Stat 414 – Day 13
Measures of Influence (Ch. 10)

Last Time: Adding contextual variables to Level 2

- Variables describing the "context" of Level 2 groups, like group means
- Within group vs. Between group equations
- Adding Level 2 variables can explain variation in intercepts; Adding cross-level interactions can explain variation in slopes
  - Should let "theory" detect fixed vs. random slopes
  - Desire to include Level 2 variables, interactions?
- Can restrict parameter covariances to be zero to reduce number of parameters

Example: Consider data on 519 students in 23 schools from the NELS-88 (National Education Longitudinal Study, 1988) data (Kreft & De Leeuw, 1998). Note for race, 1 = Asian, 2 = Hispanic, 3 = Black, 4 = White, 5 = Native American and for schtype, 1=public, 2=catholic, 3=Private other, religious, 4=Private non-religious. Suppose we want to consider the relationship between the degree of structure that schools attempt to enforce in their classrooms and students' performance on a math test. (school23.txt)

(a) Fit a multilevel model with time spent on math homework each week (timeHWwk) and structure, with random intercepts at the school level. Is *classstructure* being considered a quantitative or a categorical variable? How can you tell?

(b) What do you conclude from the p-value for *classstructure* in your model?

(c) Is this conclusion consistent with a graph of math scores vs. *classstructure*?

(d) How does the previous graph compare to one that aggregates the data to the school level?

(e) What do you conclude?

There are several ways to measure the amount of influence of observation(s) to a model, e.g.,

- *DFBetas* are measured for each observation in the dataset and indicate how much the slope coefficient of a variable changes if that observation is removed (large if > $2/\sqrt{n}$.)
- *Cook's Distance* is measured for each observation and provides an overall measure of how much all of the model predictions change when the observation is removed (large if > $4/(n-2)$). Found by considering each observation's *leverage* and *residual*.

The influence.ME package appears to be a good one for detecting influence of Level 2 groups.
https://journal.r-project.org/archive/2012-2/RJournal_2012-2_Nieuwenhuis~et~al.pdf
The influence function looks at

- DFBetas (can specify which variables/parameters you want to focus on)
- Cook's distance (normally across all variables/parameters)
- Sigtest (you can supply cut-off value for significance and then see if judgement of significance changes)

If the group sizes are not too large, you can use obs=TRUE to also (separately) analyze the influence of Level 1 observations.

(f) Use the influence function to save the influence measures

(g) Create a graph of the dfbetas for HW and classstructure. Do any schools stand out as influential on either variable?

(h) Create a graph of the Cook's distances, sorted and marked by cut-off. Which schools stand out?

(i) Does removing any of the school change the significance of the *structure* variable?

(j) What about the *homework* variable?

(k) What would it mean to have an influential observation at Level 1?

(l) Are any students identified as influential for the *structure* variable? Where does 519 come from? Does removing either observation change the significance of the *structure* variable?

(m) How would you "deal with" school 7472?