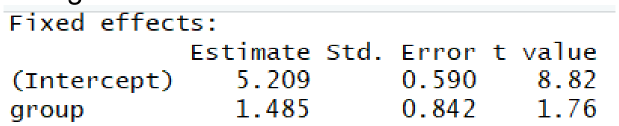
Stat 414 – Day 10
Random Slopes (5.1)

Last Time:
- Multilevel models find the "right" standard errors
  - Adjusting the standard error with design effect: $.5038 \times \sqrt{152/54.9} \approx .838$
  - Fitting a multilevel model



```
Fixed effects:
            Estimate Std. Error t value
(Intercept)    5.209      0.590    8.82
group          1.485      0.842    1.76
```

- When add variables, can compute percentage reduction in total variance or at each level
  - Changing the variance components changes the "conditional" ICC, design effect
- Three-level models
  - Allowed intercepts (think $\bar{y}$'s) to vary across schools and across classes within schools
  - Can test significance of variance component but probably best to match data structure
  - Usual methods for adding predictors, interactions
    - Adding Level 1 predictors may explain variation at each level (increase?)

Example 1: Recall the RIKZ dataset, where 5 measurements were taken on each of 9 beaches. The response variable was species richness (different number of species), and available variables were 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). (Zuur et al.)

(a) Does there appear to be a relationship between species richness and NAP? Statistically significant? In the expected direction?

negative ($t = -4.5$)

(b) Does there appear to be differences in species richness across the beaches?

highest beach 1 & 2

(c) Fit the null model that allows for the intercepts to vary by beach. Which beach has the largest intercept? Which has the smallest? What is the AIC for this model?

highest beach 1 & 2 AIC = 261

Smallest beach 4, 7

(d) Is the (conditional) association between NAP and richness statistically significant?

yes $t = -5.1$

(e) But perhaps the slopes also vary by beach. Is there evidence of this? Describe the nature of this interaction.

intercept
6.3, 9.95, 3.5, 3 4 3.1, 12.7
13.3, 10.8

slope
-2.97, -1.89, -1.3, -1.75, [illegible], -8.9
-4.2, -0.37, -1:25

(f) What are downsides to fitting a separate line for each beach?

Goal is really to find a "population line"

SD(intercepts) = 4.2 SD(slopes) = 2.6

Borrowing information, pooling, shrinkage (weight by sample size or large w/in SD)

Fitting a “random slopes” model

```
lme(fixed = Richness ~ NAP, random = ~NAP | Beach)
lmer(Richness ~ NAP + (NAP | Beach))
```

Note, the intercept is assumed.

(g) Fit the random slopes model and look at the fancy graph. Is this a better fitting model? How are you deciding?

$\hat{\tau}_0$: 3.5 : SD in intercepts beach to beach
$\hat{\tau}_1$: 1.7 : SD in slopes beach to beach
$\hat{\sigma}$: 2.7 : site to site (w/in beach) variation

(h) Identify and interpret the fixed effects.

$\widehat{\text{Richness}} = 6.6 - 2.83\,\text{NAP}$
6.6: avg richness for average beach when NAP(at tidal line ht)
2.83: avg richness ↓ by 2.83 w/ ea 1 unit increase in NAP for average beach

(i) Identify and interpret the variance components. Which is larger? What does this tell you?

points closer to lines in graph
smaller residuals

(j) Identify and interpret the new correlation parameter estimate. $\hat{\tau}_{01} \approx -.99$
e.g., Beaches with larger intercepts tends to have ~~smaller~~ ______ slopes
more negative

(k) How many parameters have you added to the model by including the random slope?

τ_1 & τ_{01} (2)

(l) Write out the level-by-level model equations and the composite model equation. $\varepsilon_{ij} \sim N(0, \sigma^2)$

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j}\,\text{NAP} + \varepsilon_{ij}$

Level 2: $\beta_{0j} = \beta_{00} + u_{0j}$ $\quad u_{0j} \sim N(0, \tau_0^2)$ $\quad \text{cov}(u_{0j}, u_{1j}) = \tau_{01}$
$\beta_{1j} = \beta_{10} + u_{1j}$ $\quad u_{1j} \sim N(0, \tau_1^2)$

Composite: $Y_{ij} = \underbrace{\beta_{00} + \beta_{10}\,\text{NAP}}_{\text{fixed}} + \underbrace{u_{1j}\,\text{NAP} + u_{0j} + \varepsilon_{ij}}_{\text{random}}$

(m) What is V(Y_{ij})? $\text{Cov}(\varepsilon_{ij}, u\text{'s}) = 0$

$$V(Y_{ij}) = \text{Var}(u_{1j}\,\text{NAP} + u_{0j}) + \text{Var}(\varepsilon_{ij})$$
$$= \text{NAP}^2\,\text{Var}(u_{1j}) + \text{Var}(u_{0j}) + 2\,\text{NAP}\,\text{Cov}(u_{1j}, u_{0j}) + \cdots$$
$$= \text{NAP}^2\,\tau_1^2 + \tau_0^2 + 2\,\text{NAP}\,\tau_{01} + \sigma^2$$

(n) Now consider adding Exposure to the model. Is this a Level 1 or Level 2 variable? What do you expect to change in the model?

Level 2, might explain beach-~~to~~-beach variation (intercepts & slopes)

(o) Write out the level-by-level model equations.

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j} NAP + \varepsilon_{ij}$

Level 2: $\beta_{0j} = \beta_{00} + \beta_{01} Exp + u_{0j}$

$\beta_{1j} = \beta_{10} + \beta_{11} Exp + u_{1j}$

(p) Summarize what you learn from the R exploration.

as for higher exposure group, the slope closer to zero

flatter relationship between NAP & richness

(q) Fit the new model including Exposure and compare it to the model without Exposure (switching to ML because now focused only on fixed effects, also using the "control" option to deal with convergence issues). What is the main impact from adding this variable?

SDs

2 32 int

1.64 slopes

2.6 $\hat{\sigma}$

explained variation in intercepts beach to beach

Note: Better statistical practice is probably is start with all potential fixed effects (including interactions), and decide on the random effects (e.g., slopes and/or intercepts). Then use that model to pare down the fixed effects.

(r) Compare and contrast model 2 and model 4 (interpretations of the models)

Note: In random slopes model, be careful with the interpretation of the intercept variance the intercept-by-slope covariance, they assume x = 0.

Example 2: Reconsider achieve.txt which contained reading ~~(gevocab)~~ geread) scores for students in different schools (school).

(a) Fit a two-level model with random slopes for gevocab. Identify and interpret the “fixed” part of the fitted model.

$\widehat{geread} = 2.01 + .52\ gevocab$

(b) Is the variation between slopes large? (How far apart might the largest and smallest slopes in the population plausibly be?)

95% of slopes should be w/in

.52 - 2 (.1386) 52 + 2(.1386)

.24 .80

(c) What is the largest source of variation in these students’ reading scores?

student to student variation within schools

(d) Interpret the correlation between the slopes and intercepts.

schools w/ big slopes tend to have small positive intercepts (corr -.86)

(e) Can we add age to the model? With random slopes? Is age significantly related to reading scores? How so? How does the random variation of coefficients for this variable relate to that of gevocab? What do you conclude?

yes, though watch convergence issues

p-value = .02 so moderate significance, w/ negative coefficient, but maybe ok to treat as fixed

corr (age slopes, gevocab slopes) = -.66

(f) How would you interpret the following models?

```
lmer(geread~gevocab+gender + (1|school) + (gender|class), data=achieve)
lmer(geread~gevocab+gender + (-1+gender|school) + (1|class), data=achieve)
lmer(geread~gevocab+gender + (1|corp) + (1|school) + (gender|class), data=achieve)
```

① random intercepts for school, class variation in slopes for gender & intercepts

② random intercepts for class, random slopes for gender across schools

③ random intercepts for school, class, corp (district) & random slopes for gender by class