Fall, 2023

Thursday, Nov. 30

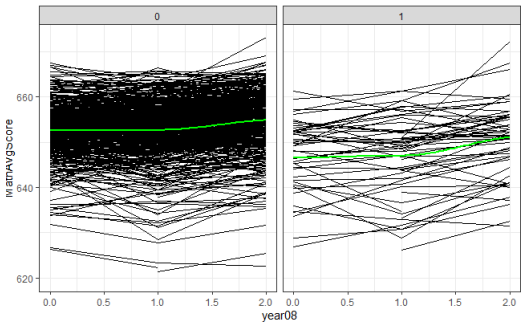
Stat 414 - Day 33
Longitudinal data - Relaxing assumptions (Ch. 15)

Last Time: Longitudinal Data

- Wide vs. Long format
- Time varying vs. time invariant variables
- Start the time variable at zero?
- Explore the raw data (graphs, correlation matrix)
- Unconditional growth model: $y_{ij} = \beta_{0j} + \beta_{1j} time_{ij} + \epsilon_{ij}$
  - Random slopes for time $V(Y_{ij}) = \tau_0^2 + 2\tau_{01}x_{ij} + \tau_1^2\sigma^2$
  - Assumes linearity in time
  - Assumes errors on the same individual are independent of each other
    $Cov(\epsilon_{ij},\ \epsilon_{kj}) = 0;\ Cov(Y_{ij}, Y_{kj}) = \tau_0^2 + \tau_{01}(x_{ij} + x_{kj}) + \tau_1^2(x + ijx_{kj}),$

Example, cont.: 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. Does the model match the data?

```
> cor(matrix, use="pairwise.complete.obs")
              MathAvgScore.0 MathAvgScore.1 MathAvgScore.2
MathAvgScore.0      1.0000000      0.8064146      0.7727215
MathAvgScore.1      0.8064146      1.0000000      0.8331408
MathAvgScore.2      0.7727215      0.8331408      1.0000000
```



(a) Fit the null model. 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).

(b) Fit the unconditional growth model (using lme). What has this changed? Is the implied correlation matrix from the model a better match to the observed correlation matrix?

So far we have assumed that the "occasion-specific" residuals (the ϵ's) are independent: $cov(\epsilon_{ij}, \epsilon_{kj})$ = 0 for any pair of occasions on the same individual.

A common alternative covariance structure is an AR(1) model for the Level 1 residuals, which assumes the covariance matrix of the errors is of the form

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

(c) What does the model assume for $Var(\epsilon_{ij})$?

(d) What does the model assume for $cov(\epsilon_{ij}, \epsilon_{kj})$? $corr(\epsilon_{ij}, \epsilon_{kj})$? How do these change the further apart the measurements in time?

(e) How many additional parameters does this add to our model?

So instead of random slopes for time, fit the AR(1) structure.

```
model2b = lme(MathAvgScore ~ year08 + I(year08^2), random = ~1 | schoolnum,
correlation=corAR1(), data = chart_long); summary(model2b)
```

(f) What is the estimated parameter of the AR(1) model ("autocorrelation"). How do you interpret it? Is it statistically significant? How are you deciding?

(g) Show how to find the correlation between year 1 and year 3 residuals based on the correlation between year 1 and year 2 residuals.

(h) Show how to find the values in the Level 1 *variance-correlation* matrix.

(i) Show how to find the "marginal" variance at time 0.

(j) Does the correlation matrix appear to be a better fit to the data?

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 here we will just consider a quadratic effect of time.

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

(l) Write out the Level 1 and Level 2 equations for a random intercepts model that includes $time$ and $time^2$. What assumptions is this "unconditional quadratic growth model" imposing on the time trends?

Fit and interpret the model specified in (l).

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

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

(n) Write out the Level 1 and Level 2 equations that allow the intercepts and both slopes to vary, but with no Level 2 covariates. Define your symbols. How many parameters does this add to the model?

(o) Can we fit the model suggested in (n)?

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

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)
```

(p) 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?

Optional: 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).

```
chart_long$ind2009 = as.numeric(chart_long$year08 == 1)
chart_long$ind2010 = as.numeric(chart_long$year08 == 2)
piecemodel = lmer(MathAvgScore ~ ind2009 + ind2010 + (1 | schoolName), data =
chart_long)
```

(q) Why does this work? How do you interpret the coefficient of ind2010? Compare this model to the quadratic model – does it describe a similar time trend? How so? How do the AIC/BIC values compare?

(r) 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.
- The AR structure does assume the observations are equally spaced in time (e.g., one year to the next/same distance apart) for all individuals. The AR model also assumes the variance is the same at the different time points, just allows for this consistent drop off in correlation as time points are further apart.
- There are more flexible structures, but "in many applications, AR(1) provides an adequate model of the within subject correlation, providing more power without sacrificing Type I error control."
- From Roback and Legler (2019): In the charter school example, as is often true in multilevel models, the choice of covariance matrix does not greatly affect estimates of fixed effects. The choice of covariance structure could potentially impact the standard errors of fixed effects, and thus the associated test statistics, but the impact appears minimal in this particular case study. In fact, the standard model typically works very well. So is it worth the time and effort to accurately model the covariance structure? If primary interest is in inference regarding fixed effects, and if the standard errors for the fixed effects appear robust to choice of covariance structure, then extensive time spent modeling the covariance structure is not advised. However, if researchers are interested in predicted random effects and estimated variance components in addition to estimated fixed effects, then choice of covariance structure can make a big difference. For instance, if researchers are interested in drawing conclusions about particular schools rather than charter schools in general, they may more carefully model the covariance structure in this study.

- 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.