Stat 414 - Day 16
Longitudinal data (Ch. 15)

Previously: Have data nested within groups. Want to include grouping variable in the model. Including it as random intercepts gives us a multilevel model, which has advantages including

- Allows separation of within group and between group variation
- Allows for inclusion of Level 1 and Level 2 variables
- Induces/Estimates within group correlation
- Including random slopes models heterogeneous responses (Level 2)
- Including cross-level interactions can explain variation in slopes (Level 2 equation)
- Does not require equal group sizes/handles missing values well

---

Multilevel models are especially helpful for "longitudinal data" (e.g., repeat observations on the same individual over time). Typically with longitudinal data we want to focus on changes over time and the effect of Level 2 variables. (We've actually already been looking at repeated measures data, but you will see some different terminology come up.)

Example: Data were collected by the Minnesota Department of Education for all Minnesota schools during the years 2008-2010 to compare charter and non-charter schools. School performance is measured by the mean score on the math portion of the Minnesota Comprehensive Assessment (MCA-II) data for the 6th grade students enrolled in 618 different Minnesota schools during the years 2008, 2009, and 2010. (MCA test scores for sixth graders are scaled to fall between 600 and 700, where scores above 650 for individual students indicate "meeting standards." Thus, schools with averages below 650 will often have increased incentive to improve their scores the following year.)
(a) Identify the Level 1 and Level 2 units.

First we want to explore how MCA math test scores relate to important Level 2 variables. This can be done using the data values for all three years or by averaging the data values for the three years into one number, or by using the 2010 values.
(b) What assumption is made by these last 2 approaches? Reasonable?

For the second approach open the "wide format" of the data (chart.wide.txt, this includes three columns for the three time points for each school) and use the SchoolAvg variable as the response. Examine the associations of these variable with several of the Level 2 variables.
(c) Which variable(s) seem(s) most useful in predicting the average math score?

Now open the "long format" of the data.
(d) Explain what year08 represents.

Create two visual representations of math scores vs. time for the first 20 schools:

- separate graphs for each school
- connecting lines or smoothers for each school overlaid on same graph (i.e., "spaghetti plot")

(e) Does it look like we will want to include random intercepts? (Meaning?) Does it look like we will want to include random slopes? (Meaning?)

(f) Produce a graph of the Math scores vs. year, separated by the charter (charter = 1) vs. public (charter = 0) schools. What do you learn?

Modeling Start with the null model.

(g) What is the ICC for these data? What does this tell you? Does this model adequately capture the behavior of our longitudinal data?

Key idea: The "exchangeability assumption" assumes the correlation between any two observations in the same cluster are the same. This is often not an appropriate assumption with longitudinal data (measures over time).

(h) How do we get variances and correlations to change over time?

Fit the "unconditional growth model" (time is only Level 1 variable, we haven't "conditioned" or "controlled" for any other possible covariates): multilevel model with year08, random intercepts, and slopes. (Be sure to use schoolnum, which are unique, not school name):

$y_{ij} =$; $where\ \epsilon_{ij} \sim$

```
Random effects:
 Groups    Name        Variance Std.Dev. Corr
 schoolnum (Intercept) 39.4410  6.2802
           year08       0.1105  0.3325   0.72
 Residual               8.8200  2.9699
Number of obs: 1733, groups:  schoolnum, 618

Fixed effects:
             Estimate Std. Error t value
(Intercept) 651.40766    0.27934 2331.96
year08        1.26495    0.08997   14.06
```

(g) Describe what this model is doing. What assumptions does this model make about the "occasion-specific" residuals? Does that seem like a reasonable assumption in this context? Interpret the variance components. What can you tell me about the populations of intercepts and slopes? How would you determine the percentage of within-school variation explained by the linear increase over time? How else can we evaluate the model?

Quiz 16 (due Wednesday 7am): What does this "random slopes" model assume for the variances and covariances of the *y* values?
Recall: $var(Y_{ij}) = \tau_0^2 + 2 \times time_{ij} \times \tau_{01} + time_{ij}^2\tau_1^2 + \sigma^2$
Recall: $cov(Y_{0j}, Y_{1j}) = \tau_0^2 + \tau_{01}(time_0 + time_1) + (time_0 \times time_1)\tau_1^2$

Adding Level 2 variable Include charter (charter schools = 1, public schools = 0) as a Level 2 variable (for both level equations, i.e., fixed effect and cross-level interaction with *year08*).

```
model2 = lmer(MathAvgScore ~ year08 + charter + charter:year08 + (year08 |
schoolnum), data = chart_long);summary(model2)
```

(h) Summarize the charter effect on the intercepts and the charter effect on the slopes. (Consistent with the graphs above?) Is either statistically significant? (Be very clear how you are deciding.) How much school-to-school variation in the intercepts has been explained by the charter school variable? What about the slopes?

Relaxing the linearity assumption The graphs of average scores over time indicated that there appeared to potentially be a nonlinear trend. There are many ways to relax the linearity assumption but first we will just consider a quadratic effect of time.

(i) *If we plan to use $time$ and $time^2$, do we need to center time first?*

(j) Write out the Level 1 and Level 2 equations that include $time$ and $time^2$, but only allow the intercepts to vary. What assumptions is this "unconditional quadratic growth model" imposing on the time trends?

Fit and interpret the model specified in (j):

```
summary(quadmodel <- lmer(MathAvgScore ~ 1 + year08 + I(year08^2) + (1 |
schoolnum), data=chart_long), corr=F)
```

(k) Is the quadratic effect statistically significant? How do you interpret the sign of the coefficient of this term?

(l) Can we fit the model that allows the slopes to vary?

Computer Problem 16 (due Wednesday 7am)

Compare the quadratic model to the model that is only linear in time, but with random slopes.

```
summary(linearmodel <- lmer(MathAvgScore ~ 1 + year08  + (1 + year08 | schoolnum), data=chart_long))
plot(allEffects(linearmodel), lines = T)
anova(quadmodel, linearmodel)
```

(a) Is it ok to do a likelihood ratio test here? How many parameters are estimated by each model? How do the AIC/BIC values compare? Which model do you recommend?

Another option is a *piecewise function.* With three time points this means we allow one slope from 2008 to 2009 and a different slope from 2009 to 2010. Create an indicator variable for 2009 and another for 2010. Include these two indicator variables (but not year08) in the model, with random intercepts (only).

```
head(chart_long$year08)
chart_long$ind2009 = as.numeric(chart_long$year08 == 1)
head(chart_long$ind2009)
chart_long$ind2010 = as.numeric(chart_long$year08 == 2)
head(chart_long$ind2010)
```

(b) Why does this work? How do you interpret the coefficient of ind2010?

Compare this model to the quadratic model –

```
piecemodel = lmer(MathAvgScore ~ ind2009 + ind2010 + (1 | schoolName), data = chart_long)
plot(allEffects(piecemodel))
```

(c) Does it describe a similar time trend? How so? How do the AIC/BIC values compare?

(d) Give a “modelling” reason to prefer the linear model to the quadratic or piecewise linear models.

Notes:

- Keep in mind the importance of the interpretability of your model, especially to non-statisticians.
- You can also consider functions that allow for “exponential growth”
- Also consider how well your model can extrapolate. It is definitely riskier to extrapolate with quadratic models.
- 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.