Stat 414 – Day 14
Longitudinal Data (Ch. 15)

Last Time: Residual plots and Measures of influence
- Interpretations of residuals plots are the same as before
  - Now include random terms at each level (e.g., conditional residuals, random effects, e.g. u_{oj})
  - Can analyze random intercepts and random slopes separately or look at the "random effect residuals" that combine these together (e.g., $u_{oj} + u_{1j}z_j$)

We have actually already been analyzing *longitudinal data* = repeat observations on the same individuals (e.g., people, plants) over time. Typically we want to focus on changes over time and the effect of Level 2 covariates. You will often see some different terminology, e.g., the model with only time as a Level 1 covariate is the *unconditional growth model* (vs. random intercepts or *unconditional means* model):

$$Y_{ij} = \beta_{0j} + \beta_{1j}time + \epsilon_{ij},\ with\ \beta_{0j} = \beta_{00} + u_{0j}, \beta_{1j} = \beta_{10} + u_{1j}$$

σ^2 τ_0^2 τ_1^2

Interpretation of the variance components:

σ^2: unexplained variation about time trend
τ_0^2: variation initial values (time = 0)
τ_1^2: variation in growth rates

Example 1: Reconsider the Minnesota math test scores (charter and non-charter schools) from HW 6. One of the big assumptions we made was that the increase per year was linear.

(a) Reconsidering the graph of scores over time in a subset of schools, does there appear to be a consistent trend? How might you change the model?

allow nonlinear relationship? w/a larger increase in scores as time goes on

There are many ways to relax the linearity assumption (splines!) but here we will just consider a quadratic effect of time.

(b) If we plan to use *time* and *time*^{2}, do we need to center time first?

should help w/ multicollinearity
we have time = 0, 1, 2 so may not be worth it

(c) Write out the Level 1 and Level 2 equations that include *time* and *time*^{2}, allowing the slopes and intercepts to vary, but with no Level 2 covariates. Define your symbols.

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j}time + \beta_{2j}time^2 + \epsilon_{ij}$

Level 2: $\beta_{0j} = \beta_{00} + u_{0j}$ $\quad u_{0j} \sim N(0, \tau_0^2)$

$\beta_{1j} = \beta_{10} + u_{1j}$ $\quad u_{1j} \sim N(0, \tau_1^2)$

$\beta_{2j} = \beta_{20} + u_{2j}$ $\quad u_{2j} \sim N(0, \tau_2^2)$

parameters: $\tau_0^2, \tau_1^2, \tau_2^2, \sigma^2, \tau_{01}, \tau_{02}, \tau_{12}$

$\beta_{00}, \beta_{10}, \beta_{20}$

(d) Write out the Level 1 and Level 2 equations that include *time* and *time*^{2}, but only allow the intercepts to vary. What is an advantage of this model over the model suggested in (c)? What assumptions is this *unconditional quadratic growth* model imposing on the time trends?

$y_{ij} = \beta_{0j} + \beta_1 time + \beta_2 time^2 + \epsilon_{ij}$

$\beta_{0j} = \beta_{00} + u_{0j}$

(e) Fit and interpret the model specified in (d). Is the quadratic effect significant? How do you interpret the sign of the coefficient of this term?

$t = 7.1$ for quadratic so significant

effect of time is increasing w/ time

(f) Compare this model to the model that is only linear in time, but with random slopes. How many parameters are estimated by each model? How do the AIC/BIC values compare? Which model do you recommend?

AIC quad: 10302

AIC linear w/ random slopes 10348 ← also required an additional parameter

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

(h) Give a "modelling" reason to prefer the linear model to the quadratic or piecewise linear models

Simpler

Extrapolation