Stat 414 – Day 15
Relaxing Exchangeability (15.3)

Last Time: Longitudinal data
- Repeat observations (level 1) of same "individuals" (level 2)
- Easily allows for missing values vs. only modeling complete cases
- May choose to make intercept correspond to first time point
- Time is a level 1 variable, other time-varying variables (level 1) and time invariant variances (level 2)
- Can allow the growth rate to vary across individuals (time | id) and/or nonlinear

Example 1: The U.S. National Longitudinal Study of Youth (NLSY) original study began in 1979 with a nationally-representative sample of nearly 13,000 young people aged 14 to 21. CurranData.txt contains data from a sub-study of children of the female NLSY respondents which began in 1986 when the children were aged between 6 and 8 years. Child assessments (e.g., reading scores) were then measured biennially in 1988, 1990 and 1992. The sub-sample of 221 children were assessed on all four occasions.

(a) Read in CurranData.txt and look at the first few rows. Which are time-varying variables and which are time-invariant? Is the dataset in long or wide format?

(b) Examine the correlations of the reading scores across the four time points. How would you summarize their behavior? Explain why the pattern in the covariation/ correlation makes sense in context.

(c) If we were to fit an OLS model to these data, what assumption would this model make about the correlations of the four observations on each person? Does that assumption seem reasonable here?

(d) Fit a random intercepts longitudinal model (= include Time). What assumptions does this model make about the correlations of the four observations on each person? Does that assumption seem reasonable here?

Multilevel models give us some different ways to model the correlation coefficients between observations changing over time.

We saw before that a **random slopes model** allows the covariance between observations to depend on the value of the explanatory variable, here time.

$$\begin{aligned}
&\text{var}(u_{0j} + u_{1j}t_{ij}) = \text{var}\,(u_{0j}) + 2t_{ij}\text{cov}\,(u_{0j}, u_{1j}) + t_{ij}^2\text{var}\,(u_{1j}) \\
&= \sigma_{u0}^2 + 2\sigma_{u01}t_{ij} + \sigma_{u1}^2 t_{ij}^2 \\
&\text{cov}\,(y_{ij}, y_{i'j}) = \text{cov}\,(u_{0j} + u_{1j}t_{ij} + e_{ij}, u_{0j} + u_{1j}t_{i'j} + e_{i'j}) \\
&= \text{var}\,(u_{0j}) + (t_{ij} + t_{i'j})\text{cov}\,(u_{0j}, u_{1j}) + t_{ij}t_{i'j}\text{var}\,(u_{1j}) \\
&= \sigma_{u0}^2 + (t_{ij} + t_{i'j})\sigma_{u01} + t_{ij}t_{i'j}\sigma_{u1}^2
\end{aligned}$$

(e) Fit a random slopes model and examine the covariance matrix. Verify the values. Does this model appear to adequately capture the shape of the individual trajectories?

A more flexible covariance structure can be fitted by including additional individual-specific random effects. Let's consider a quadratic term for time.

(f) Add a quadratic term for time and allow (to also have) random slopes. Compare this to the model without *time*^{2}, is the larger model (how much larger?) significantly better? How does the predicted correlation matrix compare?

So far we have assumed that the occasion-specific residuals are independent:

i. $\text{cov}(\varepsilon_{ij}, \varepsilon_{i'j}) = 0$ for any pair of occasions on the same individual
ii. $\text{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$ for any pair of observations for different individuals

Rather than adding *time*^{2} (and all the associated variances and covariances), we can instead model the correlation in the Level 1 residuals. A common choice is an AR(1) model for the level 1 residuals, which assumes the covariance matrix or the errors is of the form

$$\sigma_e^2 \begin{pmatrix} 1 & & & & \\ \rho & 1 & & & \\ \rho^2 & \rho & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{pmatrix}$$

(g) What does the model assume for Var(ε_{ij})?

(h) What does the model assume for Cov(ε_{ij}, $\varepsilon_{i'j}$)? Corr(ε_{ij}, $\varepsilon_{i'j}$)? How do these change the further apart the measurements in time?

(i) How many additional parameters are we now estimating?

Are more flexible structures, e.g., exponential but "in many applications, AR(1) provides an adequate model of the within subject correlation, providing more power without sacrificing Type I error control."

(j) Return to the basic two-level random intercepts model with *time* and *time*^{2} as Level 1 variables, but without random slopes. Fit the AR(1) error structure. Is this model significantly better than the "independent errors" random intercepts model?

(k) What is the estimated parameter of the AR(1) model ("autocorrelation"). How do you interpret it?

(l) What happens if you allow for the autocorrelation structure in the model with random slopes?

(m) How would we decide whether the pattern of growth (i.e., reading progress) differs between males and females? What would it mean for there to be an interaction between gender and time? Can you run this model?

From Finch and Bolin (2017):
Modeling longitudinal data in a multilevel framework has a number of advantages over more traditional methods of longitudinal analysis (e.g. ANOVA designs). For example, using a multilevel approach allows for the simultaneous modeling of both intraindividual change (how an individual changes over time), as well as interindividual change (differences in this temporal change across individuals). A particularly serious problem that afflicts many longitudinal studies is high attrition within the sample. Quite often, it is difficult for researchers to keep track of members of the sample over time, especially over a lengthy period of time. When using traditional techniques for longitudinal data analysis such as repeated measures ANOVA, only complete data cases can be analyzed. Thus, when there is a great deal of missing data, either a sophisticated missing data replacement method (e.g. multiple imputation) must be employed, or the researcher must work with a greatly reduced sample size. In contrast, multilevel models are able to use the available data from incomplete observations, thereby not reducing sample size as dramatically as do other approaches for modeling longitudinal data, nor requiring special missing data methods.

Repeated measures ANOVA is traditionally one of the most common methods for analysis of change. However, when it is used with longitudinal data, the assumptions upon which repeated measures rests may be too restrictive. In particular the assumption of sphericity (assuming equal variances of outcome variable differences) may be unreasonable given that variability can change considerably as time passes. On the other hand, analyzing longitudinal data from a multilevel modeling perspective does not require the assumption of sphericity. In addition, it also provides flexibility in model definition, thus allowing for information about the anticipated effects of time on error variability to be included in the model design. Finally, multilevel models can easily incorporate predictors from each of the data levels, thereby allowing for more complex data structures. In the context of longitudinal data, this means that it is possible to incorporate measurement occasion (level 1), individual (level 2), and cluster (level 3) characteristics. We saw an example of this type of analysis in Model5.3. On the other hand, in the context of repeated measures ANOVA or MANOVA, incorporating these various levels of the data would be much more difficult. Thus, the use of multilevel modeling in this context not only has the benefits listed above pertaining specifically to longitudinal analysis, but it brings the added capability of simultaneous analysis of multiple levels of influence.