Multilevel models

When to use them, how they differ from OLS regression, and how to implement them in Stata and R

Benjamin Rosche - benrosche.com

Cornell Population Center - Graduate Training Seminar - Spring 2022

March 4, 2022

Content

Acknowledgements and References

This presentation draws on examples and equations from:

- Bullen, Jones & Duncan (1997). Modelling complexity: analysing between-individual and between-place variation. *Environment and Planning A, 29(4), 585-609*.
- Gelman & Hill (2007): Data Analysis Using Regression and Multilevel/Hierarchical Models
- Germán Rodríguez's (Princeton) excellent website
- Goldstein (2011): Multilevel Statistical Models
- Rabe-Hesketh & Skrondal (2012): Multilevel and Longitudinal Modeling Using Stata.
- Raudenbush & Bryk (2002): Hierarchical Linear Models. Applications and Data Analysis Methods.
- Snijders & Bosker (1999): Multilevel modeling. An introduction to basic and advanced multilevel modeling.

What are multilevel structures?

What are multilevel structures?

Many kinds of data have a *multilevel* / *hierarchical* / *nested* / *clustered* structure

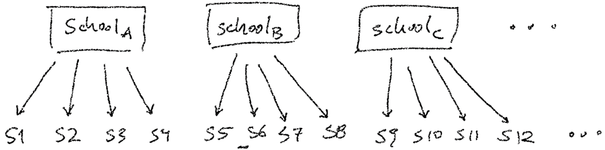

Figure 1: Examples of multilevel structures: students nested in schools, household members nested in households, citizens nested in countries.

What are multilevel structures?

Many kinds of data have a *multilevel* / *hierarchical* / *nested* / *clustered* structure

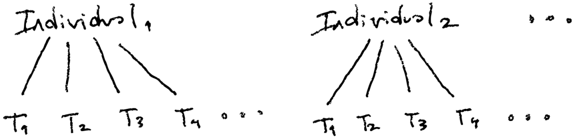

Figure 2: Panel data analysis as multilevel problem: measurement occasions nested in individuals.

What are multilevel structures?

Many kinds of data have a *multilevel* / *hierarchical* / *nested* / *clustered* structure

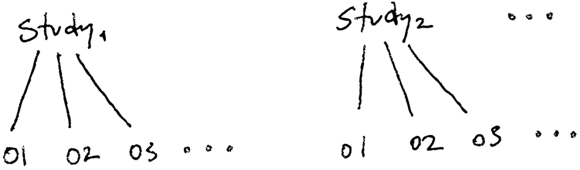

Figure 3: Meta analysis as multilevel problem: observations nested in studies.

What are multilevel structures?

Many kinds of data have a *multilevel* / *hierarchical* / *nested* / *clustered* structure

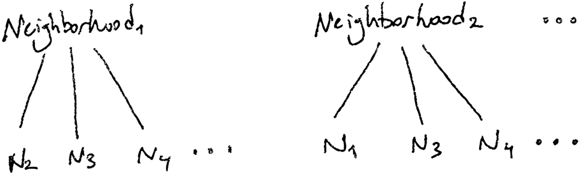

Figure 4: Spatial data analysis as multilevel problem*: neighborhoods nested in other neighborhoods.

What are multilevel structures?

Many kinds of data have a *multilevel* / *hierarchical* / *nested* / *clustered* structure

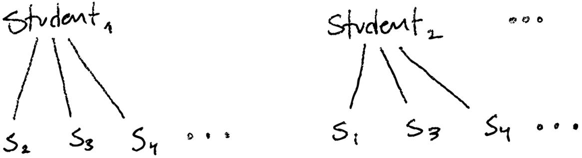

Figure 5: Network analysis data as multilevel problem*: egos nested in alters.

What are multilevel structures?

Clustering is not always *perfectly hierarchical* (= each lower-level unit is nested in one higher-level unit).

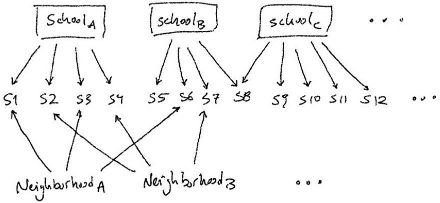

Figure 6: Students nested in schools and neighborhoods. Visible are hierarchical, cross-classified, and multiple-membership structures.

- *Cross-classified*: Lower-level units are clustered in different higher-level units (e.g., students in schools and neighborhoods).
- *Multiple-memberships*: Lower-level units are clustered in more than one higher-level unit (e.g., students have attended more than more school). With this extension, spatial and network data can be analyzed.

Why do we want to recognize multilevel structure?

- Clustering as a nuisance
  1. Properly account for uncertainty in estimation and prediction due to the clustering structure
- Clustering as an interesting phenomenon
  1. *Learn about variability within and between groups*
  2. *Learn about effect heterogeneity*
  3. *Learn whether the within-group effect and the between-group effect of a predictor differ*
  4. *Improve group-level inference and prediction*

Why do we want to recognize multilevel structure?

- Clustering as a nuisance
  1. Properly account for uncertainty in estimation and prediction due to the clustering structure
- Clustering as an interesting phenomenon
  1. *Learn about variability within and between groups*
  2. *Learn about effect heterogeneity*
  3. *Learn whether the within-group effect and the between-group effect of a predictor differ*
  4. *Improve group-level inference and prediction*

Clustering as a nuisance

Making the multilevel problem disappear

Two problematic approaches:

1. Aggregation
   - Aggregating individual-level variables changes their meaning
   - Inferences about individual-level mechanisms cannot be made from aggregated data (ecological fallacy)
   - Cross-level interactions cannot be analyzed
2. Disaggregation
   - Disaggregation of group-level data exaggerates our sample size and, therefore, induces excessive Type-I error.

→ Multilevel modeling overcomes these problems by jointly analyzing within- and between-group relationships.

Independence of observations

Standard errors in the OLS regression model require the *independence of observations*, which is violated with clustered data because observations within clusters are more similar than between clusters.

Example:

- Take y_i to be the GPA of student i nested in school j and assume the outcome is a function of a independent school-specific effect u_j and a independent student-specific effect e_i: $y_i = u_{j[i]} + e_i$.
- Accordingly, the variance in the outcome is $var(y_i) = \sigma_u^2 + \sigma_e^2$
- We can define a *variance partition coefficient* $VPC_y = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}$, which measures the proportion of variance at the 2nd level.
- The more variance at the school level, the more similar the GPA of students within the same school.

Relationship between SE_{True} and SE_{OLS}

- Consider this OLS regression model: $y_i = \beta_0 + \beta_1 X_i + e_i$
- Whether observations are independent (i.e, SE_{β_1} is correct), depends on how much variance in X and y is at the 2nd level.
- The relationship between the SE_{True} and SE_{OLS} equals:

$$SE_{True} = SE_{OLS} \times \left\{1 + VPC_X VPCy(n-1)\right\}^{\frac{1}{2}}$$

where n = number of l1 units per l2 unit

→ The SE_{OLS} will be too small as soon as there is variance in X and in y at the 2nd level.

* This equation holds for for constant n and one explanatory variable

Alternative approaches to ML modeling

- Alternatively, researchers can draw on cluster-robust SE to correct for clustering structure.
- In this strategy, an OLS regression model is estimated and then, post estimation, cluster-robust SE are calculated (see White 1984; Liang & Zeger 1986; Arellano 1987)
- Cluster-robust SE do not require specification of a model for within-cluster error correlation, but require that the number of observations *and the numbers of clusters* go to infinity.
- A practioner's guide: Cameron & Miller (2015)

The multilevel model

The varying intercept model

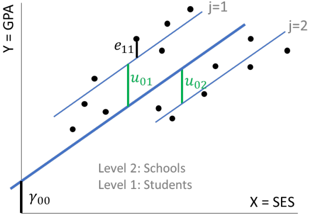

Figure 7: The effect of SES on GPA of students nested in schools. The figure shows two school-specific intercepts.

- Model without l2 predictor:

$$y_i = \beta_{0j[i]} + \beta_1 X_i + e_i \text{ with}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\rightarrow y_i = \gamma_{00} + \beta_1 X_i + u_{0j[i]} + e_i$$

- Model including l2 predictor:

$$y_i = \beta_{0j[i]} + \beta_1 X_i + e_i \text{ with}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$$
$$\rightarrow y_i = \underbrace{\gamma_{00} + \gamma_{01} Z_{j[i]} + \beta_1 X_i}_{\text{fixed part}} + \underbrace{u_{0j[i]} + e_i}_{\text{varying part}}$$

- Distributional assumptions:

$$y_i \sim N(\beta_{0j[i]} + \beta_1 X_i, \sigma_e^2)$$
$$\beta_{0j} \sim N(\gamma_{00} + \gamma_{01} Z_j, \sigma_u^2)$$

Notation: i indexes l1 units, j indexes l2 units, $j[i]$ is an indexing function returning the j in which i is nested, X is a l1 predictor, Z is a l2 predictor, β_{0j} are the varying intercepts, γ_{00} is the grand intercept, u_{0j} are the group-specific deviations from the grand intercept, and $\beta_1 + \gamma_{01}$ are regression coefficients for the l1 + l2 predictors

The varying intercept model

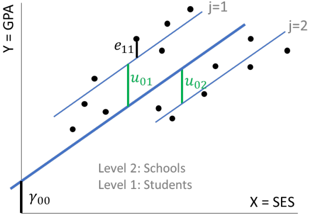

Figure 7: The effect of SES on GPA of students nested in schools. The figure shows two school-specific intercepts.

- Model without l2 predictor:

$$y_i = \beta_{0j[i]} + \beta_1 X_i + e_i \text{ with}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\rightarrow y_i = \gamma_{00} + \beta_1 X_i + u_{0j[i]} + e_i$$

- Model including l2 predictor:

$$y_i = \beta_{0j[i]} + \beta_1 X_i + e_i \text{ with}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$$
$$\rightarrow y_i = \underbrace{\gamma_{00} + \gamma_{01} Z_{j[i]} + \beta_1 X_i}_{\text{fixed part}} + \underbrace{u_{0j[i]} + e_i}_{\text{varying part}}$$

- Distributional assumptions:

$$y_i \sim N(\beta_{0j[i]} + \beta_1 X_i, \sigma_e^2)$$
$$\beta_{0j} \sim N(\gamma_{00} + \gamma_{01} Z_j, \sigma_u^2)$$

Notation: i indexes l1 units, j indexes l2 units, $j[i]$ is an indexing function returning the j in which i is nested, X is a l1 predictor, Z is a l2 predictor, β_{0j} are the varying intercepts, γ_{00} is the grand intercept, u_{0j} are the group-specific deviations from the grand intercept, and $\beta_1 + \gamma_{01}$ are regression coefficients for the l1 + l2 predictors

The varying intercept model

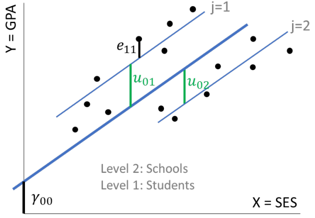

Figure 7: The effect of SES on GPA of students nested in schools. The figure shows two school-specific intercepts.

- Model without l2 predictor:

$$y_i = \beta_{0j[i]} + \beta_1 X_i + e_i \text{ with}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\rightarrow y_i = \gamma_{00} + \beta_1 X_i + u_{0j[i]} + e_i$$

- Model including l2 predictor:

$$y_i = \beta_{0j[i]} + \beta_1 X_i + e_i \text{ with}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$$
$$\rightarrow y_i = \underbrace{\gamma_{00} + \gamma_{01} Z_{j[i]} + \beta_1 X_i}_{\text{fixed part}} + \underbrace{u_{0j[i]} + e_i}_{\text{varying part}}$$

- Distributional assumptions:

$$y_i \sim N(\beta_{0j[i]} + \beta_1 X_i, \sigma_e^2)$$
$$\beta_{0j} \sim N(\gamma_{00} + \gamma_{01} Z_j, \sigma_u^2)$$

Notation: i indexes l1 units, j indexes l2 units, $j[j]$ is an indexing function returning the j in which i is nested, X is a l1 predictor, Z is a l2 predictor, β_{0j} are the varying intercepts, γ_{00} is the grand intercept, u_{0j} are the group-specific deviations from the grand intercept, and $\beta_1 + \gamma_{01}$ are regression coefficients for the l1 + l2 predictors

The varying-intercept model in Stata

Stata commands

```
mixed y X Z || gid:
xtreg y X Z, re i(gid) // can only do random intercepts
```

Example (Dataset from Snijders & Bosker 1999):

```
mixed gpa ses clubs || schoolnr:

Mixed-effects ML regression                     Number of obs     =      2,287
Group variable: schoolnr                        Number of groups  =        131

------------------------------------------------------------------------------
         gpa |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ses |   .3574069   .0210423    16.99   0.000     .3161648     .398649
       clubs |   .0787655    .043304     1.82   0.069    -.0061087    .1636397
       _cons |  -.0350527   .0423598    -0.83   0.408    -.1180764    .0479711
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
schoolnr: Identity           |
                  var(_cons) |   .1851497    .029573      .1353833      .25321
-----------------------------+------------------------------------------------
               var(Residual) |   .7030494   .0214484      .6622435    .7463696
------------------------------------------------------------------------------
LR test vs. linear model: chibar2(01) = 272.99        Prob >= chibar2 = 0.0000
```

The varying-intercept model in R

R commands:

```
library(lme4)
lmer(y ~ 1 + X + Z + (1 | gid), ...)
```

Example:

```
summary(lmer(gpa ~ 1 + ses + clubs + (1 | schoolnr), REML=F, dat))

Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: gpa ~ 1 + ses + clubs + (1 | schoolnr)
   Data: dat

Random effects:
 Groups   Name        Variance Std.Dev.
 schoolnr (Intercept) 0.1851   0.4303
 Residual             0.7030   0.8385
Number of obs: 2287, groups:  schoolnr, 131

Fixed effects:
            Estimate Std. Error t value
(Intercept) -0.03505    0.04236  -0.827
ses          0.35741    0.02104  16.985
clubs        0.07877    0.04330   1.819
```

The varying intercepts visualized

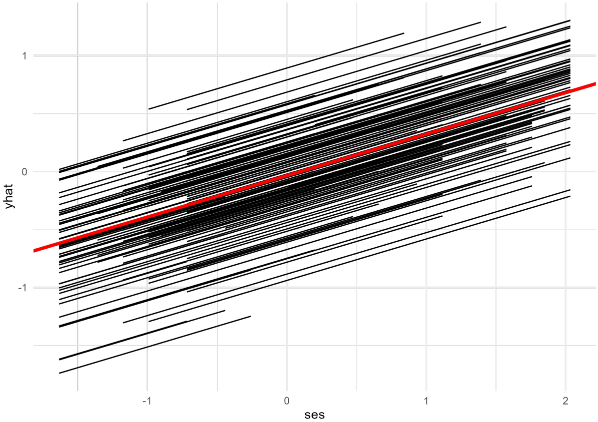

Figure 8: The variance around the grand intercept (red) is estimated to be 0.185. The variance around each school-specific intercepts is estimated to be 0.703.

The varying slope model

- Without l2 predictor:

$$y_i = \beta_{0j[i]} + \beta_{1j[i]} X_i + e_i \text{ with}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$
$$\rightarrow y_i = \underbrace{\gamma_{00} + \gamma_{10} X_i}_{\text{fixed part}} + \underbrace{u_{0j[i]} + u_{1j[i]} X_i + e_i}_{\text{varying part}}$$

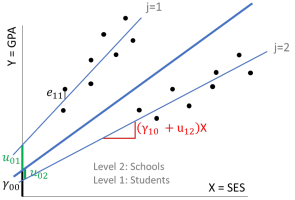

Figure 9: The effect of SES on GPA depends on the school

- Including l2 predictor:

$$y_i = \beta_{0j[i]} + \beta_{1j[i]} X_i + e_i \text{ with}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}$$
$$y_i = \underbrace{(\gamma_{00} + \gamma_{01} Z_j + u_{0j[i]})}_{\text{intercept}} + \underbrace{(\gamma_{10} X_i + \gamma_{11} Z_{j[i]} X_i + u_{1j[i]} X_i)}_{\text{slope}} + e_i$$

- $\gamma_{11} Z_{j[i]} X_i$ is called a *cross-level interaction, which explains the group-specific slope.*

The varying slope model

- Without l2 predictor:

$$y_i = \beta_{0j[i]} + \beta_{1j[i]} X_i + e_i \text{ with}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$
$$\rightarrow y_i = \underbrace{\gamma_{00} + \gamma_{10} X_i}_{\text{fixed part}} + \underbrace{u_{0j[i]} + u_{1j[i]} X_i + e_i}_{\text{varying part}}$$

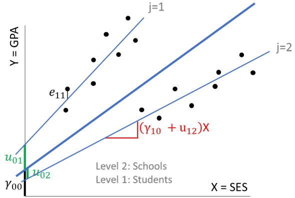

Figure 9: The effect of SES on GPA depends on the school

- Including l2 predictor:

$$y_i = \beta_{0j[i]} + \beta_{1j[i]} X_i + e_i \text{ with}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}$$
$$y_i = \underbrace{(\gamma_{00} + \gamma_{01} Z_j + u_{0j[i]})}_{\text{intercept}} + \underbrace{(\gamma_{10} X_i + \gamma_{11} Z_{j[i]} X_i + u_{1j[i]} X_i)}_{\text{slope}} + e_i$$

- $\gamma_{11} Z_{j[i]} X_i$ is called a *cross-level interaction, which explains the group-specific slope.*

The varying-slope model in Stata

Stata commands:

```
mixed y X || gid: X // random slope for X
mixed y X Z X#Z || gid: X // Z explaining random intercept and random slope (=cross-level interaction)
```

Example:

```
mixed gpa c.ses c.clubs c.ses#c.clubs || schoolnr: ses, mle covariance(unstructured)

Mixed-effects ML regression                     Number of obs     =      2,287
Group variable: schoolnr                        Number of groups  =        131

--------------------------------------------------------------------------------
            gpa |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+---------------------------------------------------------------
            ses |   .3687384   .0225306    16.37   0.000     .3245791    .4128976
          clubs |   .0710318   .0422582     1.68   0.093    -.0117927    .1538564
                |
  c.ses#c.clubs |  -.0611543   .0222428    -2.75   0.006    -.1047494   -.0175592
                |
          _cons |  -.0124706   .0423211    -0.29   0.768    -.0954185    .0704773
--------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
schoolnr: Unstructured       |
                    var(ses) |   .0073425   .0067279      .0012187    .0442381
                  var(_cons) |   .1736029   .0277884      .1268547    .2375789
              cov(ses,_cons) |  -.0283662   .0106466      -.049233   -.0074993
-----------------------------+------------------------------------------------
               var(Residual) |   .6969668   .0216296      .6558371    .7406759
```

The varying-slope model in R

R commands:

```
library(lme4)
lmer(y ~ 1 + X + (1 + X | gid), ...) # random slope for X
lmer(y ~ 1 + X + Z + X*Z + (1 + X | gid), ...) # Z explaining random intercept and random slope
```

Example:

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: gpa ~ 1 + ses + clubs + ses*clubs + (1 + ses | schoolnr)
   Data: dat

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 schoolnr (Intercept) 0.173597 0.41665
          ses         0.007341 0.08568  -0.79
 Residual             0.696968 0.83485
Number of obs: 2287, groups:  schoolnr, 131
```

The varying slopes visualized

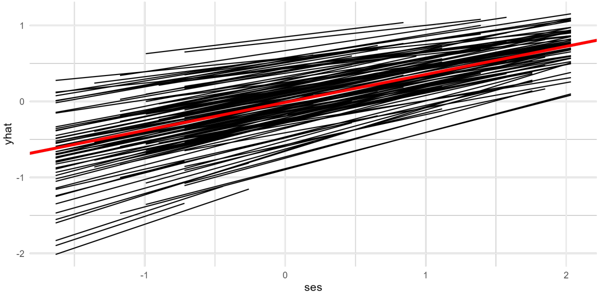

Figure 10: The variance of the intercepts is estimated to be 0.174. The variance of the slopes is estimated to be 0.007. The covariance between intercepts and slopes is estimated to be -.0284. That is, the slope is steeper for groups with lower intercepts and vice versa.

Comparison of model assumptions

- OLS and multilevel regression have the same type of assumptions:
  1. Functional form (linear predictor) is appropriate
  2. Independence of errors (= independence of observations given the linear predictor)*
  3. Constant variance of errors (homoscedasticity)*
  4. Normality of errors

  → MLM relaxes assumptions 2 + 3

  → MLM extends assumptions 4 to two "error" terms
- *OLS regression:* $e_i \sim N(0, \sigma_e^2)$
- *Varying intercept model:*
  $e_i \sim N(0, \sigma_e^2), u_{0j} \sim N(0, \sigma_u^2), Cov(e_i, u_{0j[i]}) = 0$
- *Varying intercept + slope model:*
  $e_i \sim N(0, \sigma_e^2), \left[u_{0j}, u_{1j}\right] \sim N(0, \Sigma)$ with $\Sigma = \begin{bmatrix} \sigma_{00}^2 & \\ \sigma_{10}^2 & \sigma_{11}^2 \end{bmatrix}$,
  $Cov(e_i, \mathbf{u}_{j[i]}) = 0$

MLM relaxes assumptions 2 + 3

- Covariance matrix of 4 students nested in 2 schools (students 1-2 in school 1 and students 3-4 in school 2) for a variance-component model:

$$\Sigma_{OLS} = \begin{bmatrix} \sigma_e^2 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 \\ 0 & 0 & 0 & \sigma_e^2 \end{bmatrix}, \Sigma_{MLM} = \begin{bmatrix} \sigma_u^2 + \sigma_e^2 & \sigma_u^2 & 0 & 0 \\ \sigma_u^2 & \sigma_u^2 + \sigma_e^2 & 0 & 0 \\ 0 & 0 & \sigma_u^2 + \sigma_e^2 & \sigma_u^2 \\ 0 & 0 & \sigma_u^2 & \sigma_u^2 + \sigma_e^2 \end{bmatrix}$$

  $\rightarrow$ MLM allows for covariance of students within the same school (e.g., student 1+2): $Cov(u_1 + e_1, u_1 + e_2) = cov(u_1, u_1) = \sigma_u^2$.

- The varying slope model relaxes the *homoscedasticity assumption* by allowing the "error" variance to depend on X:
  $y_i = (\gamma_{00} + \gamma_{10} X_i) + (u_{0j[i]} + u_{1j[i]} X_i + e_i)$
  $\rightarrow var(e_i) = \sigma_e^2$
  $\rightarrow var(u_{0j[i]} + u_{1j[i]} X_i) = \sigma_{00}^2 + 2\sigma_{u10} X_i + \sigma_{11}^2 X_i^2$

Modeled heteroscedasticity

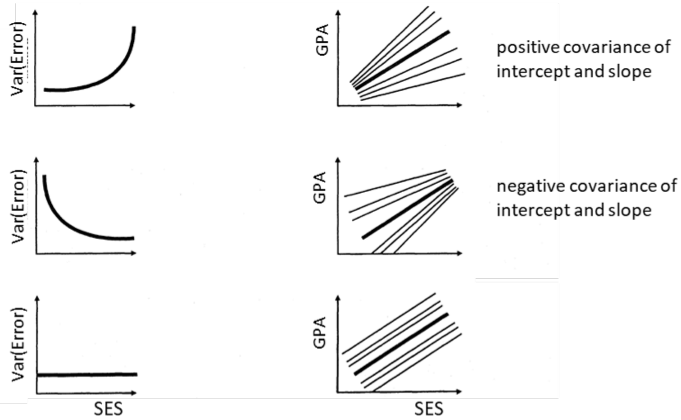

Figure 11: Different types of heteroscedasticity lead to different varying intercept and varying slope estimates. Figure adapted from Bullen, Jones & Duncan (1997).

Clustering as an interesting phenomenon

Clustering as an interesting phenomenon

1. *Learning about variability within and between groups*
2. *Learning about effect heterogeneity*
3. *Learning whether the within-group effect and the between-group effect of a predictor differ*
4. *Improving group-level inference and prediction*

Learning about variability within and between groups

- In my own work, I analyze the survival of coalition governments in Europe and measure the proportion of variance *within and between countries*.
- I then examine how much of this variance at each level can be explained by country differences in the funding structure of parties

Table: Variance estimates at each level

Level	M1: variance component model	M1: % of total variance	M2: incl. party funding variable
Country (σ_u^2)	0.66	33	0.54
Government (σ_e^2)	1.13	67	1.13

Figure 12: Simplified example. For more information: Rosche (2020): A multilevel model for coalition governments: Uncovering dependencies within and across governments due to parties.

Clustering as an interesting phenomenon

1. *Learning about variability within and between groups*
2. *Learning about effect heterogeneity*
3. *Learning whether the within-group effect and the between-group effect of a predictor differ*
4. *Improving group-level inference and prediction*

Learning about effect heterogeneity

- Predictor effects may vary by group, which is difficult to analyze with OLS regression when the number of groups are large and the number of observations per group are small.
- With multilevel modeling, we can specify *varying slopes* to allow predictor effects to vary by group. Moreover, by adding cross-level interactions, this variation can be explained.

Clustering as an interesting phenomenon

1. *Learning about variability within and between groups*
2. *Learning about effect heterogeneity*
3. *Learning whether the within-group effect and the between-group effect of a predictor differ*
4. *Improving group-level inference and prediction*

Within- and between-group predictor effects

- Consider a situation where the within-group effect of a predictor differs from its between-group effect:

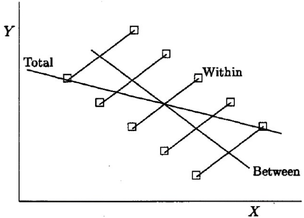

Figure 13: The within-effect of X (β^W) differs from the between-effect of X(β^B). (Snijders & Bosker 1999: 28)

- Any model simply including X: $y_i = \beta_0 + \beta_1^* X + e_i$ will estimate a weighted average of within- and between-group effect: $\beta_1^* = \phi\beta_1^W + (1-\phi)\beta_1^B$.
- The weighting ϕ will depend on the proportion of variance within and between groups and the ensuing precision of β^W and β^B.

Within- and between-group predictor effects

- Any pooled model will estimate the weighted average:
  - Pooled OLS model: $y_i = \beta_0 + \beta_1^* X_i + e_i$
  - Pooled ML model: $y_i = \gamma_{00} + \beta_1^* X_i + u_{0j[i]} + e_i$
    $\rightarrow$ If we know that $\beta^* = \beta^W = \beta^B$ or we are interested in the pooled effect β^*, the ML estimator β^*_{ML} varies less across samples and is thus more efficient than β^*_{OLS}.
- The within-group model (*"FE model"*) is a different estimator:
  $(y_i - \bar{y}_{j[i]}) = \beta_1^W (X_i - \bar{X}_j[i]) + (e_i - \bar{e}_{j[i]})$
- IMO a better solution: the *within-between ML model*
  $y_i = \beta_{0j[i]} + \beta_1^W (X_i - \bar{X}_{j[i]}) + \beta_1^B \bar{X}_{j[i]} + u_{0j[i]} + e_i$
  $\rightarrow$ Estimates the same within-group effect as the FE model
  $\rightarrow$ Estimates the between-group effect
  $\rightarrow$ Keeps the variance at each level

Within- and between-group predictor effects

- Any pooled model will estimate the weighted average:
  - Pooled OLS model: $y_i = \beta_0 + \beta_1^* X_i + e_i$
  - Pooled ML model: $y_i = \gamma_{00} + \beta_1^* X_i + u_{0j[i]} + e_i$

    $\rightarrow$ If we know that $\beta^* = \beta^W = \beta^B$ or we are interested in the pooled effect β^*, the ML estimator β^*_{ML} varies less across samples and is thus more efficient than β^*_{OLS}.
- The within-group model (*"FE model"*) is a different estimator:

  $(y_i - \bar{y}_{j[i]}) = \beta_1^W (X_i - \bar{X}_j[i]) + (e_i - \bar{e}_{j[i]})$
- IMO a better solution: the *within-between ML model*

  $y_i = \beta_{0j[i]} + \beta_1^W (X_i - \bar{X}_{j[i]}) + \beta_1^B \bar{X}_{j[i]} + u_{0j[i]} + e_i$

  $\rightarrow$ Estimates the same within-group effect as the FE model

  $\rightarrow$ Estimates the between-group effect

  $\rightarrow$ Keeps the variance at each level

Clustering as an interesting phenomenon

1. *Learning about variability within and between groups*
2. *Learning about effect heterogeneity*
3. *Learning whether the within-group effect and the between-group effect of a predictor differ*
4. *Improving group-level inference and prediction*

Improving group-level inference and prediction

- Varying intercept (and slope) estimates are especially relevant when researchers are interested in predicting $\hat{y}$

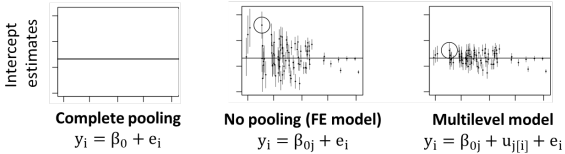

Figure 14: Adapted from Gelman & Hill (2002: 253)

- Compared to a model in which only 1 intercept is estimated ("*complete pooling*") and a model in which J intercepts are directly estimated ("*no pooling*"), the MLM models β_{0j} and estimates their mean and variance: $\beta_{0j} \sim N(\mu, \sigma_u^2)$
- While *no pooling* overstates the group-level variation (overfits) and *complete pooling* ignores it (underfits), the MLM estimates a weighted average of group-specific and overall intercept.

Shrinkage estimation

- For an intercept-only model: $\hat{\beta}_{0j} \propto \frac{n_j}{\sigma_e^2}\bar{y}_j + \frac{1}{\sigma_u^2}\bar{y}$
- The MLM "borrows strength" from groups with more information to improve the prediction of groups with less information. Predictions are therefore often more accurate. This feature is called *shrinkage estimation*
- As the MLM takes into account uncertainty at each level, predictive intervals are also often more accurate (for in-sample and out-of-sample prediction).

Take home message

- To use multilevel modeling, the number of groups should be larger than ≈ 10. With less, there likely is not enough information to reliably estimate the variance between groups. In that case, OLS regression with group-level indicators (*"fixed effects"*) should be employed. MLM, however, can be used with very small numbers of observations within (some) groups.
- For panel data, the within-between ML model is a good choice
- MLM is a powerful tool that is able to integrate many different statistical models:
  - Skrondal & Rabe-Hesketh (2002): Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.
  - Hodges (2013): Richly Parameterized Linear Models. Additive, Time Series, and Spatial Models Using Random Effects.