Stat 414 - Day 23
Modeling Heteroscedasticity (5.1)

Last Time

- Adding a Level 2 variable can explain variation in intercepts; Adding a cross-level interaction can explain variation in slopes
  - With a Level 2 variable, we are thinking of the intercepts and slopes as "outcomes" and then running a regression model to explain that variation
  - Can add cross-level interactions even without random slopes and/or a cross-level interaction can explain "all" the variation in random slopes
- Also keep in mind that the variance of the intercepts is at "x = 0"

Example 1: Return to our Beach data

(a) Fit the random intercepts using the lme function from the nlme package.

```
library(nlme)
summary(model1 <- lme(Richness ~ NAP, random = ~ 1 | Beach, data = rikzdata))
```

Nice that it gives us AIC, BIC, log likelihood automatically, but we still need to count up the number of parameters estimated and it only gives us SDs for the random components.

(b) What is the estimated Level 1 variance and the estimated Level 2 variance? What is the total variance? What is the ICC?

The nlme package allows us to see that variance-covariance matrix for each beach. Here is that matrix for the five observations in Beach 1, and then the correlation matrix.

```
vcm = getVarCov(model1, type = "marginal"); vcm
cov2cor(vcm[[1]])
```

(d) What are the values along the diagonal of the vcm matrix? The off-diagonal values?

(e) What are the off-diagonal values after running cov2cor? How do we convert?

Now let's look at the random coefficients model:

```
model2 = lme(Richness ~ NAP, random = ~ 1 + NAP | Beach, data = rikzdata)
```

(f) Based on the graph of the model, is the variance in the responses the same for each NAP value?

Looking at the variance covariance matrix:

(g) According to the fitted model, is the variance constant? Which observations in Beach 1 have larger variance?

Examine the data for the 5 observations for beach 1:

(h) What is true about the NAP values for the observations with higher predicted variance? The smallest predicted variance? In other words, the variance in the predicted Richness values (increases/decreases) with NAP?

(i) According to the fitted model, is the correlation between two observations within beach 1 the same for any two observations, or does it vary depending on which two observations you are pairing? Identify two observations in beach 1 that are more highly correlated, and two observations in beach 1 that are less correlated. (Do you see a pattern in their NAP values?)

The point is that a random slopes model also allows us to model heterogeneity in the data (y_{ij}) and that the amount of correlation between two observations depends on the corresponding x_{ij} values. On HW 6, you will show that the variance is a quadratic function in NAP: $\tau_0^2 + x_{ij}^2\tau_1^2 + 2x_{ij}\tau_{01} + \sigma^2$ so is minimized at $x_{ij} = -b/2a = -2\tau_{01}/(2\tau_1^2) = -\tau_{01}/\tau_1^2$

(j) Suggest 3 different ways to find $\hat{\tau}_{01}$for this model.

(k) Find the value of NAP that minimizes $Var(y_{ij})$ for our fitted model. Is this a value in the range of our data?? (Does your answer agree with the graph of the model?)

The idea is when the correlation between the intercepts and slopes is negative, the lines are "fanning in" and variability is smaller for larger x values. If the correlation between the slopes and intercepts is positive, then the lines will "fan out" and variability in y is increasing for larger x values. But also watch for the point where they switch from fanning in to fanning out… If the correlation is close to zero, then there is no fanning, and you will have a scatter of positive and negative lines.

You will show in HW 6, that the covariance between two observations also depends on the x values: $Cov(y_{ij}, y_{kj}) = \tau_0^2 + (x_{ij} + x_{kj})\tau_{01} + +x_{ij}x_{kj}\tau_1^2$

(m) What happens to the covariance between two observations when NAP = 0 (for both observations)? What about the correlation?

Notes:

- Bottom line: the variance and covariance in our data (y_{ij}) values now depend on the x_{ij} values, but τ_0^2 represents the variation in the intercepts when $x = 0$ and $(\tau_0^2)/(\tau_0^2 + \sigma^2)$ is the correlation of two measurements on the same beach with $x = 0$.
- But in general now have "fanning lines" and it may not make sense to calculate ICC. Or do so conditional on a particular value of x. In general, be more detailed when talking about "variability explained."