Multilevel Modeling (Part 1)

Null Random-Intercept Modeling

Desmond D. Wallace

Department of Political Science
The University of Iowa
Iowa City, IA

March 23, 2018

Regression Highlights

- A way to summarize the relationship between variables.
- Assuming there is a relationship between Y and the independent variable(s).
- Relationship may be linear (OLS) or non-linear (CLDV).
- Regression helps our understanding of how our dependent variable of interest changes when one or more independent variables vary, while holding remaining variables fixed.
- PRF: $y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon_i$
- SRF: $y_i = b_0 + b_1 x_1 + \cdots + b_k x_k + e_i$

Model Specification

1. Linearity in the parameters
2. The number of observations n must be greater than the number of parameters to be estimated
3. The regression model is correctly specified

Independent Variable(s)

1. X values are fixed in repeated sampling
2. Variability in X values
3. There is no perfect multicollinearity

Error Term

1. Zero mean value of error (e_i)
2. **Homoscedasticity or equal variance** of e_i
3. No autocorrelation between the errors
4. Zero covariance between e_i and X_i

Additional Assumptions

1. Errors are normally distributed
2. **Errors for any two observations are independent of one another**

Multi-Stage Sampling

- OLS assumptions imply utilization of Simple Random Sampling (SRS)
- However, due to cost-efficiency, multi-stage sampling approaches may be utilized instead.
- Researcher may randomly sample grouping units instead of individuals (cluster sampling)
- Examples
    - Students nested in schools
    - Respondents nested in states (countries)
    - Patients nested in hospitals

Applying OLS to Multilevel Data

- Biased standard errors
- Model Misspecification

New Approach

- Best approach to analyzing nested data is a statistical approach that accounts for both within-group and between-group variation *simultaneously*
- One approach is to conceive within-group and between-group variation as random variability
- One can achieve this by including *random coefficient(s)* in the statistical model

Multilevel Model

- Multilevel Model (MLM) is a model where the parameters vary at more than one level
- Features more than one error term
- Variation can occur wit respect to the intercept (*random intercept*) and/or the slope (*random slope*)
- This approach leads to corrected standard errors and correct model specification

Multilevel Model

- Model coefficients are now a combination of both fixed and random components
  - Fixed Coefficient – An unknown constant of nature
  - Random Coefficient – One which varies from sample of groups to sample of groups
- Random coefficients are not estimated, and are instead predicted
- Instead of BLUE (Best Linear Unbiased Estimator) coefficients, we know have BLUP (Best Linear Unbiased Prediction) coefficients

Null (Variance Components) Model

- Predicting the outcome from only an intercept that varies between groups
- The 2-level null model takes the following form:
  - Level-1 Model: $y_{ij} = \beta_{0j} + \varepsilon_{ij}$
  - Level-2 Model: $\beta_{0j} = \gamma_{00} + U_{0j}$ where
    - β_{0j} – Average (general) intercept holding across all groups (fixed effect)
    - U_{0j} – Group-specific effect on the intercept (random effect)
  - Full Specification: $y_{ij} = \gamma_{00} + U_{0j} + \varepsilon_{ij}$
- Interested in general mean value for y_{ij} (γ_{00}) and deviation between overall mean and group-specific effects for the intercept (U_{0j})

Null Model Assumptions

1. Groups are a random sample from the population of all possible groups
2. U_{0j} is randomly drawn from a population distribution with mean 0 and variance τ_0^2
3. τ_0^2 (Variance of U_{0j}) and σ^2 (variance of ε_{ij}) are uncorrelated

Variance of y_{ij}

- Variance of y_{ij} is just the sum of the level-two and level-one variances
  - $Var(y_{ij}) = Var(U_{0j}) + Var(\varepsilon_{ij}) = \tau_0^2 + \sigma^2$
- Covariance between two observations from the same group (ij and $i'j$) is equal to the variance of the contribution U_{0j} that is shared by these observations
  - $Cov(y_{ij}, y_{i'j}) = Var(U_{0j}) = \tau_0^2$

Intraclass Correlation Coefficient

- A measure of the proportion of variation in the dependent variable that occurs *between* groups versus the total variation (*between* and *within*)
- Correlation between two randomly drawn observations from in one randomly drawn group
- Ranges between 0 (no variation between groups) and 1 (all between-group variation but no within-group variation)
- $\rho(y_{ij}, y_{i'j}) = \frac{\tau_0^2}{\tau_0^2 + \sigma^2}$

Example #1: Math Achievement

- Goal: Partition the value of students' math achievement scores into an overall mean and group-specific random effects.
- Model Specification
  - Level-1 Model (student-level): $\text{mathach}_{ij} = \beta_{0j} + \varepsilon_{ij}$
  - Level-2 Model (school-level): $\beta_{0j} = \gamma_{00} + U_{0j}$
  - Full Model: $\text{mathach}_{ij} = \gamma_{00} + U_{0j} + \varepsilon_{ij}$
- Parameter Interpretation
  - γ_{00}: Overall mean of student's math achievement scores
  - U_{0j}: Unique effect of school j on mean math achievement score

Example #2: Gross State Product

- Goal: Partition the value of states' gross domestic product into an overall mean, state-specific random effects, and region-specific random effects.
- Model Specification
  - Level-1 Model (repeated measures): $\text{GSP}_{ijk} = \beta_{0jk} + \varepsilon_{ijk}$
  - Level-2 Model (state-level): $\beta_{0jk} = \gamma_{00k} + U_{0jk}$
  - Level-3 Model (region-level) : $\gamma_{00k} = \delta_{000} + V_{00k}$
  - Full Model: $\text{GSP}_{ijk} = \delta_{000} + V_{00k} + U_{0jk} + \varepsilon_{ijk}$
- Parameter Interpretation
  - δ_{000}: Overall mean of GSP
  - U_{0jk}: Unique effect of state j on mean GSP
  - V_{00k}: Unique effect of region k on mean GSP

Email: desmond-wallace@uiowa.edu
Any Questions?