



Why Multilevel Analysis?

Multilevel Data Structures and the Limits of Regression

Marco R. Steenbergen



University of
Zurich ^{UZH}

Institut für Politikwissenschaft

Part I

Multilevel Data Structures



Multilevel Data Defined

Definition

Multilevel data consist of multiple units of analysis, which stand at least in a partial hierarchy and contain one or more outcome variables that vary across the units.

Some Jargon

Each unit of analysis is defined in terms of its **level** in the data structure, with units placed higher in the hierarchy receiving a higher level.

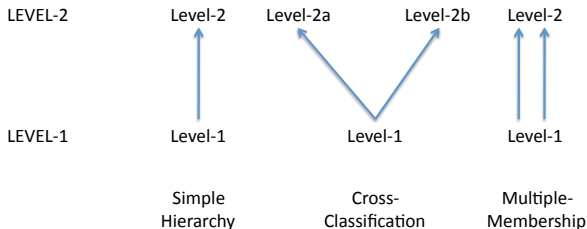


Varieties of Multilevel Data

1. Hierarchical data structures
2. Cross-classified data structures
3. Multiple membership data structures



Classification Diagrams



After Browne, Goldstein, and Rasbash (2001).



Common Levels in Electoral Research

Level-1

- Measures (in panel data)
- Individuals (in cross-sectional data)

level-2

- Individuals (in panel data)
- Geographic units (polling districts, regions, countries, etc.)
- Temporal units (in repeated cross-sections)
- Social units (networks)



Why Multilevel Data?

- Rich understanding of different sources of variation.
- Exploring “causal” heterogeneity (Western 1998).
- Increasing the sample size.



Course Limitations

- Hierarchical data structures only.
- 2-level structures only.
- Cross-sectional data only.



University of
Zurich ^{UZH}

Institut für Politikwissenschaft

Part II

The Limits of Regression Analysis



Some Notation

- Let Y_{ij} be the outcome for level-1 unit i nested inside level-2 unit j .
- Let X_{ij} denote a level-1 predictor.
- Let Z_j denote a level-2 predictor:
 - Derived variables—summaries of level-1 predictors
 - Integral variables—level-2 constructs that have no level-1 equivalents

(The terminology is based on Diez Roux 2003a, b and deviates from Hox 2010.)



Regression Analysis

- We could run the following regression analysis:

$$y_{ij} = \beta_0 + \beta_1 z_j + \beta_2 x_{ij} + \epsilon_{ij}$$

- This would be a pooled analysis (cf. panel data analysis).
- It suffers from two major problems:
 - Bias
 - Incorrect standard errors

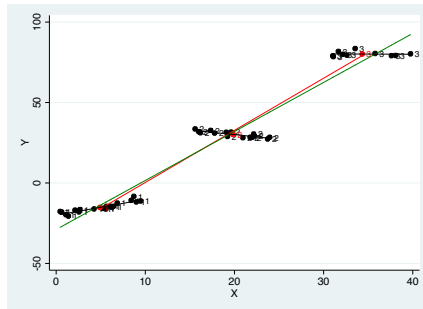


Bias

- Bias arises due to “causal” heterogeneity—the effect of X varies, so that we need an interaction.
- If no level-2 predictors are included, this may also be a source of bias.
- In both cases, we are dealing with an omitted variables bias.



An Illustration of “Causal” Heterogeneity



The black lines are the within regressions; the red line is the between regression; the green line is the pooled regression, which is a “compromise” between the within and between regression.



Incorrect Standard Errors

- The standard errors are incorrect due to **clustering**:
 - The phenomenon that observations from within the same context are not independent
- The degree of clustering is measured through the **intra-class correlation (ICC)**:
 - A correlational measure of how homogeneous observations from within the same context are—how much do they have in common?
- In nearly all social science applications, the ICC is positive.
- This means that OLS standard errors will be *underestimated*.
- Alternatively, one can say that the **effective sample size** is smaller than the nominal size.



The Scope of the Problem

Table: Ratio of True over OLS Standard Errors

ρ	$n_j = 2$	$n_j = 5$	$n_j = 10$	$n_j = 50$	$n_j = 100$	$n_j = 1000$
.05	1.001	1.005	1.011	1.059	1.117	1.870
.10	1.005	1.020	1.044	1.221	1.411	3.315
.25	1.031	1.118	1.250	2.016	2.681	7.965
.50	1.118	1.414	1.803	3.640	5.074	15.835
.75	1.250	1.803	2.462	5.344	7.529	23.726
1.00	1.414	2.236	3.162	7.071	10.000	31.623

Notes: Results for a simple regression model under the assumption that the ICC (ρ) is the same for X and Y . n_j is the cluster size, which is assumed to be constant across clusters.



Solution 1: Fixed Effects

- Introduce context dummies, possibly interacted with level-1 predictors.
- But:
 - This is quite inefficient
 - It eliminates the opportunity to introduce substantive level-2 predictors
 - It is not clear whether this can be used in nonlinear models (incidental parameter problem)



Solution 2: Cluster-Corrected Standard Errors

- Treat clustering as a statistical nuisance that is addressed by applying cluster-corrected standard errors.
- But:
 - This addresses the problem of standard errors only
 - It is often not clear how to take care of clustering in the level-1 predictors
 - This does not easily generalize to higher level models
 - This does not easily generalize to nonlinear models



Solution 3: Multilevel Analysis

- This is a good general purpose solution for models of all kinds.
- It works well when the model is correctly specified and when other conditions, to be discussed later, are satisfied.