>*: fixed + random effects)

$\Omega^2 = 0.24000$  (= 1 - proportion of unexplained variance)

ANOVA Table:

	Sum Sq	Mean Sq	NumDF	DenDF	F	p
ses_groupmc	12378.08	12378.08	1.00	153.45	337.23	<.001 ***
sector	1519.54	1519.54	1.00	155.59	41.40	<.001 ***
ses_groupmc:sector	1224.67	1224.67	1.00	153.45	33.36	<.001 ***

Fixed Effects:

Unstandardized Coefficients (b or γ):

Outcome Variable: mathach

	b/y	S.E.	t	df	p	[95% CI of b/y]
(Intercept)	11.394 (0.291)	39.17	160.4	<.001 ***	[10.819, 11.968]	
ses_groupmc	2.803 (0.154)	18.21	143.0	<.001 ***	[ 2.499, 3.107]	
sectorCatholic	2.807 (0.436)	6.43	155.6	<.001 ***	[ 1.946, 3.669]	
ses_groupmc:sectorCatholic	-1.341 (0.232)	-5.78	153.4	<.001 ***	[-1.800, -0.883]	

'df' is estimated by Satterthwaite approximation.

Standardized Coefficients (β):

Outcome Variable: mathach

	β	S.E.	t	df	p	[95% CI of β]
ses_groupmc	0.269 (0.015)	18.21	143.0	<.001 ***	[ 0.240, 0.298]	
sectorCatholic	0.204 (0.032)	6.43	155.6	<.001 ***	[ 0.141, 0.267]	
ses_groupmc:sectorCatholic	-0.085 (0.015)	-5.78	153.4	<.001 ***	[-0.114, -0.056]	

Random Effects:

Cluster	K	Parameter	Variance	ICC
id	160	(Intercept)	6.64044	0.15320
		ses_groupmc	0.23991	
		Residual	36.70554	

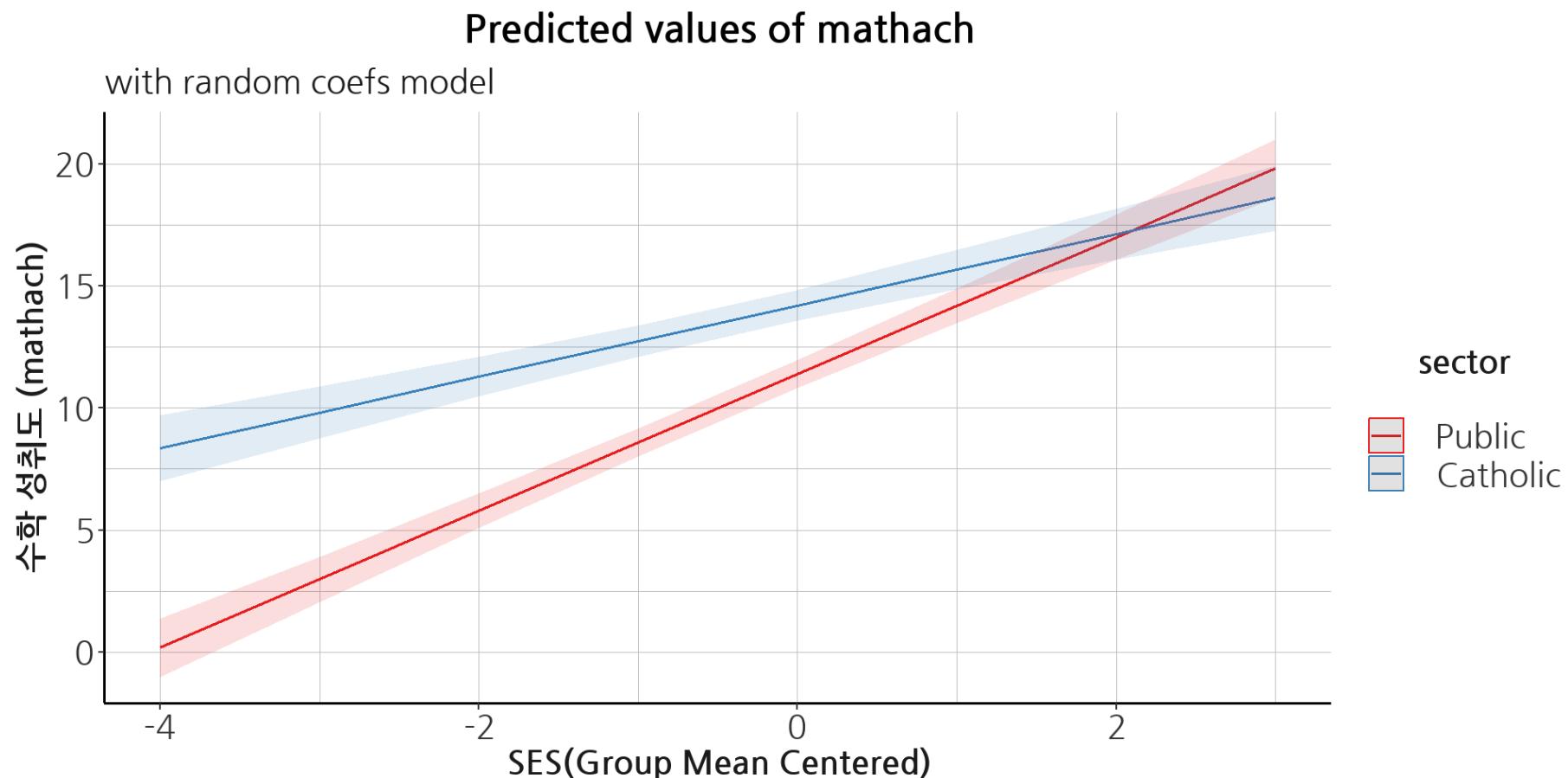
ANOVA-like table for random-effects: Single term deletions

Model:

```
mathach ~ ses_groupmc + sector + (ses_groupmc | id) + ses_groupmc:sector
      npar logLik AIC LRT Df Pr(>Chisq)
<none>          8 -23317 46650
ses_groupmc in (ses_groupmc | id)   6 -23323 46658 11.83 2  0.002699 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Model 5-2. Slopes-as-Outcomes: Marginsplot

```
1 # Marginsplot
2 # install.packages("sjPlot")
3 # install.packages("effects")
4 library(sjPlot)
5 plot_model(model8, type = 'pred', terms = c('ses_groupmc','sector'), ci.lvl = 0.95)
```



# Model Specification Summary

Model comparison over Mathach ~ SES

	Group mean		Individual mathach								
	OLS	OLS	linear			mixed-effects					
	Between OLS Pooled OLS One-way-ANOVA Random Intercept Random Coefficient Intercept-as-Outcome Contextual Effects Slope-as-Outcome Slope-as-Outcome with Random Coeff		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
SES		2.95*** (0.10)						2.39*** (0.12)			
Mean SES		5.39*** (0.38)						5.87*** (0.36)		5.24*** (0.36)	
Group Mean Centered SES						2.19*** (0.13)		2.19*** (0.11)	2.78*** (0.14)	2.79*** (0.15)	
Grand Mean Centered SES				2.39*** (0.11)							
Sector (1 = Catholic)	1.22*** (0.32)	1.94*** (0.15)					2.54*** (0.34)		2.81*** (0.44)	1.25*** (0.30)	
Group Mean Centered SES*Sector								-1.35*** (0.22)	-1.35*** (0.23)		
Constant	12.12*** (0.20)	11.79*** (0.11)	12.64*** (0.24)	12.66*** (0.19)	12.64*** (0.24)	11.47*** (0.23)	12.68*** (0.15)	11.39*** (0.29)	12.12*** (0.20)		
Observations	160	7185	7185	7185	7185	7185	7185	7185	7185	7185	
R-squared	0.65	0.15									
Adj. R-squared	0.64	0.15									
Log Likelihood		-23558.40	-23322.58	-23357.12	-23300.93	-23281.90	-23322.86	-23254.04			
AIC		47122.79	46653.17	46726.23	46615.85	46573.81	46657.71	46526.08			
BIC		47143.43	46680.69	46767.51	46664.01	46608.21	46698.99	46587.99			

p < .05; p < .01; p < .001

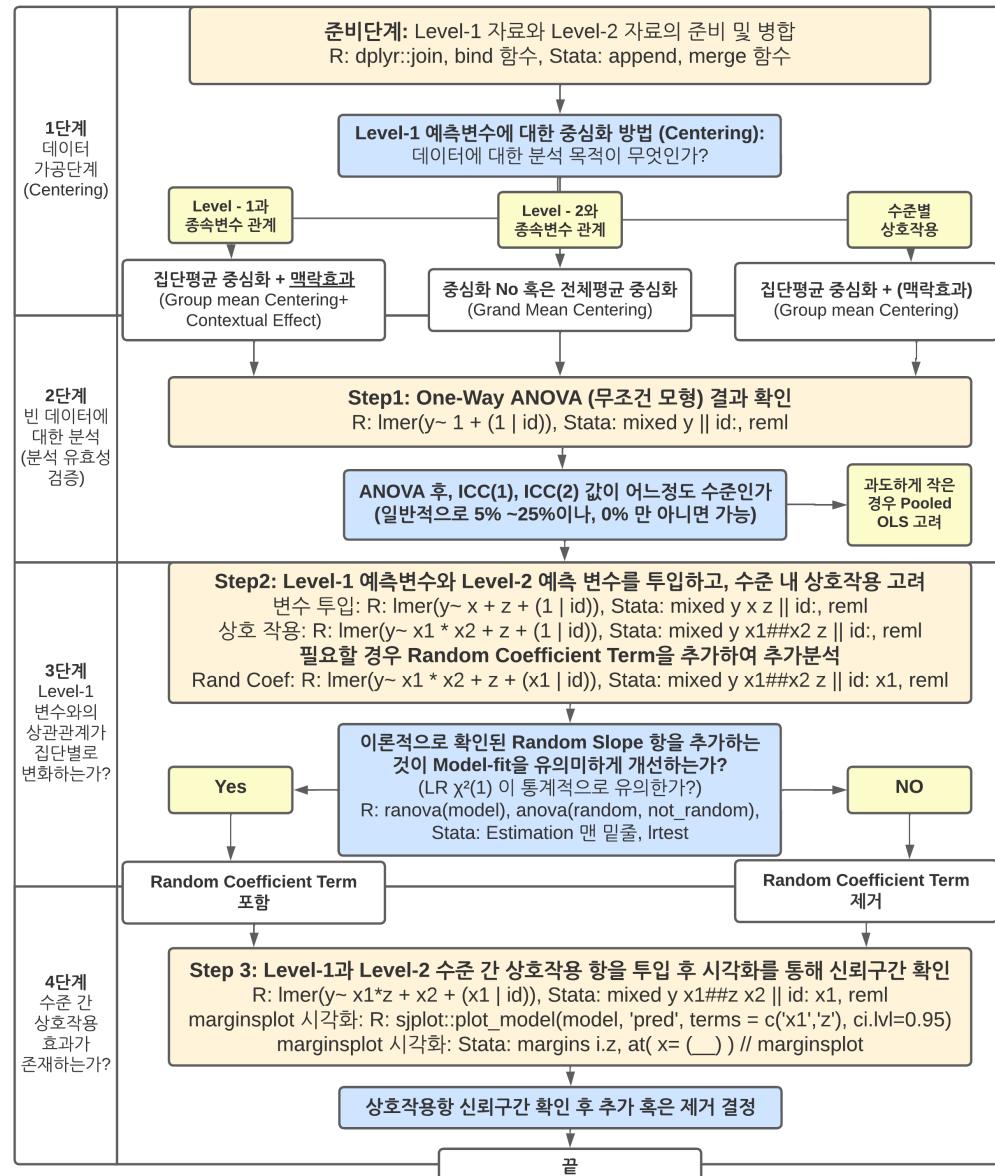
# Model Specification Summary (Cont'd)

Model comparison over Mathach ~ SES

Predictors	OLS (dv = mathch)			HLM (dv = mathch)		
	Estimates	std. Error	CI	Estimates	std. Error	CI
ses	2.95 ***	0.10	2.76 - 3.14			
ses groupmc				2.79 ***	0.15	2.48 - 3.09
meanses				5.24 ***	0.36	4.53 - 5.96
sector [Catholic]	1.94 ***	0.15	1.64 - 2.23	1.25 ***	0.30	0.66 - 1.85
ses groupmc × sector [Catholic]				-1.35 ***	0.23	-1.81 - -0.89
(Intercept)	11.79 ***	0.11	11.59 - 12.00	12.12 ***	0.20	11.73 - 12.50
<b>Random Effects</b>						
$\sigma^2$			36.71			
$\tau_{00}$			2.32 id			
$\tau_{11}$			0.25 id.ses_groupmc			
$\rho_{01}$			0.26 id			
ICC			0.06			
N			160 id			
Observations	7185		7185			
R <sup>2</sup> / R <sup>2</sup> adjusted	0.149 / 0.149		0.176 / 0.227			
AIC	46946.637		46532.575			
log-Likelihood	-23469.318		-23254.038			

\* p<0.05 \*\* p<0.01 \*\*\* p<0.001

# Module III: Sum-up



# E.O.D.

# 참고: R vs. STATA

```
1 cd "E:/OneDrive - SNU/(B) 대학원/세미나/HLM/3주차 HLM"
2 use HSB1.dta, clear
3 merge m:1 id using ".\HSB2.dta"
4 bysort id :egen meanses = mean(ses)
5 gen ses_groupmc= ses - meanses
6 eststo final_rand: mixed mathach c.ses_groupmc##ib0.sector meanses|| id: ses_groupmc, mle var cov(un)
```

```
Result                      Number of obs
-----
Not matched                  0
Matched                     7,185  (_merge==3)
-----
```

Performing EM optimization ...

Performing gradient-based optimization:

```
Iteration 0:  log likelihood = -23254.609
Iteration 1:  log likelihood = -23254.04
Iteration 2:  log likelihood = -23254.038
```

Computing standard errors ...

```
Mixed-effects ML regression
Group variable: id
Number of obs      =      7,185
Number of groups  =        160
Obs per group:
               min =          14
               avg =        44.9
               max =          67
Wald chi2(4)      =     707.99
Prob > chi2       =     0.0000
```

Log likelihood = -23254.038

---

mathach	Coefficient	Std. err.	z	P> z	[95% conf. interval]
---------	-------------	-----------	---	------	----------------------

---

# 참고: R vs. STATA (Summary & Fit Statistics)

Hierarchical Linear Model (HLM)  
(also known as) Linear Mixed Model (LMM)  
(also known as) Multilevel Linear Model (MLM)

Model Information:

Formula: mathach ~ ses\_groupmc \* sector + meanses + (ses\_groupmc | id)

Level-1 Observations:  $N = 7185$

Level-2 Groups/Clusters: id, 160

Model Fit:

AIC = 46526.077

BIC = 46587.994

$R_{(m)}^2 = 0.17615$  (*Marginal R<sup>2</sup>: fixed effects*)

$R_{(c)}^2 = 0.22721$  (*Conditional R<sup>2</sup>: fixed + random effects*)

$\Omega^2 = 0.23844$  (= 1 - proportion of unexplained variance)

```
. mixed mathach c.ses_groupmc##ib0.sector meanses|| id: ses_groupmc, mle var cov(un)
```

Performing EM optimization ...

Performing gradient-based optimization:

Iteration 0: log likelihood = -23254.609

Iteration 1: log likelihood = -23254.04

Iteration 2: log likelihood = -23254.038

Computing standard errors ...

Mixed-effects ML regression

Group variable: id

Number of obs = 7,185

Number of groups = 160

Obs per group:

min = 14

avg = 44.9

max = 67

Wald chi2(4) = 707.99

Prob > chi2 = 0.0000

Log likelihood = -23254.038

. estat ic

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	7,185	.	-23254.04	9	46526.08	46587.99

Note: BIC uses N = number of observations. See [R] BIC note.

# 참고: R vs. STATA (Fixed Effects)

Fixed Effects: Unstandardized Coefficients (b or y): Outcome Variable: mathach						
	b/y	S.E.	t	df	p	[95% CI of b/y]
(Intercept)	12.116	(0.197)	61.38	163.1	<.001	** [11.726, 12.506]
ses_groupmc	2.788	(0.155)	18.01	145.6	<.001	** [ 2.482, 3.094]
sectorCatholic	1.253	(0.303)	4.13	152.5	<.001	** [ 0.654, 1.852]
meanses	5.244	(0.365)	14.38	154.2	<.001	** [ 4.524, 5.964]
ses_groupmc:sectorCatholic	-1.348	(0.233)	-5.78	154.9	<.001	** [-1.809, -0.887]

'df' is estimated by Satterthwaite approximation.

Standardized Coefficients ( $\beta$ ): Outcome Variable: mathach						
	$\beta$	S.E.	t	df	p	[95% CI of $\beta$ ]
ses_groupmc	0.268	(0.015)	18.01	145.6	<.001	** [ 0.238, 0.297]
sectorCatholic	0.091	(0.022)	4.13	152.5	<.001	** [ 0.048, 0.135]
meanses	0.315	(0.022)	14.38	154.2	<.001	** [ 0.272, 0.359]
ses_groupmc:sectorCatholic	-0.086	(0.015)	-5.78	154.9	<.001	** [-0.115, -0.056]

mathach	Coefficient	Std. err.	z	P> z	[95% conf. interval]
ses_groupmc	2.787651	.1548387	18.00	0.000	2.484173 3.091129
1.sector	1.252655	.3032032	4.13	0.000	.658388 1.846923
sector#c.ses_groupmc	-1.3481	.2333073	-5.78	0.000	-1.805374 -.8908265
1	5.244067	.3646407	14.38	0.000	4.529384 5.958749
_cons	12.11605	.1973894	61.38	0.000	11.72917 12.50292

# 참고: R vs. STATA (Random Effects)

ICC

$\tau_{00}$

$\tau_{11}$

$\tau_{01}(\rho_{01})$

$\sigma^2$

Random Effects:				
Cluster	K	Parameter	Variance	ICC
id	160	(Intercept)	2.31794	0.05940
		ses_groupmc	0.24589	
Residual			36.70766	

```
> bdiag(VarCorr(model_final))
2 x 2 sparse Matrix of class "dsCMatrix"
      (Intercept) ses_groupmc
(Intercept) 2.3179364 0.1925678
ses_groupmc 0.1925678 0.2458869
```

## Random Coefficient Test

```
> ranova(model11)
ANOVA-like table for random-effects: Single term deletions

Model:
mathach ~ ses_groupmc + sector + meances + (ses_groupmc | id) + ses_groupmc:sector
          npar logLik   AIC     LRT Df Pr(>Chisq)
<none>           9 -23254 46526
ses_groupmc in (ses_groupmc | id)    7 -23255 46524 2.2248  2     0.3288
.

> anova(model10, model11)
Data: data_merged
Models:
model10: mathach ~ ses_groupmc * sector + meances + (1 | id)
model11: mathach ~ ses_groupmc * sector + meances + (ses_groupmc | id)
          npar   AIC   BIC logLik deviance Chisq Df Pr(>Chisq)
model10  7 46524 46572 -23255   46510
model11  9 46526 46588 -23254   46508 2.2248  2     0.3288
```

Random-effects parameters	Estimate	Std. err.	[95% conf. interval]
<b>id: Unstructured</b>			
var(ses_groupmc)	.2462764	.2248209	.0411507 1.473901
var(_cons)	2.318041	.3610305	1.708237 3.145532
cov(ses_groupmc,_cons)	.1924682	.2154886	-.2298816 .6148181
var(Residual)	36.70748	.6256972	35.50139 37.95454

LR test vs. linear model:  $\text{chi2}(3) = 217.17$  Prob > chi2 = 0.0000

## Random Intercept + Coefficient Test

Conditional intraclass correlation

Level	ICC	Std. err.	[95% conf. interval]
id	.0593981	.0087957	.0443271 .0791687

Note: ICC is conditional on all other variables.

## Random Coefficient Test

```
. lrtest final_rand final_nonrand, stats
Likelihood-ratio test
Assumption: final_nonrand nested within final_rand

LR chi2(2) = 2.22
Prob > chi2 = 0.3288
```

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
final_nonrand	7,185	.	-23255.15	7	46524.3	46572.46
final_rand	7,185	.	-23254.04	9	46526.08	46587.99

# 참고: REML vs. MLE

- REML vs. MLE
  - REML이 Mixed Effect 모형 추정시 default 추정 방법
  - REML이나 MLE나 비슷한 회귀계수를 추정함
  - REML과 MLE의 경우 variance component 추정에서 차이가 존재
  - Level-2 unit들의 수가 적을때, MLE 분산에 대한 추정치가 REML보다 작게 계산되고, 결과적으로 좁은신뢰 구간과 biased된 유의성 검정으로 이어짐
- Likelihood Ratio test for nested models로 검정
  - 표본의 수가 작은 경우에 REML 사용을 통해 MLE를 보정, 표본의 수가 충분히 많을 경우 MLE
  - 만약 두 모형의 fixed effect들의 값이 동일하고, random effect의 값이 적다면 REML과 MLE는 혼용가능
  - 만약 두 모형의 fixed effect들의 값이 다르고, random effect의 값이 적다면 MLE를 사용해야 함
  - 현실적으로는 수준 간 상호작용 같은 경우 MLE 사용 - estimation 속도

# 참고 (샘플1)

TABLE 3  
Results of Hierarchical Linear Modeling Analyses of Expatriate Work Adjustment and Job Performance<sup>a</sup>

Variables	Work Adjustment			Job Performance		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
<i>Level 1 main effects</i>						
Age	.00 (.00)	.00 (.00)	.00 (.00)	-.01* (.00)	-.01* (.00)	-.01* (.00)
Marital status	-.07 (.08)	-.07 (.07)	-.07 (.07)	.20* (.08)	.19* (.08)	.20* (.08)
Prior Intl. experience	.00 (.01)	.00 (.01)	.00 (.01)	.01 (.01)	.01 (.01)	.01 (.01)
Assignment tenure in years	.02* (.01)	.02* (.01)	.02* (.01)	.03* (.01)	.03* (.01)	.03* (.01)
Language proficiency	.01 (.02)	.01 (.02)	.01 (.02)	.01 (.02)	.01 (.02)	.00 (.02)
Openness to experience	.38* (.06)	.31* (.07)	.30* (.06)	.04 (.07)	-.01 (.07)	-.05 (.07)
Emotional stability	.39* (.05)	.35* (.05)	.33* (.05)	-.01 (.05)	-.03 (.06)	-.09 (.06)
Job performance, 2006	.03 (.03)	.02 (.03)	.01 (.03)	.17* (.04)	.17* (.04)	.17* (.04)
Perceived support	.09* (.04)	.07 (.04)	.05 (.04)	.05 (.05)	.04 (.05)	.03 (.05)
Perceived cultural distance	-.06* (.03)	-.06* (.03)	-.06* (.03)	-.02 (.04)	-.02 (.03)	-.01 (.03)
Cross-cultural motivation					.12* (.05)	.09 (.05)
Work adjustment						.15* (.04)
<i>Level 2 main effects</i>						
Subsidiary support	-.09 (.16)	-.06 (.16)	-.04 (.17)	.27 (.15)	.28 (.15)	.28 (.15)
Cultural distance	.01 (.06)	.02 (.07)	.00 (.07)	.04 (.06)	.05 (.05)	.05 (.06)
<i>Cross-level interactions</i>						
Cross-cultural motivation × subsidiary support				-.66* (.23)		
Cross-cultural motivation × cultural distance				-.24* (.06)		
Pseudo R <sup>2</sup>	.19	.20	.22	.09	.10	.11

<sup>a</sup> n = 556 expatriates (level 1) in 31 host countries/foreign subsidiaries (level 2). Unstandardized estimates (based on grand-mean centering) are reported, with standard errors in parentheses. Pseudo R<sup>2</sup> values estimate the amount of total variance (both level 1 and level 2) in the dependent variable captured by predictors in the model.

\* p < .05  
Two-tailed test.

prior research on overall support climate (Vanderwerf & Bigley, 2002; Takeuchi et al., 2009). Both intermember reliability indexes (ICC1 = .06, ICC2 = .55,  $F_{30, 555} = 2.20, p < .05$ ) and Interrater agreement (median  $r_{wg(j)} = .95$ ) provided support for aggregating individual support scores to the subsidiary level.

Level-2 변수의 투입에 관심

(Chen et al, 2010) AOM

# 참고 (샘플2)

**Table 3** Results of Multilevel Analysis of the Effect of Monitoring and Sanctioning and Compliant (Noncompliant) Behavior of Leaders and Peers on Respondents' Willingness to Refuse Bribes

	Null Model		Model 1		Model 2		Model 3		Model 4	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
<b>Fixed Part</b>										
Constant	3.43***	.04	3.47***	.05	3.00***	.05	2.99***	.07	2.96***	.47
<b>Level 1 Main design parameters</b>										
Monitoring and sanctioning weak	-.26***	.04	-.25***	.04	-.31**	.11	-.43***	.11		
Monitoring strong and sanctioning weak	-.13**	.04	-.13**	.04	-.10	.08	-.10	.08		
Monitoring and sanctioning strong	.25***	.04	.25***	.04	.34**	.11	.23*	.11		
Peer refused gift										
Leader refused gift										
Peer refused gift * Leader refused gift	.61***	.04	.53***	.11	.43***	.11				
	.08	.06	.33*	.18	.55***	.18				
<b>Level 1 Interaction parameters</b>										
Monitoring and sanctioning weak* Peer refused gift										
Monitoring strong and sanctioning weak* Peer refused gift										
Monitoring and sanctioning strong* Peer refused gift										
Monitoring and sanctioning weak* Leader refused gift										
Monitoring strong and sanctioning weak* Leader refused gift										
Monitoring and sanctioning strong* Leader refused gift										
Monitoring and sanctioning weak* Peer and leader refused gift										
Monitoring and sanctioning strong* Peer and leader refused gift										
	.18	.18	.40*	.18						
	-.12	.12	-.11	.12						
	-.03	.18	.17	.18						
	.15	.18	.36*	.18						
	.24*	.12	.23*	.12						
	-.04	.18	.15	.18						
	-.44	.32	-.85**	.31						
	-.37*	.16	-.37*	.17						
	-.19	.32	-.58*	.31						
<b>Level 2 Respondent characteristics</b>										
Trust in management										
Trust in peer										
Work relation with leader										
Work relation with peer										
Reward satisfaction										
Job satisfaction										
Knowledge of unethical cases in organization										
Perception on scenarios presented										
Gender (ref: male)										
Female										
Education (ref: undergraduate)										
Graduate degree										
Government level (ref: central government)										
Local government (province and district)										
Work experience in present position (ref: 0–12 months)										
> 12 months										
Number of staff (ref: 0–50)										
51–100										
>100										
Position (ref: diplomat and other position)										
Position at local government										
Position at central government										
Central at local government										
Control function										
<b>Random Part</b>										
Level 2: Respondent	.64	.05	.65	.05	.67	.05	.67	.05	.57	.04
Level 1: Vignette	1.15	.03	1.11	.03	.96	.02	.95	.02	.96	.02
-2*log likelihood:	14,684.2		14,542.0		13,953.4		13,923.2		13,442.0	
N Respondent	577		577		577		577		557	
N Vignette	4,602		4,602		4,602		4,602		4,452	

Deviance

L2 Intercept:  $\tau_{00}$   
L1 Intercept :  $\sigma^2$

## Level-1 Interaction

Note: \*  $p < .10$ ; \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$  (two tailed).  
SCS: Senior civil servant.

(Silitonga et al, 2019) PAR

# 참고 (샘플3)

**Table 3.** Multilevel Analysis of Certifier Sector and Respondent-Level Characteristic Relationships With Regulatory Approach Constructs

	Strictness of Regulatory Interpretation	Administer Sanctions	Flexibility	Provide Technical Assistance
Level 2: certifier attributes				
Nonprofit certifier (reference = for-profit)	-0.06 (0.08)	0.13 (0.09)	0.17 (0.11)	-0.08 (0.13)
Public certifier (reference = for-profit)	-0.13 (0.09)	0.17 (0.11)	-0.13 (0.13)	-0.37 (0.15)***
State market share	0.05 (0.15)	-0.31 (0.16)*	0.09 (0.20)	0.11 (0.21)
Number of clients (log transformed)	-0.04 (0.03)	0.08 (0.04)**	-0.11 (0.05)**	-0.11 (0.06)*
Level 1: respondent attributes				
State HHI	-0.02 (0.19)	0.31 (0.21)	-0.10 (0.26)	0.12 (0.29)
Profit motive	0.02 (0.02)	0.05 (0.02)**	0.01 (0.03)	-0.03 (0.03)
Organic movement motive	0.10 (0.02)***	0.03 (0.02)	0.07 (0.03)**	0.07 (0.03)**
Regulatory assessment	0.11 (0.03)***	0.03 (0.03)	0.07 (0.04)*	0.16 (0.04)***
Regulatory experience	0.02 (0.03)	0.07 (0.03)**	-0.02 (0.04)	-0.02 (0.04)
Predisposition towards laws	0.00 (0.03)	0.05 (0.03)	-0.04 (0.04)	-0.01 (0.04)
Operation size	0.03 (0.03)	0.08 (0.03)**	-0.06 (0.04)	-0.04 (0.04)
Pre-NOP certification	-0.04 (0.06)	0.08 (0.06)	-0.05 (0.08)	-0.12 (0.08)
Handling certification scope	0.07 (0.07)	0.00 (0.08)	0.09 (0.10)	-0.08 (0.10)
Constant	1.00 (0.25)***	-0.18 (0.30)	0.92 (0.38)**	1.31 (0.43)***
Percent variation explained by level 2 (certifier)	0.0%	1.9%	1.6%	2.8%
Log likelihood	-869	-911	-1,076	-1,112
N	784	766	761	777

Note: SEs in parentheses; level 2 (certifier) N = 41.

\*Significant at  $p < .10$ ; \*\*significant at  $p < .05$ ; \*\*\*significant at  $p < .01$ .

# 참고 (샘플4)

Table 2. Summary of hierarchical linear modelling results.

	Model 1			Model 2			Model 3			Model 4		
	$\beta$	SE	t									
<i>Control variables</i>												
(Level 1)												
Intercept	-.03	.05	-.54	-.03	.05	-.54	-.03	.05	-.55	-.02	.05	-.48
Gender	-.11**	.02	-4.26	-.08**	.02	-3.46	-.08**	.02	-3.45	-.08**	.02	-3.48
Age	.07	.06	1.29	.07	.05	1.33	.07	.05	1.32	.07	.05	1.37
Education	.02	.02	.82	.01	.02	.59	.01	.02	.58	.01	.02	.5
Work experience	.02	.06	.27	.01	.05	.24	.01	.05	.25	.01	.05	.22
Job rank	-.01	.03	-.18	-.03	.03	-1.08	-.03	.03	-1.06	-.03	.03	-1.11
(Level 2)												
Organization type	-.09	.05	-1.69	-.05	.05	-1.09	-.05	.05	-.89	-.05	.05	-1.02
<i>Predictor variables</i>												
(Level 1)												
Peer satisfaction				.46**	.02	21.79	.46**	.02	21.78	.45**	.02	21.42
(Level 2)												
Agency power												
Interaction effect												
Peer satisfaction x Agency power												
Cross-level Interaction												
N	1781			1781			1781			1781		
R <sup>2</sup>	.03			.23			.23			.24		
AIC	4957.05			4544.07			4549.73			4551.62		

\* p < .05, \*\* p < .01.

Fit

(Kim and Eun, 2022) JAPP

