Fall, 2024 *Monday, Oct. 21*

Stat 414 - Day 9
Model Diagnostics (Ch. 10)

Last Time: Adding Level 1 and Level 2 variables to random intercept models

- Level 1 variables may explain association at Level 1 and/or Level 2 (if the distribution of the variable differs across the Level 2 units)
  - Can also increase unexplained variation at Level 2
  - Some can only explain Level 1 variation (e.g., income percentile, z-score in country)
- Level 2 variables only explain association at Level 2 (e.g., country uranium)
- Aggregating a Level 1 variable to Level 2 gives a nice "contextual" variable
  - Does coefficient of group mean variable represent the additional contribution or the group level effect?
  - $\hat{\beta}_1 x_{ij} + \hat{\beta}_2 \bar{x}_j$ ($\hat{\beta}_2$ is the additional contribution at Level 2)
    - If not significant, the Level 1 and Level 2 associations are "the same"
  - $\hat{\beta}_1 \left(x_{ij} - \bar{x}_j\right) + \hat{\beta}_2 \bar{x}_{ij}$($\hat{\beta}_2$ is the effect at Level 2)
    - Nicely separates the Level 1 and Level 2 associations

Example 1: Recall our beach data. The response variable was species richness (number of different species), and available variables are NAP (the height of the sampling station relative to the mean tidal level), and Exposure (a composite measure of wave action, length of the surf zone, slope, grain size, and the depth of the anaerobic layer).

(a) What are the Level 1 units in this study? How many are there? Any Level 1 variables?

(b) What are the Level 2 units in this study? How many are there? Any Level 2 variables?

```
plot(rikzdata)
```

(c) Does Richness vary by beach? Does Richness vary with NAP? Does Richness vary with Exposure?

```
modelOLS = lm(Richness ~ 1 + BeachF, data = rikzdata)
```

(d) What is the coefficient for Beach 9? How do we interpret this coefficient?

Let's consider instead the the "random intercepts" or "variance components" or "unconditional means" model: $Y_{ij} = \beta_0 + u_j + \epsilon_{ij}$ for the i^{th} site on j^{th} beach

(e) How will $\hat{u}_1$ compare to $\hat{\beta}_1$?

Fit the null model:

(f) According to this model, how much of the variation in Richness is due to the different beaches?

Fit the multilevel model that also allows Richness to vary with NAP, after adjusting for beach:

```
model1 = lmer(Richness ~ NAP + (1 | Beach), data  = rikzdata)
```

(g) What does this model look like?

Only one beach has Exposure = 8, so we are going to combine that with Exposure 10 make this a binary variable (the rest are exposure 11).

```
rikzdata$ExposureCat = (rikzdata$Exposure > 10)
```

Fit the multilevel model that also allows Richness to vary with NAP and Exposure, after adjusting for beach:

```
model2 = lmer(Richness ~ NAP + ExposureCat + (1 | Beach), data  = rikzdata)
```

(h) What is the main change? As expected? How do we interpret the coefficient of exposure? How do we interpret the $\hat{u}_i$ values?

(i) What does this model predict for the Richness when NAP = 0.045 for the "average" beach with low exposure? What does this model predict for the first observation in the first beach? What is the first observation in the first beach? What is the residual for this observation?

Is this model valid?

```
performance::check_model(model2)
```

(j) What do you learn from this output?

(k) Which one is observation 22? Is it influential?

Notes: 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
  - "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)

To do 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?

```
(model.matrix(model2) %*% fixef(model2))[1,1]
fitted.values(model2)[1]
residuals(model2)[1]
ranef(model2)[[1]][1,1]
(rikzdata$Richness[1] - model.matrix(model2) %*% fixef(model2))[1,1]
```