Nested Data Structures

Michael P. Fix
Georgia State University
mfix@gsu.edu

Week 11

Multilevel/Nested Data

- Most frequently contain observations that are nested within larger spatial categories or groupings
  - Examples: individuals within states, households within counties
- Also contain observations that are nested temporally
  - Example: Annual gross domestic product
- May even contain observations that are nested in larger spatial groupings, across time
  - Example: individual responses within surveys within years

Dealing with Multilevel/Nested Data

- Disaggregate Group Data to Individual Level
  - Example: Individual data nested with in states, include state level variables at individual level with same values for all individuals in a given state
- Problem:
  - All unmodeled contextual information (usually macro effects) ends up in the error term
  - Individuals within same macro group then have correlated errors (violates OLS assumption)
  - Can we get around this? Is there a way to account for this unmodeled contextual information?

Solution 1: Fixed Effects

- Essentially, this approach adds an additional dummy variable for each macro-level grouping to account for the contextual variation
- This prevents to correlated error issue, but does so at a cost
- Model estimates will be inefficient as N-1 new independent variables are added to the model, burning N-1 degrees of freedom (where N is the number of macro-level groupings)

Solution 2: Random Effects

- Like fixed effects, random effects allows the estimation of different intercepts for each macro-level group
- It avoids the inefficiency problem by assuming these intercepts are randomly drawn for a given (usually normal) distribution
- Estimates are likely to be biased though

Solution 3: Clustering

- Clustering essentially is a statistical “fix” of the problem by allowing a compound error term that accounts for the macro-level information
- This is a variation on the commonly used Huber-White robust standard errors
- Like random effects models it allows off diagonial elements in teh variance covariance matrix to be non-0
- See Primo, David M, Matthew L. Jacobsmeier, and Jeffrey Milyo. 2007. “Estimating the Impact of State Policies and Institutions with Mixed-Level Data” *State Politics and Policy Quarterly* 7(Winter): 446–459.

Solution 4: Multilevel Modeling

- Also known as Hierarchical Linear Modeling or Mixed Effects Modeling
- Heavily used in educational research to look at students nested in classrooms, nested within schools, nested within districts, etc
- Goal is to predict influences on a dependent variable using independent variables from several contexts (individual and macro)

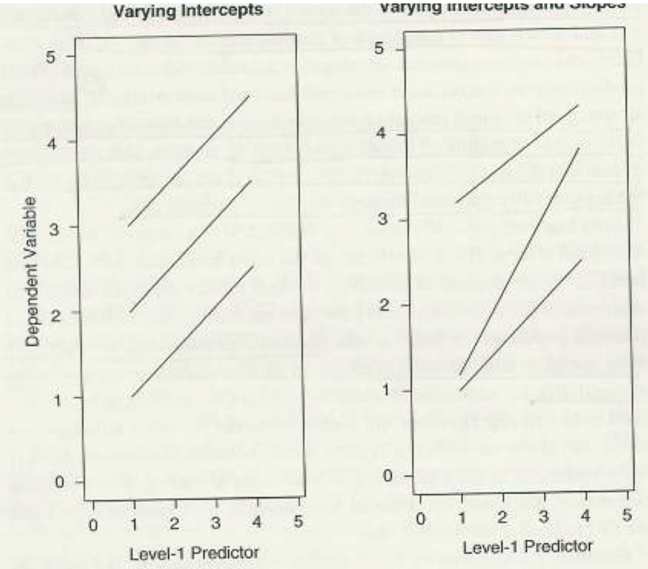
Multilevel Modeling — Basic Structure

- Consider the following equations:
- Level 1: $y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \epsilon_{ij}$
- Level 2:
  - $\beta_{0j} = \gamma_{00} + \eta_{0j}$
- Where:
  - i = individuals
  - j = groups

Multilevel Modeling Considerations

- How many levels are in the data?
  - Most social science contains only 2 or 3
- How many predictors for each level are needed?
  - Model becomes increasingly more complex as these numbers increase (especially for macro-level predictors)
  - Are any cross-level interactions hypothesized
- Which parts of the model will include random effects?
- What structural form will you use?
  - Varying intercepts only
  - Varying slopes only
  - Varying intercepts and slopes

Varying Intercepts vs Slope and Intercept



Varying Intercepts and Slopes

- Varying intercept and slope adds an additional level of modeling complexity
- Our level 1 equation remains unchanged:
- $y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \epsilon_{ij}$
- However, our level 2 model now accounts for the fact that we are allowing both the intercept and the slope to vary at level 2:
    - $\beta_{0j} = \gamma_{00} + \eta_{0j}$
    - $\beta_{1j} = \gamma_{10} + \eta_{1j}$

Multilevel Modeling Considerations

- Number of Groups
  - Some argue that a minimum number of groups is needed for multilevel modeling
  - However, even with a small number of groups, a multilevel regression will simply reduce to a classical regression
  - Therefore, the number of groups is a limitation, only in that it estimation of between-group variation will be limited
- Number of Observations per Group
  - Another issue that some scholars present as an issue even though none exists
  - With small numbers of observations in some groups, estimates of the α parameters for those groups will be imprecise
  - Also, if there is significant imbalance there can be issues with random effects estimates

Estimation in Stata

- The basic syntax for estimating a mixed effects linear regression in Stata is:
  `mixed depvar fe_equation [|| re_equation]`
  `[|| re_equation ...] [,options]`
- where:
  - `fe_equation` syntax is:
    `[indepvars] [if] [in] [weight] [, fe_options]`
  - and
  - `re_equation` syntax is:
    `levelvar: [varlist] [, re_options]`

Estimation in Stata: Example

```
. mixed cites readpca nytSalience MOWmq MinWin precedentAlteration age || court: r
> eadpca

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -305736.02
Iteration 1:   log likelihood = -305736.02

Computing standard errors:

Mixed-effects ML regression                     Number of obs     =     86,517
Group variable: court                           Number of groups  =         52

                                                Obs per group:
                                                              min =      1,663
                                                              avg =    1,663.8
                                                              max =      1,664

                                                Wald chi2(6)      =     287.32
Log likelihood = -305736.02                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
           cites |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
         readpca |   .0355832   .0108527     3.28   0.001     .0143123    .0568542
     nytSalience |   .6068268   .0808438     7.51   0.000     .4483759    .7652776
           MOWmq |   .0249962   .0135759     1.84   0.066    -.001612     .0516045
          MinWin |    .570886   .0728748     7.83   0.000     .4280539     .713718
precedentAlte~n |   1.600817   .1969919     8.13   0.000      1.21472    1.986914
             age |   .0430976   .0052521     8.21   0.000     .0328036    .0533915
           _cons |  -.2450255   .1392622    -1.76   0.079    -.5179744    .0279234
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
court: Independent           |
                var(readpca) |   .0027766   .0010918      .0012847    .0060008
                 var(_cons)  |   .2340195   .0542114       .148617    .3684985
-----------------------------+------------------------------------------------
               var(Residual) |   68.59489   .3299952      67.95115    69.24472
------------------------------------------------------------------------------
LR test vs. linear model: chi2(2) = 223.67               Prob > chi2 = 0.0000
```

Estimation in Stata: Beyond Linear Regression

- Moving from a multilevel linear regression to more complex multilevel models is quite straightforward in Stata
  - However, you will want to ensure that you understand what you are doing. Simply because the code runs, doesn't mean something is properly modeled
- Examples:
  - Multilevel logit: `melogit`
  - Multilevel probit: `meprobit`
  - Multilevel poisson: `mepoisson`
  - Multilevel negative binomial: `menbreg`
  - Multilevel ordered logit: `meologit`
  - Multilevel ordered probit: `meoprobit`