Stat 414 - Day 8
Adding Level 1 and Level 2 Predictors (4.4, 4.5)

Last Time: Random intercept models $\hat{y}_{ij} = \beta_0 + u_j + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2), Cov(u_j, \epsilon_{ij}) = 0$

- ranef(model) displays the estimated group effects
  - ranef(model, condVar = TRUE) with lmer displays the standard errors of those effects
    $SE(\hat{u}_j) = \sqrt{1/\left(\frac{1}{\tau^2} + \frac{n_j}{\sigma^2}\right)}$ (measure of random sampling error in estimate)
  - assume $u_j \sim N(0, \tau^2)$; τ^2 = measure of unit to unit variation in population
  - assume β_{0j} follow a normal distribution with mean β_0 and variance τ^2
- Caterpillar plot
- An advantage of assuming the random effects follow a normal distribution is we can apply everything we know about normal distributions
- lme vs. lmer

Example 1: Netherlands language scores cont

Now we want to include pupil (verbal) IQ as a predictor of language test performance.

(a) Is this a Level 1 or Level 2 predictor?

(b) Write out an appropriate statistical "random intercepts" model (both composite and level equations) including verbal IQ. How many parameters are to be estimated?

Include the IQ variable (which has been centered (though before students with missing values were removed)) in the model.

```
model1 = lmer(langPOST ~ 1 + IQ_verb + (1|schoolnr), data = neth, REML=F)
```

(c) Provide interpretations of the estimated slope and intercept. (Hint: Remember lessons learned!)

(d) Is IQ-verb statistically significant? How are you deciding?

(e) What is the estimated variation in responses for a particular value of IQ_verb?

(f) What percentage of the Level 1 variance was explained by verbal IQ?

(g) What percentage of the Level 2 variance was explained by verbal IQ?

(h) Which has changed more, the estimated within-group variation or the estimated between-group variation? Does this make sense in context? Is it possible for both of them to decrease? What does that mean?

(i) What percentage of the total variance was explained by verbal IQ?

(j) What is the new value of the ICC? How do you interpret this? What would it mean for this value to be super close to zero?

(k) What would a graph of this model look like? What if we had treated the schools as fixed effects? What if school wasn't in the model?

Notes:

- We can think of ICC as the proportion of total variance explained by the grouping variable and R^2 as the proportion explained by the fixed effects.
- The difference between adjusted/unadjusted ICC is whether you take into account the "variance explained by the fixed effects" in the denominator.
  - The adjusted intraclass correlation coefficient is often smaller than the "raw" (null model) intraclass correlation coefficient.
  - Performance package: The adjusted ICC is what we would calculate "by hand" which just uses the variance components after adding the covariate into the model. The unadjusted ICC takes "fixed effect" variance into account (in the denominator) as well (see insight::get_variance(model)) ($var(X\beta)$= the change in unexplained variation when the fixed effect is added to the model). See handout in Canvas)
  - We will focus more on the adjusted ICC, if that. Of real interest to us is the unadjusted ICC from the null model, but you can look at the ICC in other models to see how that has impacted the "unexplained" group to group variation.
- The difference between conditional and marginal R^2 is whether you account for the random effects in the numerator (conditional - marginal = unadjusted ICC).
  - Marginal R^2 measures the variance explained by the fixed effects as a proportion of the sum of all the variance components ($\hat{\sigma}^2 + \hat{\tau}^2 + var(X\beta)$ ("The fixed effects explain …")
  - Conditional R^2 measures the variance explained by both the fixed and random effects in the model. ("The fixed and random effects explain …")
  - The unadjusted ICC is the difference between these! (the contribution of the random effects…)
- There is also some lack of agreement ("the literature does not seem to have converged on this topic") in how to calculate R^2 values for these models as the formulas provided here can actually turn out to be negative!

Reference: Nakagawa S, Johnson P, Schielzeth H (2017) The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisted and expanded. J. R. Soc. Interface 14. doi: 10.1098/rsif.2017.0213

Example 2: Radon (note: this data file may not match the one from HW)
Radon comes from underground and can enter more easily when a house is built into the ground (i.e., has a basement). In this dataset (for 919 homes across 85 counties in MN), *floor* indicates whether the measurement was taken in the basement or the first floor, and *basement* indicates whether the house had a basement.
(a) What is the estimated mean (log) radon ?

(b) What is the basemeas variable about?

(c) Do you predict higher or lower radon levels if the house has a basement and the measurement was taken in the basement? What if the house doesn't have a basement?

(d) Check your predictions.

(e) What do the estimated random effects $\hat{u}_j$ represent in this model?

(f) How does adding the basemeas to the model change the variance components?

(h) Now we want to add the soil uranium level to the model. Is this a level 1 or level 2 variable? How can you tell?

(i) Show how to add this variable to the model equations. How many terms will it add to the model?

(j) What do we have to do differently in the lmer command to tell R about this variable?

(k) Interpret the coefficient of uranium in this model. Do you consider it statistically significant?

(l) How did the variance components change? Do these changes make sense? Explain.

```
performance::r2(model1, by_group = TRUE)
```

Example 3: Recall the Netherlands study on language test performance for Grade 8 students nested within schools. Another important type of Level 2 variable is aggregating a Level 1 variable, e.g., average school IQ. These are sometimes referred to as "contextual effects."
(a) Is the school mean verbal IQ related to (average) performance score?

(b) Adding sch_iqv, what are the Level 1 and Level 2 model equations?

(c) Interpret the slope of the sch_iqv variable in this model in context.

(d) Based on the output you have, is this new variable a significant addition to the model? How are you deciding?

(e) How much Level 1 variability did we explain? How much Level 2?

(f) Which do we understand more, why some schools have higher language scores or why some students have higher language scores? Explain your reasoning.

What if we change to the "deviation" variable, verbal_IQ - sch_iqv? (aka Group Mean Centering)
(g) Interpret the slope coefficients for this model.

(h) Does the deviation variable explain variation at Level 1 and/or Level 2?

Notes

- It's probably a good idea to grand mean center all explanatory variables before you start your analysis.
- "Group mean centering" (as opposed to grand mean centering) creates a "within" group variable.
- Some recommend calculating group means before observations with missing values are deleted (or better yet, use imputation)
- When using x and x-bar (rather than deviation and x-bar), the coefficient of x-bar is the difference between the within and between group effect. The significance of the group mean variable is akin to the Hausman specification test in econometrics: Is the difference between the "within group" and "between group" effects statistically significant?