Stat 414 - Day 25
Model Diagnostics: Residuals (Ch. 10)

Last Time: Multiple random slopes

- Try to minimize use of random slopes or model gets very complicated very quickly
  - Can zero out covariances to simplify model but makes sense in context?
- Centering variables can sometimes help with convergence

Multilevel Model Assumptions: LINE for each level, independence of errors across levels

Example 1: Beach data Reconsider the random intercepts model for our beach data

```
model2 = lmer(Richness ~ 1 + NAP + (1 | rikzdata$Beach), data = rikzdata)
plot(model2)
head(ranef(model2)[[1]])
```

(a) What do you learn from the residuals plot? How do you interpret these residuals?

(b) What does this model predict for the Richness when NAP = 0.045 for the "average" beach? What does this model predict for this observations in Beach 1? What is the first observation in the first beach? How do we calculate the residual for this observation?

We will consider three different types of residuals!

- *Conditional residuals*: our usual level 1 residuals, the prediction errors within a particular level 2 group
  - These are what R returns with residuals(model)
  - Check for normality, equal variance
  - Can also plot residuals vs. other variables, use smoothers
- *Level 2 residuals*: our estimated random effects.
  - This is what R returns with ranef(model)
  - Check for normality but doesn't always guarantee real effects follow normal distribution, check for outliers
  - Useful to plot the Level 2 random effects vs. Level 2 units, other Level 2 variables
  - Can also plot squared Level 2 residuals against Level 2 variables to check for heteroscedasticity (similar to "Levene's Test")
  - "Random effect residuals" = response – fixed effects – conditional residuals
- *Marginal residuals*: prediction errors from overall model
  - In R: response - model.matrix(model) %*% fixef(model)
  - Accounts for (confounds) both random effects and random error
  - Check for unusual observations
  - Can be informative to plot these across the groups (probably differ)

(d) Verify the calculation of the conditional residual, the level 2 residual, and the marginal residual for the first observation in the first beach. Which is largest? Why?

Example 2: Consider a hypothetical dataset with 10 subjects with 4 temporal-based observations (one every year). Each person has data for age, sex, average number of cigarettes smoked each week, level of nicotine dependence from the Fagerstrom Test of Nicotine Dependence (FTND), ratings of depressive symptoms from the Beck Depression Inventory (BDI), and a count variable for the total number of lifetime major depressive episodes suffered up to that point of data collection.

(a) Fit a model predicting the number of cigarettes smoked each month use based on time and self-reported depression (BDI), including their interaction, and FTND score, with random intercepts for subjects (subjectID):

```
model1 = lmer(Cigs ~ Time + BDI + Time:BDI + FTND + (1 |subjectID))
```

(b) Predict the average number of cigarettes smoked each week for the first time point for everyone with BDI = 7 and FTND = 6.

(c) Calculate the conditional and marginal residuals for the first person in the dataset.

(d) Plot the conditional residuals vs. the predicted values. Look for equal variance across the fitted values, outliers. Also check for normality. What do you conclude?

```
Plot(residuals(model1) ~ fitted.values(model1)); qqnorm(residuals(model1))
residuals(model1)
which(residuals(model1) > 20)
```

(e) Examine a graph of the marginal residuals vs. the marginal fitted values. Do you see evidence of outliers?

```
margfits = model.matrix(model1) %*% fixef(model1)
margresids = cigstudy$Cigs - margfits
plot(margresids~ margfits)
which(margresids > 40)
```

(f) Investigate and discuss the nature of this outlier. Which subject does this outlier belong to? Compare the conditional and marginal residual for this observation. Which is larger? Why?

```
cigstudy
margresids[28,1]
residuals(model1)[28]
```