Multilevel data and traditional analytic approaches

Lecture 17

Data Analysis III

Psychology 613 – Spring 2022

Multilevel data

Also known as:

- random coefficient model
- hierarchical linear model (HLM)
- variance components model
- mixed model

An appropriate technique for the analysis of *nested* or *clustered* data

Multilevel data

Examples: students within classrooms
siblings within families
workers within companies
respondents within interviews
repeated measures within individuals

Notation: Level 1 / Level 2 (often “L1/L2”)
lower / upper levels
micro / macro levels

Group similarity

Natural groups of individuals are often similar

Why? Shared group experiences

Interaction within the group

Non-randomly distributed background variables

→ Observations are **not independent**

The ICC

"Intraclass correlation coefficient":

An index of *within-group similarity*

Denoted as "ICC", ρ, r

Based on ANOVA-style partitioning of var:

Total variance =variance between groups +
variance within groups

ICC = between group variance / total variance
= between / (between + within)

The ICC: Example

In the context of a one-way ANOVA with 2 groups:

How much variance of the total variance (100%) can be attributed to the "clustering" by groups?

Y	x_code
1	0
2	0
3	0
4	0
5	0
6	1
7	1
8	1
9	1
10	1

Group 1

Group 2

The ICC: Example

Compute a one-way ANOVA with 2 groups:

model = lm(Y ~ factor(x_code))

anova(model)

```
> anova(model)
Analysis of Variance Table

Response: y
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(code)  1   62.5    62.5      25 0.001053 **
Residuals     8   20.0     2.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ICC: Example

ICC for our one-way ANOVA with 2 groups:

SS-between	= 62.5
SS-within (i.e., SS-error)	= 20.0
SS-total	= 82.5

ICC = 62.5/ 82.5 = .758 = **76%**

Interpreting the ICC

ICCs range from -1 to 1, but effectively 0 to 1.

= Proportion of total variance that is between groups

→ *Inversely related to within-group noise*

= Expected correlation between randomly chosen pairs of observations in the same group

Interpreting the ICC

If ICC = 0

→ All variance is within-group variance

i.e., the grouping doesn't matter at all

If ICC = 1

→ All variance is between groups

i.e., everyone in a group has exactly the same score

Can conduct significance test of ICC

Example: Applying OLS models to clustered data

Effect of educational attainment on income

4313 workers nested within twelve industries

Significant ICC (i.e., industry matters)

Artificially restructure the dataset into single level (two ways of doing this)

Way 1: Disaggregation

Disaggregation = ignore the clustering

Run analysis as if observations were independent

This would be fine if and only if ICC=0

If not, then this violates the GLM independence-of-errors assumption…

→ Parameter estimates will be OK, BUT

→ SEs will be too small, test stats too big

→ Type I error rate is inflated

Type 1 error rate: How bad?

Depends on (1) ICC and (2) group size

Barcikowski (1981) provide an empirical α for T-versus-C comparisons by group size and ICC:

	ICC			
N per cell	.00	.01	.05	.20
10	.05	.06	.11	.28
25	.05	.08	.19	.46
50	.05	.11	.30	.59
100	.05	.17	.43	.70

Disaggregated Example

OLS regression (disaggregated) in the education / income dataset:

$Y_{ij} = 8.98 + 0.70*X_{ij}$ (n > 4300)

Significant but small positive relation: higher individual education is related to higher individual income

BUT, significance test is suspect (why?)

Way 2: Aggregation

Aggregation = examine only the group means

- Run analysis by averaging the IV and DV of the observations within each group.
  - *This would be fine if ICC=1*

Treats the *group means* as independent observations (and ignores within-group var.)

→ Losing the advantage of a large N

→ SEs will be too big, test stats too small

→ Type II error rate is inflated

Aggregated Example

OLS regression (aggregated) in the education / income dataset:

$$Y_{.j} = 27.71 - 5.91 * X_{.j} \qquad (n = 12)$$

Significant negative relationship: industries with high average levels of education tend to have relatively low average income levels

BUT, this doesn't mean the same thing…

Problems with aggregation

Discards individual variability within group (e.g., within an industry)

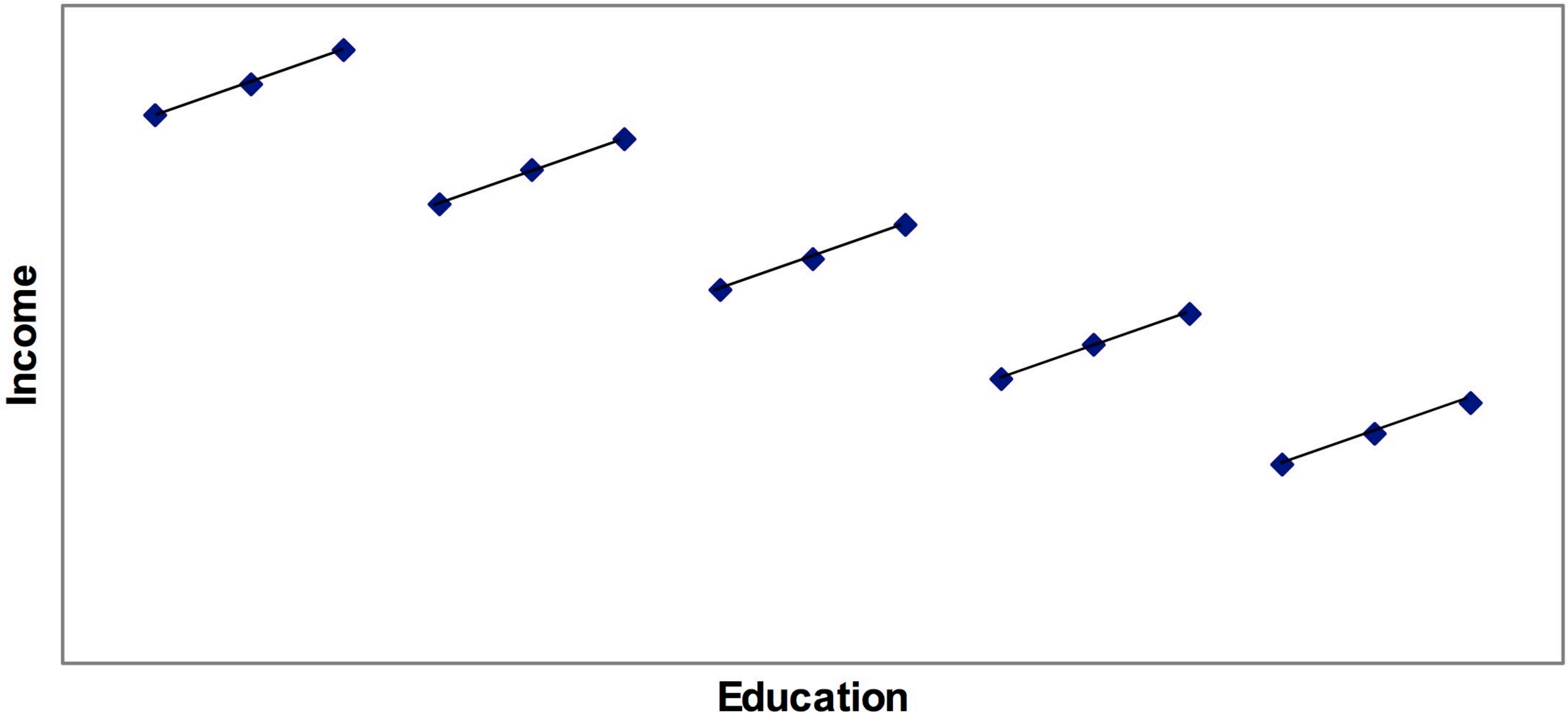
Df (and power) are reduced

- Offset by increased reliability of group means?

Shift in *meaning*: aggregated variable may not represent the same construct as individual level

Ecological fallacy: aggregate level effects do not necessarily parallel individual level effects

(Exaggerated) graph of the effects



Total, between, and within analysis

“WABA”: within and between analysis

AKA contextual analysis

AKA Cronbach analysis

Advantage: Distinguish within-group and between-group processes in a single model

Total, between, and within analysis

How?

- Calculate deviation score of each observation from group mean: $X_{ij} - X_{.j}$ (i.e., ε in ANOVA)

  Estimates a within-group effect

  Assumed identical for all groups

- Calculate a deviation score of each group mean from the grand mean: $X_{.j} - X_{..}$

  Estimates a group effect

- Enter both effects simultaneously

WABA example

In the education / income example:

“Within-group” effect

$$Y_{ij} = 27.72 + 1.10(X_{ij} - X_{.j}) - 5.91(X_{.j} - X_{..})$$

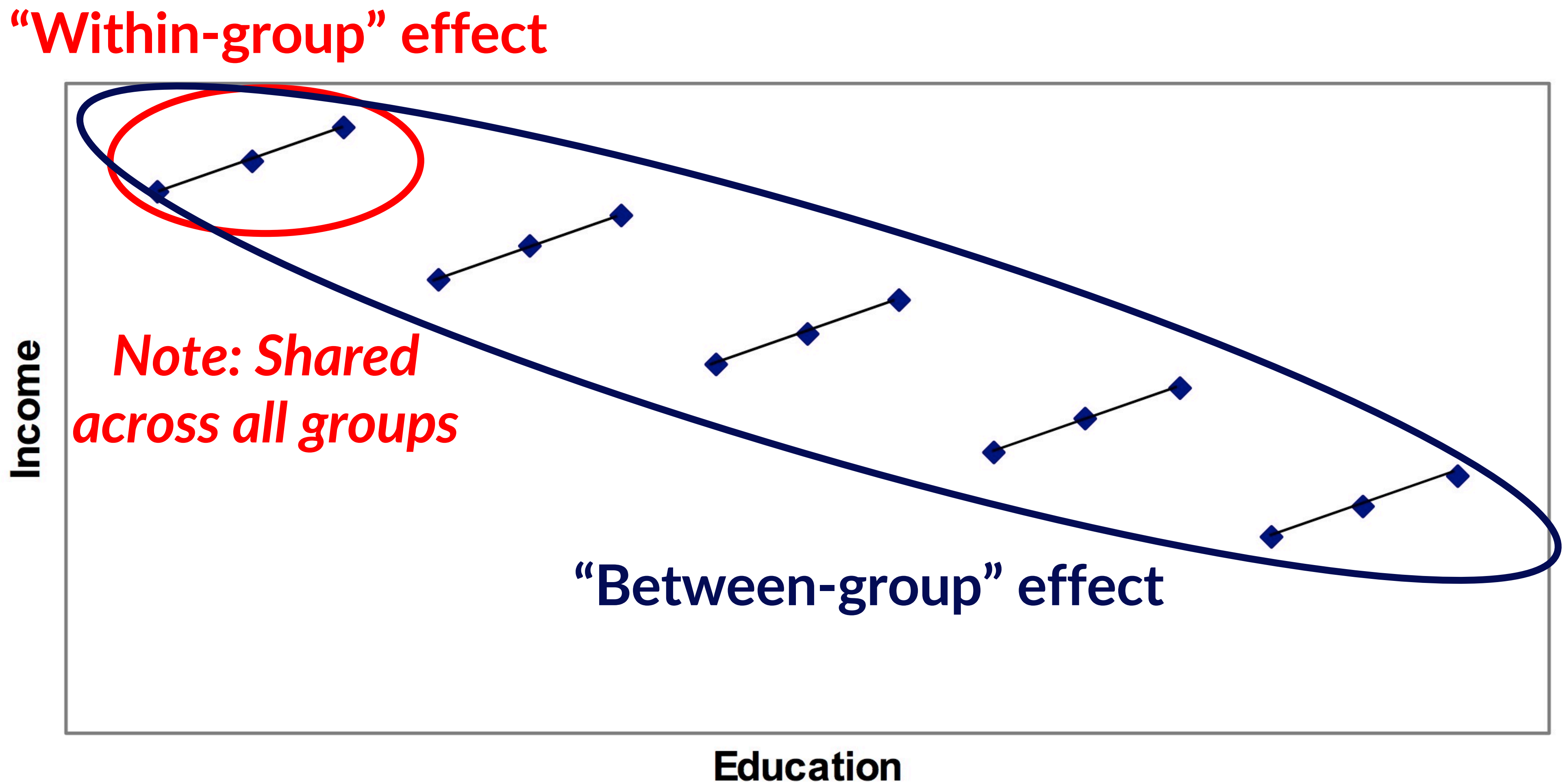
(N > 4300)

“Between-group” effect

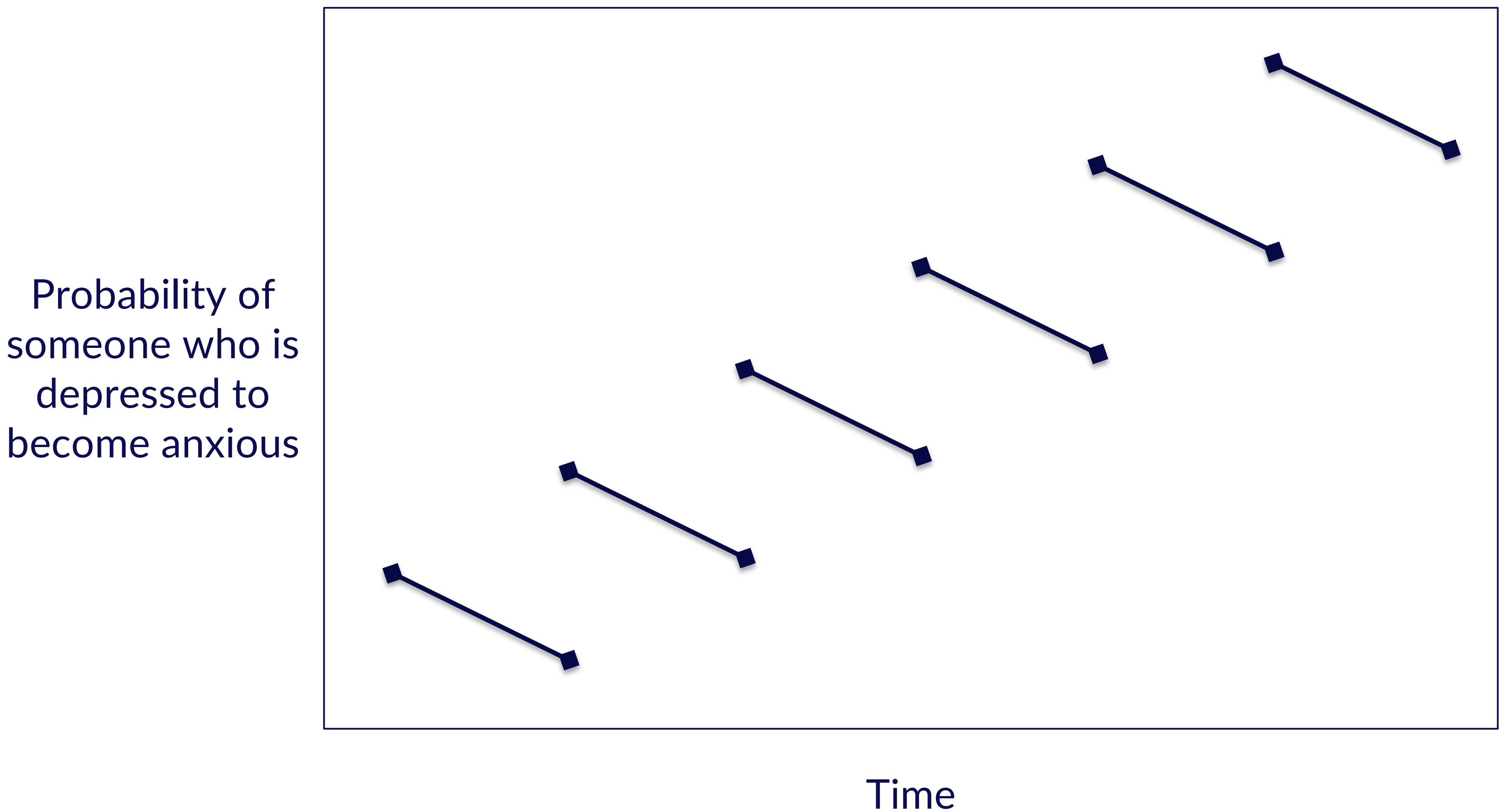
Within group effect is significant and positive

Between group effect is significant and negative

(Exaggerated) graph of the effects



(Another) example: Anxiety and depression across time



Total, between, and within effects

WABA tests within- and between-group effects

Aggregate analysis tests between-group effect

Disaggregate tests the total effect: A weighted combination of the within- and between-group effects

The higher the ICC, the more heavily the between-group effect is weighted

Problems with WABA

WABA treats grouped observations as independent

Still a disaggregation model in that sense

Miraculous multiplication of data:

group	sub	Y_{ij}	X_{ij}	$X_{.j}$	W_j
1	1	5.2	3	4	3
1	2	3.1	5	4	3
2	3	4.1	6	5	1
2	4	2.1	4	5	1

Only 2 unique means, but 4 observations → Inaccurate *df*

Applies to all group-level variables (W_j) in disaggregation models