Unit 3:
Multilevel Panel and Growth Models
II. Estimation and Interpretation of Mixed Models

PS2701-2019

Longitudinal Analysis

Week 10

Professor Steven Finkel

The "Mixed" or Hierarchical Growth Model

- Level 1: Intra-Individual Change Over Time

$$(1) \quad Y_{it} = \beta_{0i} + \beta_{1i} Time_i + \varepsilon_{it}$$

- Level 2: Inter-Individual Differences in Level 1 Parameters

$$(2a) \quad \beta_{0i} = \beta_{00} + \beta_{01} X_{1i} + \beta_{02} X_{2i} + \zeta_{0i}$$

$$(2b) \quad \beta_{1i} = \beta_{10} + \beta_{11} X_{1i} + \beta_{12} X_{2i} + \zeta_{1i}$$

- Mixed Formulation:

$$(3) \quad Y_{it} = \beta_{00} + \beta_{01} X_{1i} + \beta_{02} X_{2i} + \beta_{10} Time_i + \beta_{11} X_{1i} Time_i + \beta_{12} X_{2i} Time_i + (\zeta_{0i} + \zeta_{1i} * Time_i + \varepsilon_{it})$$

- with three random effects and their variances:
  - σ_0^2: the variance of ζ_{0i}, the random (unexplained) portion of the unit-level intercept
  - σ_1^2: the variance of ζ_{1i}, the random (unexplained) portion of the unit-level slope for time
  - σ_ε^2: the variance of the idiosyncratic error term ε_{it} for a given unit at a given time

Fixed and Random Effects: Youth Delinquency Example

$$Y_{it} = \beta_{00} + \beta_{01}X_{1i} + \beta_{02}X_{2i} + \beta_{10} * Time_i + \beta_{11}X_{1i} * Time_i + \beta_{12}X_{2i} * Time_i + \ (\zeta_{0i} + \zeta_{1i} * Time_i + \varepsilon_{it})$$

- Level 2 variables: (X_1) Divorced Parents (yes/no), and (X_2) Number of Moves pre-Age 13.
- **FIXED EFFECTS IN THE MODEL**
  - β_{00}: Overall or Average Delinquency Intercept for all Units when X_1 and X_2 both equal 0
  - $\beta_{01:}$ Difference in the Unit-Level Delinquency Intercept for Children of Divorced Parents (i.e. when X_1=1)
  - β_{02}: Effect on the Unit-Level Delinquency Intercept for each Additional Move Made by Child's Family Before Age 13
  - β_{10}: Overall or Average Slope Effect for Time for All Units when X_1 and X_2 both equal 0
  - $\beta_{11:}$ Difference in the Unit-Level Slope for Time for Children of Divorced Parents (i.e. when X_1=1). If this coefficient is positive, it means that children of divorced parents *increase* in delinquent behavior over time more so than children whose parents are not divorced.
  - β_{12}: Effect on the Unit-Level Slope for Time for each Additional Move Made by Child's Family Before Age 13
- **RANDOM EFFECTS IN THE MODEL**
  - ζ_{0i}: the Random Portion of the Unit-Level Intercept (with variance σ_0^2)
  - ζ_{1i}: the Random Portion of the Unit-Level Slope for Time (with variance σ_1^2)
  - ε_{it}: the idiosyncratic error term for a given unit at a given time (with variance σ_ε^2)

Predictions in the Growth Model

- "**Fixed Predicted"** growth trajectory for a given individual is based on the fixed portion of the model. All individuals with identical values on the Xs will have the same "fixed predicted" intercept and the same "fixed predicted" slope, and hence the same "fixed predicted" growth trajectory. This can be viewed as the "average" growth trajectory for all individuals with identical values on the Xs.
- "**Full Predicted"** growth trajectory of Y will deviate from the fixed predicted value, depending on the size of the random components ζ_{0i} and $\zeta_{1i.}$ If ζ_{0i} is large, the unit's intercept will be bigger than that predicted by the Xs, and if ζ_{1i} is large, the unit's predicted slope will also be bigger than predicted by the Xs. So every unit has some additional random effect that affects the magnitude of both its intercept and slope in the growth curve.
- "**Actual"** value of Y at any given time will equal the predicted Y based on the **full predicted** growth trajectory plus the idiosyncratic unit-time error term $\varepsilon_{it.}$

Maximum Likelihood Estimation of Mixed Models

- For our "simple" mixed model, we need to estimate:
  - 6 fixed parameters (β_{00}, β_{01}, β_{02}, β_{10}, β_{11}, β_{12}); these are (usually) the primary theoretical parameters of interest.
  - But we must do so in the context of a composite error term that has three separate random components, and that is by construction **heteroskedastic** (because ζ_{1i} *Time varies over time) and **autocorrelated** (because of the presence of ζ_{0i} and ζ_{1i} in the error term for each unit at every point in time). This makes estimation difficult and sometimes very slow, depending on the number of random effects included in the models.
  - Moral: Choose the random effects wisely (and judiciously)!
- ML methods estimate the values of these population parameters that *maximize the likelihood* of observing the sample data that we, in fact, did observe.
  - In SEM we maximized the likelihood of observing the sample variances and covariances that we did observe, given the implied V-C matrix of our SEM model; here we will maximize the likelihood of observing the value of the individual Y_{it} in our sample

- This proceeds by writing out the probability of observing the outcome Y_{it} for a particular case that we did observe, given that the data were generated from combining the parameters in the model (the β and the σ^2) with the X independent variables.
  - The probability of observing the outcome Y for a given case is a function of its distance from the mean (like a "z-score") in a normal distribution; hence the importance of the normality assumption for the variables and random effects
- We then aggregate the individual probabilities (through multiplication) to arrive at a *joint likelihood function* of observing the given sample, given the estimated parameters.
- Iterative procedures find the estimated parameters which yield the highest joint likelihood of having observed the set of outcomes that comprise the sample data. These parameters yield predictions of Y_{it} that are as close as possible to the actual observed Y_{it}.
  - In practice, the process maximizes the *log* of the joint likelihood to make the mathematics more tractable, which is why summary statistics for ML estimation are given as log-likelihoods.

- Complications in mixed models stem from the multiple random effects, such that the probability of observing Y for a given case is a function of Xs, the β, and its place on *three separate normal distributions*, those of the ζ_{0i}, ζ_{1i}, and ε_{it}. (In "regular" regression we worried only about X, β and ε_{it}).
- So we need to consider in the estimation process the variances and covariances of the random effects, or the **"variance components"** of the model. In this case we have:
  - the ε_{it}, which we assume to be normal, homoskedastic with variance σ_ε^2, and having no correlation with previous values over time, and:
  - the variances for the two Level 2 random effects and their covariance (σ_{01}). We assume that the variances are also normally distributed and independent of the σ_ε^2 errors, though the covariance between the two Level 2 random effects may be nonzero.

- Formally, we assume the following for the Level 2 random effects:

(4) $$\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right)$$

- In combination with the assumptions for ε_{it}, this produces in the mixed growth model, or what Singer and Willett call the "standard multilevel model for change", a complex variance-covariance matrix of the compound error term

$$(\zeta_{0i} + \zeta_{1i} * Time_i + \varepsilon_{it})$$ in (3) above. For a three-wave example:

(5) $$\Sigma = \begin{bmatrix} \sigma_0^2 + \sigma_1^2(1)^2 + 2\sigma_0(1) + \sigma_\varepsilon^2 & & \\ \dfrac{\sigma_0^2 + \sigma_{01}(1+2) + \sigma_1^2(1*2)^2}{\sqrt{(\sigma_{c1}^2\sigma_{c2}^2}} & \sigma_0^2 + \sigma_1^2(2)^2 + 2\sigma_0(2) + \sigma_\varepsilon^2 & \\ \dfrac{\sigma_0^2 + \sigma_{01}(1+3) + \sigma_1^2(1*3)^2}{\sqrt{(\sigma_{c1}^2\sigma_{c3}^2}} & \dfrac{\sigma_0^2 + \sigma_{01}(2+3) + \sigma_1^2(2*3)^2}{\sqrt{(\sigma_{c2}^2\sigma_{c3}^2}} & \sigma_0^2 + \sigma_1^2(3)^2 + 2\sigma_0(3) + \sigma_\varepsilon^2 \end{bmatrix}$$

- where the terms in brackets signify the time period in question (wave 1, 2, 3), and the terms under the square root signs signify the composite error variance for a given point in time. So in element (2,1), the terms under the square root sign signify the wave 1 composite variance, i.e, the diagonal element (1,1), and the wave 2 composite variance, i.e. the diagonal element (2,2).
- **NOTE: You can see the intrinsic heteroskedasticity and autocorrelation built into the model (right?)**
- So what we want is for ML to estimate those values of the β fixed effects, and those values of the four variance components (σ_ε^2, σ_0^2, σ_1^2, σ_{01}) that produce the highest overall likelihood of observing the data that we did observe, given the following distribution for Y that is assumed in our model:

$$(6) \quad Y \sim N(\beta_{00} + \beta_{01}X_{1i} + \beta_{02}X_{2i} + \beta_{10}Time_i + \beta_{11}X_{1i}Time_i + \beta_{12}X_{2i}Time_i, \Sigma)$$

- ML will:
  - estimate the β fixed effects that, in combination with the σ random effects, generate Y_{it} that come as close as possible to the actual unit-time values of Y_{it}. This is what it means to "maximize the likelihood of observing the data." By maximizing the likelihood in this fashion, the estimation procedures also minimizes the variance of the Level 1 residuals ε_{it}.
  - estimate "fixed predicted growth trajectories" from the β fixed effects that come as close as possible to the average growth trajectories of all cases with the same values of X (so that the overall deviations of the actual trajectories from the "fixed predicted" regression lines will be as small as possible).
  - estimate Level 2 variance components that describe the amount of variation of the units around the fixed predicted growth trajectories in terms of the intercepts and the slopes, and the amount of idiosyncratic variation around the "actual predicted trajectories," given the normality assumptions and the form of the Σ variance-covariance matrix that is specified in the model.

Full (FML) versus Restricted (REML) Maximum Likelihood Estimation

- One important technical issue: should one use FML or REML to estimate the model parameters?
- You can think of FML as estimating *all* of the fixed and random effects simultaneously in an effort to maximize the overall likelihood of observing the sample. But actually, we need the fixed effects first to solve for the random effects, and there is some uncertainty about the *estimated* fixed effects that should (but is not) taken into account by FML. This is somewhat like the issues involved in "Feasible" GLS, when we used an estimate from our sample of some parameter (like the autocorrelation parameter ϱ) instead of the "real" population value. So the FML estimates of the random effects should be adjusted somewhat because they are based on estimates of the fixed effects ***as if*** we knew their true value. FML does not do this, and so it underestimates to some degree the size of the variance components σ.

- Restricted Maximum Likelihood (REML) corrects for this problem by eliminating the fixed effects from the likelihood function altogether. What it does is use some other consistent procedure to estimate the fixed effects (such as Generalized Least Squares or even OLS), and then calculates a residual for each unit at each point in time. It then maximizes the likelihood of observing those *residuals* across the entire sample, given the assumptions about the distribution of the model's composite residuals in (5). This is a more conservative estimation procedure for the random effects variance components.
- The downside of REML is that ***one cannot compare models with tests based on the (log)-likelihood function (like χ^2 difference tests) unless the fixed components are identical.*** This follows from the descriptions of each method above, as REML estimates are based on the residuals from a model with *specific* fixed effects – these residuals would obviously be different if different fixed effects were included. So if we wish to test the fit of alternative models that contain *different* fixed components, we need to estimate the alternative models with Full Maximum Likelihood Methods, recognizing the potential deficiencies of this method for the estimation of the random effects. You can change the estimation procedure used in Stata by modifying the command syntax. Full ML is Stata's current default estimation option.

Interpreting the Estimated Fixed Effects in the Mixed Model

- We interpret the β fixed effects exactly as we do in "regular" regression, i.e. the effect of one unit change in the Xs on change in the outcome variable. But remember that the outcome variable can be either viewed as the Level 1 intercept and slope parameters β_{0i} or β_{1i}, **OR**, in the mixed formulation, as the ultimate dependent variable which is Y_{it}. The coefficient is exactly the same, but of course you need to be cognizant of which way you are interpreting the effect at any given time.
- So assume that we have the mixed model from our Stata example where Y is a country's POLITY democracy score, X_1 is Years of Prior Democracy (1972-1989), and X_2 is a country's UN Human Development Score in 1990. We estimate an effect for β_{01} of .3, and an estimate of β_{11} of -.5.

- In the Level 1/Level 2 way of looking at the mixed model, this means:
  - Each additional year between 1972 and 1989 that a country was democratic is associated with an increase of .3 on the country's growth trajectory *intercept*, controlling for the UNDP Human Development index (X_2).
  - Each additional year between 1972 and 1989 that a country was democratic is associated with an increase of -.5 in the country's growth trajectory *slope*, controlling for the UNDP Human Development index (X_2).
- But if you want to interpret the coefficients in the "mixed" formulation, it gets a bit trickier because the mixed model is a fully multiplicative model (with TIME as the interactive variable). That means the effects of one variable are dependent on the levels of the other variables that make up the multiplicative components. It also means that you need to know the coefficients associated with all of the components in order to make the proper interpretations of the effects.

- For example, if we want to describe the effect of YEARS OF PRIOR DEMOCRACY (X_{1i}) on Y_{it}, we can rearrange the mixed formulation to isolate the X_{1i} effect as:

(7) $$Y_{it} = (\beta_{01} + \beta_{11} Time_i) X_{1i} + \beta_{10} Time_i + \beta_{00} + \beta_{02} X_{2i} + \beta_{12} X_{2i} Time_i + (\zeta_{0i} + \zeta_{1i} * Time_i + \varepsilon_{it})$$

- So the effect of X_{1i} on Y_{it} depends on the time period you are talking about. When TIME=0, the effect of X_{1i} on Y_{it} is equal to β_{01}, which equals .3, which is the effect of X_{1i} on the intercept as just noted. When TIME=2, however, the effect of a unit-change in X_{1i} on Y_{it} is equal to (.3+ 2*-.5)=-.7; when TIME=5, the effect is -2.2, at Time=10, the effect is -4.7, and so on. So if you want to interpret an X variable's effect directly on the Y outcome, as opposed to its effect on the slope of TIME, you need to take into account the intercept effect *and* the slope effect to make the overall interpretation. And of course all of this takes place "holding the other variables and their effects on the growth trajectories constant."

- It might be important in some applications to calculate *conditional standard errors* at different levels of X_{1i}; if so, see the formula in Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14 (1):63-82.

- This also means that the effects in mixed models will depend on where the "Zero Point" of a variable is. For example, assume that the overall intercept in the model (β_{00}) is .75. This is actually the intercept when all other variables at Level 2 are 0. At TIME=5, we would add the effect associated with β_{10} (for example, 2) and multiply it by 5 to produce the overall predicted value of Y at TIME=5 of 10.75, with all other variables being 0. This means that if we had centered time around 1995, such that 1995=0, 1996=1, 1997=2, 1994=-1, etc., we would have estimated β_{00} to be 10.75.

- ***So the estimated effects depend on where "0" is on TIME and on all other relevant variables.*** This may lead you to make inaccurate conclusions if you are not cognizant of the implications of the zero point on your variables, as well as the fully interactive nature of the mixed model.
- Other variables could be centered as well to facilitate the interpretation of the effects. For example, for the UNDP Human Capital measure, the impact of TIME (by itself) would be interpreted as when UNDP=0. This may be unrealistic since there are no countries where UNDP=0. If we centered the UNDP variable at its "grand mean," then the impact of TIME (by itself) would be interpreted as when "centered UNDP" was 0, in which case you could say when UNDP is at its "average" value. This may make interpretations easier in the mixed model and be useful for you and for the reader.

Interpreting the Estimated Random Effects and the "Explained Variance(s)" in the Mixed Model

- There are two general kinds of random effects: the variance terms σ^2, and the individual random effects ζ_i that are estimated for each unit, though STATA does not provide the individual effects in the output unless you ask for them.
- The interpretation of the variance terms provides a starting point for discussing "explained" and "unexplained" variation, and "R-squared" in the mixed model framework. All of the variance components make these interpretations somewhat more complicated than in "normal" regression.

Interpreting the Variance Components

- $\boldsymbol{\sigma_\varepsilon^2}$: tells us how much Level 1 error variance there is in the model, that is, how much (squared) difference there is on average between the predicted and actual Y_{it}. If you would divide this value by the overall variance in Y_{it}, you would arrive at a predicted l-R^2 for the Level 1 equation.
- $\boldsymbol{\sigma_0^2}$: tells us how much Level 2 error variation there is in equation 2a predicting the growth trajectory intercepts, that is, how much (squared) difference there is on average between the "fixed predicted" β_{0i} and the "actual predicted" β_{0i} that includes the unit effects ζ_{0i}. If this value is statistically significant, it means that there *is* random Level 2 variation in the estimated intercepts, controlling for the Xs; if the value is not significant, you cannot reject the hypothesis that all of the units with identical values of the Xs share a common intercept.

- $\boldsymbol{\sigma_1^2}$: tells us how much Level 2 error variation there is in equation 2b predicting the growth trajectory slopes, that is, how much (squared) difference there is on average between the "fixed predicted" β_{1i} and the "actual predicted" β_{1i} that includes the unit effects ζ_{1i}. If this value is statistically significant, it means that there *is* random Level 2 variation in the estimated slopes, controlling for the Xs; if the value is not significant, you cannot reject the hypothesis that all of the units with identical values of the Xs share a common slope.
- $\boldsymbol{\sigma_{01}}$: tells us how much covariation there is in the two Level 2 random effects. If the value is positive, it means that as the unit effect for the intercept (ζ_{0i}) gets bigger, the unit effect for the slope (ζ_{1i}) also gets bigger; when the value is negative, it means that as the unit effect for the intercept gets bigger, the unit effect for the slope gets smaller. It is like a correlation between the two ζ. Often in growth models you see a negative correlation between initial status and rate of change – this is how "regression to the mean" effects manifest themselves within this analytic framework.

- The first three variance components thus give you the unexplained variation in each of the three equations of the multilevel model: (1) for the Level 1 error variation, and (2a) and (2b) for the Level 2 error variation. So you naturally want to talk about the absolute level of "explained variation" in each of these equations as well.
- This makes *some* sense for the Level 1 residuals because there *is* an observed variance in Y_{it} to compare the value of σ_ε^2 with. So an "R^2" at Level 1 is relatively straightforward conceptually. But since we do not *observe* a true distribution for the β_i variables, it is not clear what the actual variances are that could be used compare σ_0^2 and σ_1^2 with to arrive at Level 2 "R-Squared" values.
- The upshot of this is that we do not really talk about "R-squared" in mixed models; rather, we talk about **"Pseudo-R-squared"** statistics that reflect intuitively some of the same ideas as normal R-squared but do not share their exact statistical (and nice algebraic) properties.

- NOTE: Another potential problem arises with R-squared in that the estimates of all the error variances depend on the estimates of the others, as they are produced simultaneously in the ML procedures. This means that the error variances are not based on algebraic calculations like in normal regression (i.e. Sums of Squared Errors divided by degrees of freedom), but rather they are estimated from the complex likelihood maximization that includes every other estimate in the model.
- So the estimated error variances do not necessarily add up in predictable ways, and in fact may change in unusual (i.e. larger, not smaller) ways when you add more Level 2 predictors. This can happen because the fixed portion of the model, and hence the starting point for the REML estimation of the residuals, changes each time you add more variables.
- This is another reason for caution in interpreting these measures in mixed models!!!

- What is done in the mixed modeling tradition is to compare the estimated Level 2 variance components to their counterparts in models that have *no independent variables at Level 2* at all; that is, to equations (2a) and (2b) with no Xs and only random ζ_{0i} and ζ_{1i}terms in their respective model.
- Intuitively, we then base our notion of R-squared at Level 2 on the idea of **"proportional reduction of error" (PRE)**, which we know is a perfectly valid way of looking at R-squared. We say "how much does the addition of explanatory variables X_1, X_2, etc. reduce the error variation in the Level 2 intercept (slope) equations?" If we reduce the error variance by a lot through adding the X variables, we say that these variables "explain" the variation in the Level 2 equations. If not, we still have much unexplained variation in those parameters.

- We can start with a model that has no IVs at Level 2 **and** no effect of TIME. This model reduces to:

(8a) Level 1: $Y_{it} = \beta_{0i} + \varepsilon_{it}$

(8b) Level 2: $\beta_{0i} = \beta_{00} + \zeta_{0i}$

(8c) Mixed: $Y_{it} = \beta_{00} + \zeta_{0i} + \varepsilon_{it}$

- This is the **"Unconditional Means Model" (UM)**. It says that Y_{it} is a function of an overall population mean, a random unit effect that represents the intercept difference for unit *i* from the population mean, and a random level 1 residual. This model will produce two variance components:
  - σ_ε^2: Level 1 error variance in the individual Y_{it} (from equation 8a)
  - σ_0^2: Level 2 error variance in the β_{0i} intercepts (from equation 8b)
- This is the first "baseline model" for Pseudo R-squared

- Forgetting about growth models for the moment, consider a new model that tries to explain the magnitude of the Level 1 intercept by adding two new variables at Level 2.

(9a) Level 1: $Y_{it} = \beta_{0i} + \varepsilon_{it}$

(9b) Level 2: $\beta_{0i} = \beta_{00} + \beta_{01}X_{1i} + \beta_{02}X_{2i} + \zeta_{0i}$

(9c) Mixed: $Y_{it} = \beta_{00} + \beta_{01}X_{1i} + \beta_{02}X_{2i} + \zeta_{0i} + \varepsilon_{it}$

- This model (call it **Model "A"**) will also produce estimates of
  - σ_ε^2: Level 1 error variance in the individual Y_{it} (from equation 9a)
  - σ_0^2: Level 2 variance in the error for the β_{0i} intercepts (ζ_0 from equation 9b)
- The estimate of σ_ε^2 should not change (too much) from equation (8a or c) to equation (9a or c) because nothing in the model has been added at Level 1.

- But we *can* compare the estimate of σ_0^2 from equation (8b) to equation (9b), and this (proportional) difference will tell you how much the addition of the two Xs reduced the error variance of the Level 2 intercepts, or, in other words, how much the two Xs "explain" the Level 2 intercept variation.

- Formula:

$$"Pseudo\ \text{Intercept}\ R-squared" = \frac{\sigma^2_{0_{UM}} - \sigma^2_{0_A}}{\sigma^2_{0_{UM}}}$$

- We can extend these ideas to the mixed growth model. First, we can ask: How much does the inclusion of TIME, and its associated random effect ζ_{1i} reduce the overall error variation in the Level 1 residuals?

(10a) Level 1: $Y_{it} = \beta_{0i} + \beta_{1i} Time_i + \varepsilon_{it}$

(10b) Level 2: $\beta_{0i} = \beta_{00} + \zeta_{0i}$

$\beta_{1i} = \beta_{10} + \zeta_{1i}$

(10c) Mixed: $Y_{it} = \beta_{00} + \beta_{10} Time_i + (\zeta_{0i} + \zeta_{1i} Time_i + \varepsilon_{it})$

- This is called the **"Unconditional Growth Model" (UG)**. It allows for a common intercept and common fixed effect for time, and then four random effects:
  - σ_ε^2: Level 1 error variance in the individual Y_{it} (from equation 10a)
  - σ_0^2: Level 2 variance in the error for the β_{0i} intercepts (ζ_0 from equation 10b)
  - σ_1^2: Level 2 variance in the error for theβ_{1i} slopes (ζ_1 from equation 10b)
  - σ_{01}: covariance between the ζ_{0i} and ζ_{1i} random effects

- We can calculate the proportional difference between Level 1 residuals from this equation (10a or c) with its value from the UM model in (8a or c), and this will tell you how much the addition of TIME reduced the Level 1 error variance --- in other words, how much variation TIME (and its associated random effect ζ_{1i}) "explain" at Level 1 compared to a model without these terms.
- Formula:

$$"Pseudo \text{ Level 1 } R-squared" = \frac{\sigma^2_{\varepsilon_{UM}} - \sigma^2_{\varepsilon_{UG}}}{\sigma^2_{\varepsilon_{UM}}}$$

- If we add additional Level 1 covariates to the Level 1 model, we can go through the same procedures to determine their relative explanatory power.

- Finally, we can compare any subsequent model that includes Level 2 explanatory variables in the growth framework to the UG model, and see how much reduction in variance in both the intercept and slope they may produce. So, if we call the model we have been working with – with two independent variables predicting the Level 2 intercepts and slopes – **Model "B"**, we can use the variance components from that estimation to arrive at two "Pseudo R-squared" values, one for the intercept and one for the slope, as:

$$"Pseudo\ \text{Intercept}\ R-squared" = \frac{\sigma^2_{0_{UG}} - \sigma^2_{0_B}}{\sigma^2_{0_{UG}}}$$

$$"Pseudo\ \text{Slope}\ R-squared" = \frac{\sigma^2_{1_{UG}} - \sigma^2_{1_B}}{\sigma^2_{1_{UG}}}$$

Assessing the Overall Fit of Mixed Models

- We can use the ML estimation procedures to derive summary statistics of the overall fit of mixed models. These are analogous to the statistics we used in SEM. We use these statistics in the model building process to arrive at the "best" model among the several that we might estimate in a given situation.
- The most common measure for mixed models is called the model **"Deviance"**, and it is analogous to the Model χ^2 from SEM.

  We begin with the final maximized (log)-likelihood function from the ML estimation procedure. This is the log of the maximized likelihood of observing the data in the entire sample that we did observe, given the estimated fixed and random effects. Call this value the LL_C, for "current" model . It is the largest value (i.e., least negative value) possible, given the assumptions of the model and the sample data.

- We can also conceive of the "best possible model," which would produce a predicted likelihood of 1 of observing this sample of Y_{it}, for a LL of 0 (as the log of 1 is 0). This is analogous to the "saturated" model of SEM that fully accounted for the observed variances and covariances between the variables in the model. In this case, it would be a model that perfectly predicts each unit's Y at each point in time. Call this value the LL_S, for "saturated."
- The "model χ^2" or "Deviance" is:

  $-2(LL_C - LL_S) = -2*LL_C$
- This value follows a χ^2 distribution, as we discussed earlier. More importantly, the difference in Model Deviances between *any two nested models* also follows a χ^2 distribution, so you can test whether relaxing constraints from one model to another results in a significant improvement in overall fit. The difference between the two Model Deviances will have degrees of freedom equal to the number of relaxed constraints.

- For example, we can compare a "constrained" model that has a common slope for TIME (i.e., where σ_1^2 is assumed to be 0) with an "unconstrained" one that has non-zero random unit-level variation in the slope for TIME. We know that we can arrive at the constrained model simply by imposing the constraint that $\sigma_1^2 = 0$, so that model is *nested* in the model with randomly varying slopes. This would be calculated as:

  $-2(LL_{Constrained} - LL_{Unconstrained})$, with 1 df
- Since Unconstrained LL will always be larger (less negative) or equal to the Constrained LL, this expression will be greater than or equal to zero.
- NOTE: The difference in Model Deviances between nested models with different *fixed effects* must be based on the LLs that you obtain through Full Maximum Likelihood Methods (FML). If you use "Reduced ML" methods (which are advantageous in point estimation), you can only test nested models that differ in their *random* effects, for reasons noted above. So it is often the case that you will estimate models using both FML and REML methods, the latter to provide the best point estimates for the fixed and random effects, the former to use in testing alternative models with different fixed effects.

Comparing Non-Nested Models

- If you have non-nested models, you cannot use the Model Deviance differences because these differences follow no known distribution. But you can use what we discussed in SEM, Information-Based Goodness of Fit Indices, to compare non-nested models that are estimated via ML methods.
- Note that these must be fit on the same data, so you cannot use them in completely different situations. But the models you compare with the same data need not be nested within one another.
- Two common measures (both with smaller values being "better"):
  - Akaike Information Criterion (AIC): Model Deviance plus a penalty equal to the number of estimated parameters
  - Bayesian Information Criterion (BIC): Model Deviance plus penalty equal to:

    2 times the number of estimated parameters *.5(ln(N))

Calculating Individual Random Effects.

- It is sometimes useful to estimate the unit effect, either the intercept or the slope or both, for individual cases in the analysis.
- We know, for example, that all individuals with the same values of the Xs will have the same predicted growth trajectory. But we also know that there will be a distribution around these growth trajectories, with some individuals having a larger or smaller intercept than predicted based on their ζ_{0i}, and some individuals having a larger or smaller slope than predicted based on their ζ_{1i}.
- We also know that the variance of the ζ_{0i} is estimated to be σ_0^2, and we know that the variance of the ζ_{1i} is estimated to be σ_1^2. But what about an *individual* value for ζ_{0i} and ζ_{1i}? Where can we estimate these from?

- One way: estimate the values of ζ_{0i} and ζ_{1i} for a given case by running an OLS regression predicting Y from TIME on that case only. (This is what STATA did with the 'STATSBY" routine in the beginning of the DO file we are working with in class).
- Do this for the entire sample, one individual at a time, or estimate all of the effects simultaneously with a series of dummy variables to estimate each individual intercept, and a series of the dummies multiplied by TIME to estimate the individual slopes. This would give us, in effect, the "fixed effects" version of the individual unit effects, to borrow the language of Unit 1.
- But these estimates are not totally trustworthy because they are not terribly precise, given the relatively small number of observations for each unit and the consequent likelihood that the estimates would be influenced by small amounts of truly random variation that would go unrecognized by the FE procedure. (We discussed this problem earlier in the class).

- So *all* random effects models – including the mixed growth models – attempt to improve on the LSDV/FE procedure through the idea of "borrowing strength", meaning that our estimates of each individual case's unit effects "borrow" from the overall population values, with the borrowing dependent on how reliable we estimate the original LSDV/FE estimates to be. We can also say that the RE estimates "shrink" the LSDV/FE estimates back to the overall population values, depending again on the amount of unreliability in the individual LSDV/FE values.
- In RE models, the predictions of the random effects for given cases are called "Empirical Bayes" Estimates, or "Model-Based" estimates.
- Idea: We take the FE estimate of a unit intercept or slope effect, and we "shrink" it back to the population average to the extent that the unit effects are not reliable. We "borrow" from the population to get a more precise estimate of the particular unit effects – this is why random effects estimators are more efficient than fixed effects (assuming no correlation between the Xs and the error terms).

- When are fixed effects (dummy variable) estimates of unit effects likely to be less reliable?
  - When there is little overall unit-level variation in the population – in that case the little bit of variation that exists in our sample might be random noise. The larger the "true" unit-level variation, the more we trust the unit effects that a dummy variable LSDVor FE estimation would produce;
  - When there is much Level 1 variation in the population – in that case the variation in unit effects is swamped by the variation of observations *around* the unit effect, and from sample to sample we might see much different estimated unit effects because of the large variance. So we again have less trust in a dummy variable LSDV/FE estimation of the unit effects; and
  - When there is small T – that means that the unit effects that look to be real might just be a couple of random blips – if they hold up with larger T we have more confidence.

- So in RE models, we take the LSDV/FE dummy effects and "shrink back" to zero as the conditions in a-c above are met. If there is little unit-level variation, much Level 1 variation, and small T, we shrink the estimates a lot to produce the "Empirical Bayes" RE estimates; if the opposite, then we say that the LSDV/FE dummy effects are reliable and we keep them
- The "Shrinkage Factor" for a given unit-level intercept effect with variance σ_0^2 is based on the formula:

$$(11) \quad \frac{\sigma_0^2}{\sigma_0^2 + \frac{\sigma_\varepsilon^2}{T}}$$

- As this value "shrinks" (as it were), the Empirical Bayes estimate will correspondingly shrink more towards 0. You can see that there will be more shrinkage when there is less predicted unit-level variance, as there is more Level-1 residual variance, and as there are fewer observations. **(NOTE: This is all very similar to what "Theta" was doing in the original RE exposition in unit 1!).**

- So our Empirical Bayes (EB) Prediction of the random effect is equal to:

(12) EB Estimate= FE Estimate*Shrinkage Factor

$$=\text{FE Estimate}^* \frac{\sigma_0^2}{\sigma_0^2 + \frac{\sigma_\varepsilon^2}{T}}$$

- These estimates are also sometimes called the "Best Linear Unbiased Predictions" from the model (**BLUP**).
- We can do the same thing for the RE associated with the Time Trend (ζ_{1i}) or any other random effect in the model. We start with a FE/LSDV effect, and shrink back to zero depending on the conditions listed above to get the EB, or BLUP estimate.

Further Issues in Mixed Models

- Non-linear growth models. In these models we do not assume linear growth across the population, but rather some kind of polynomial, quadratic, logarithmic, etc. Can add a squared terms for TIME or some other transformation of TIME to model the substantive process you think is operating. (See Chapter 6 of Singer and Willett).
- Time-Varying Covariates at Level 1. There is no reason to exclude other Level 1 predictors aside from TIME from the specification. This would mean that, controlling for the fixed predicted growth trajectory and the random effects for individual *i*, there is still an additional source of Level 1 variation based on the values of the time-varying Level 1 covariate. If you have these kinds of variables in your model, you can, in addition to examining statistical significance, see how they affect the reduction of Level 1 residuals in assessing their importance. (Example: USAID expenditures on democracy as a Level 1 time-varying covariate in Finkel *et al.* (2007)).

- You can extend the interaction models to include possible interactions between Level 1 time-varying covariates and TIME, and/or between Level 1 time-varying covariates and Level 2 predictors as well. In our study, e.g., we examine how the impact of AID differs under different Level 1 conditions (good economic performance versus bad, large US Military expenditures versus small), as well as under different Level 2 conditions (lots of Human Capital in the country versus little, lots of ethnic fractionalization versus little).
- You can also modify the assumptions regarding the error covariances to accommodate more complex patterns in the model's disturbances

- Recall that the variances-covariances of the composite error term for the "standard" mixed growth model was given as equation (5) above for the three-wave case as:

$$\Sigma = \begin{matrix} \sigma_0^2 + \sigma_1^2(1)^2 + 2\sigma_0(1) + \sigma_\varepsilon^2 & & \\ \dfrac{\sigma_0^2 + \sigma_{01}(1+2) + \sigma_1^2(1*2)^2}{\sqrt{(\sigma_{c1}^2\sigma_{c2}^2}} & \sigma_0^2 + \sigma_1^2(2)^2 + 2\sigma_0(2) + \sigma_\varepsilon^2 & \\ \dfrac{\sigma_0^2 + \sigma_{01}(1+3) + \sigma_1^2(1*3)^2}{\sqrt{(\sigma_{c1}^2\sigma_{c3}^2}} & \dfrac{\sigma_0^2 + \sigma_{01}(2+3) + \sigma_1^2(2*3)^2}{\sqrt{(\sigma_{c2}^2\sigma_{c3}^2}} & \sigma_0^2 + \sigma_1^2(3)^2 + 2\sigma_0(3) + \sigma_\varepsilon^2 \end{matrix}$$

- It may be, however, that the model does a poor job of reproducing the observed composite errors because the assumptions of the "standard" model do not hold in a given situation. For example, there may be autocorrelation in the idiosyncratic errors ε_{it}, or there may be time-wise heteroskedasticity in the idiosyncratic errors as well. There are many other possibilities as well (see Chapter 7 of Singer and Willett, or pp. 293-325 in Rabe-Hesketh/Skrondal).