# Chapter 11: Multilevel Structures

This chapter covers basic mulitlevel models, surveys data that have multilevel structure (grouped data, repeated measurements, time-series crosss sections, non-nested structures), and outlines costs and benefits of multilevel modeling over classical regresion.

## Basic regression models for grouped data

For data in which there are $J$ groups, there are three basic regression models: varying-intercept, varying-slope, and varying-intercet and varying-slope.

**Varying-intercept model:**

$$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i$$

**Varying-slope model:**

$$y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i$$

**Varying-intercept, varying-slope model:**

$$y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$$

Estimating all $\alpha_j$'s and $\beta_j$'s can be challenging. The rough method in multilevel modeling is to first set up a regression with varying coefficients (which vary by group). The next step is to fit a regression model for the coefficients.

## Modeling setups for clustered (grouped) data (child support enforcement in cities)

Example: You're given data on an observational study on the effect of city policies on enforcing child support payments from unmarried fathers. Enforcement/treatment occurs at the city level, and outcomes are measured by looking at individual families.

Given city- and individual-level predictors for each family, the goal is to estimate the probability that the mother receives informal support.

The layout of the data consists of a individual-level (family-level) data matrix, where city indicators are one of the predictors. Then there is a city-level data matrix.

### Individual- and group-level models

Possible ways to analyze the above data… as a lead-in to multilevel modeling

### Individual-level regression

We can construct an individual-level logistic regression, where the target variable is whether or not the mother receives informal support. City-level information can be joined to family-level data for this regression. However, the individual-level regression ignores city-level variation beyond what's explained by the city-level predictors in the model.

In reality, two cities that have similar predictor values may actually have different effects on the target variable.

### Group-level regression on city averages

In another approach, you could perform logistic regression at the city level. You can take the city data, and append an aggregated form of the family-level data to this city data to predict the proportion of families in which fathers provide informal support.

In this regression, you can get group-level inferences. But by aggregating the family data, you lose the ability to predict individual outcomes.

### Individual-level regression with city indicators, followed by group-level regression of estimated city effects

This approach involves two steps.

First, fit a logistic regression to individual data using individual predictors, and group/city indicators. Then, perform linear regression at the city level, using the coefficients of the city indicators as the target and city-level data s predictors. This two-step analysis can run into issues when sample sizes are small in particular groups (indicator coefficients will be noisy), or if there are interactions between individual- and group-level predictors (leading to misattribution).

### Multilevel models

Multilevel modeling looks similar to the above strategy, but both steps are fitted simultaneously. There are two components. First, there is a logistic regression at the individual-level with an intercept that varies by city. Then, there is a city-level linear regression that predicts the city intercepts using city-level predictors.

The logistic regression will look like

$$P(y=1) = \text{logit}^{-1}(X_i\beta + \alpha_{j[i]}) \qquad \text{for } j = 1, ..., n$$

where $X$ refers to individual-level predictors. The second part of this model is the regression of city coefficients (intercepts)

$$\alpha_j \sim \text{N}(U_j\gamma, \sigma_\alpha^2) \qquad \text{for } j = 1, ..., 20$$

where $U$ is the matrix of city-level predictors, $\gamma$ is the vector of coefficients for the city-level regression, and $\sigma_\alpha$ is the standard deviation of unexplained group-level errors.

The model for $\alpha$ allows the inclusion of varying intercepts, without collinearity issues between city-level and individual-level predictors. I think this is because there isn't any specification of individual-level predictors in this model.

## Repeated measurements, time-series cross sections, and other non-nested structures

A survey of multilevel data structures beyond grouped data.

### Repeated measurements (longitudinal data)

Repeated measurements – in which multiple measurements are made on the same subject – are another multilevel data structure. This is also referred to as longitudinal data. This is a typical data structure for scientific/causal experiments.

### Time-series cross-sectional data

Time series is similar to repeated measurements data, but with measurements taken at more frequent and regular intervals. Here, overall time trends are of interest.

### Other non-nested structures

Non-nested data can also arise when individuals have overlapping categories of attributes. For example, a person could be classified by job and by state, although these classification levels are not necessarily nested.

## Indicator variables and fixed or random effects

In a typical classical regression, we typically include an indicator for $N-1$ categories to prevent collinearity from occurring. However, this isn't a problem in multilevel modeling because indicators are themselves modeled based on group-level idstributions.

## Costs and benefits of multilevel modeling

What classical regression achieves:

- Prediction of continuous or discrete outcomes
- Fitting nonlinear relations with transformations
- Inclusion of categorical predictors using indicator variables
- Modeling interactions between inputs
- Causal inference

### Motivations for multilevel modeling

Causal inference, study of variation, prediction of future outcomes are siome reasons.

### Benefits

- Accounting for individual- and group-level variation together while estimating group-level regression coefficients
- Modeling variation among individual-level regression coefficients; in particular, when we want to model variation of these coefficients across groups, make predictions for new groups, or account for group-level variation in uncertainty for individual-level coefficients
- Estimating regression coefficents for particular groups

### Potential costs

A potential drawback to multilevel modeling is the complexity added by modeling groups.

Another is that multilevel modeling requires additional assumptions for each level of the model, such as additivity, linearity, independence, equal variance, and normality.

### When to use multilevel modeling

When there is little group-level variation, multilevel models reduce to classicial regression with no group indicators; when group-level coefficeints vary greatly, multilevel modeling reduces to classical regression with group indicators.

When the number of groups is small, there is typically not enough information to estimate group-level variation.