Stat 414 - Day 10
Level 2 variables (Ch. 4)

Previously

- Basic multilevel model: $Y_{ij} = \beta_0 + u_j + \epsilon_{ij}$ where we are assuming $\epsilon_{ij} \sim N(0, \sigma^2)$ and $u_j \sim N(0, \tau^2)$ and $cov(\epsilon_{ij}, u_j) = 0$
- Assessing whether variance components are statistically significant. In other words, is the group mean variation significant, a LRT with df = 1, and describing the "model distribution" of the random effects (e.g., the median school).
- Assessing whether adding variables to the model explains significantly more variation in the response including percentage of Level 1 and Level 2 variation explained (and overall) and t-test/partial F-test or LRT.

Example 1: Beach data

Data were collected from nine beaches along the Dutch coast. Five readings were taken for each beach, measuring the species richness (number of different species).

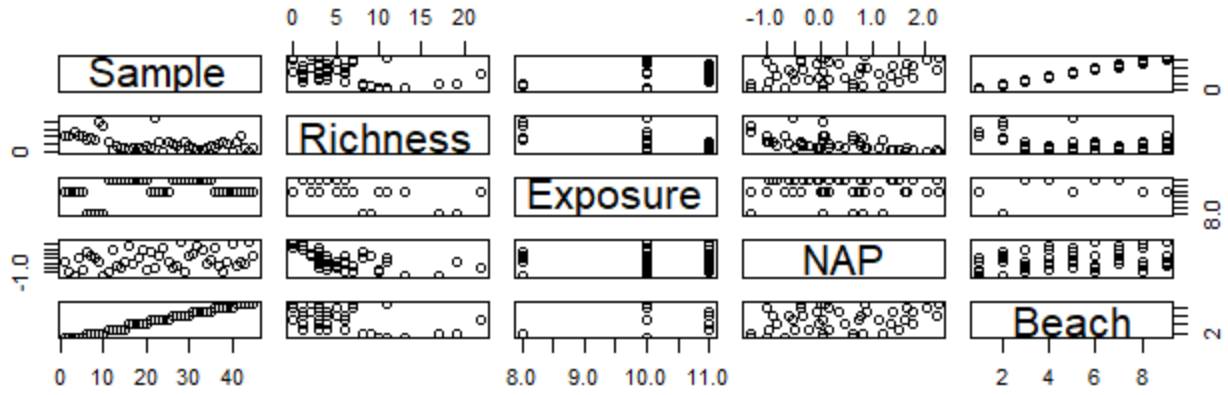
```
rikzdata<-read.table("http://www.rossmanchance.com/stat414/data/RIKZ.txt",header=T)
head(rikzdata)
  Sample Richness Exposure    NAP Beach
1      1       11       10  0.045     1
2      2       10       10 -1.036     1
3      3       13       10 -1.336     1
4      4       11       10  0.616     1
5      5       10       10 -0.684     1
6      6        8        8  1.190     2
rikzdata$Beach = factor(rikzdata$Beach)
```

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

Level 1 units = beach sites (45 total); Level 2 units = beaches (9); Level 1 variable is NAP and Exposure is Level 2 variable (doesn't change across sites, only across beaches).

Examine the data

```
plot(rikzdata)
```



(b) Does Richness vary by beach? Does Richness vary with NAP? Does Richness vary with Exposure? Does NAP appear to vary with Exposure? Do the natures of these associations make sense in context? Ask ChatGPT?

There is a negative association between NAP and Richness. Beaches with less exposure appear to have more Richness. Beaches with more exposure tend to have higher NAP? That's important to note as they may explain some of the same variation in Richness.

Probably should have mentioned - also note that the NAP (means) vary across the beaches.

(c) Using good statistical notation, write out the model equation for the model using random intercepts for the beaches.

$Y_{ij} = \beta_0 + u_j + \epsilon_{ij}$ with $u_j \sim N(0, \tau^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$

Alternatively, we can write the **null model** as two "level equations":

Level 1: $Y_{ij} = \beta_{0j} + \epsilon_{ij}$ with $\epsilon_{ij} \sim N(0, \sigma^2)$

Level 2: $\beta_{0j} = \beta_{00} + u_j$ with $u_j \sim N(0, \tau^2)$

Confirm that these equations match the "composite" equation.

(d) How would you add the NAP variable to the level-model equations? (Which level? What indices? what are the parameters in the model?)

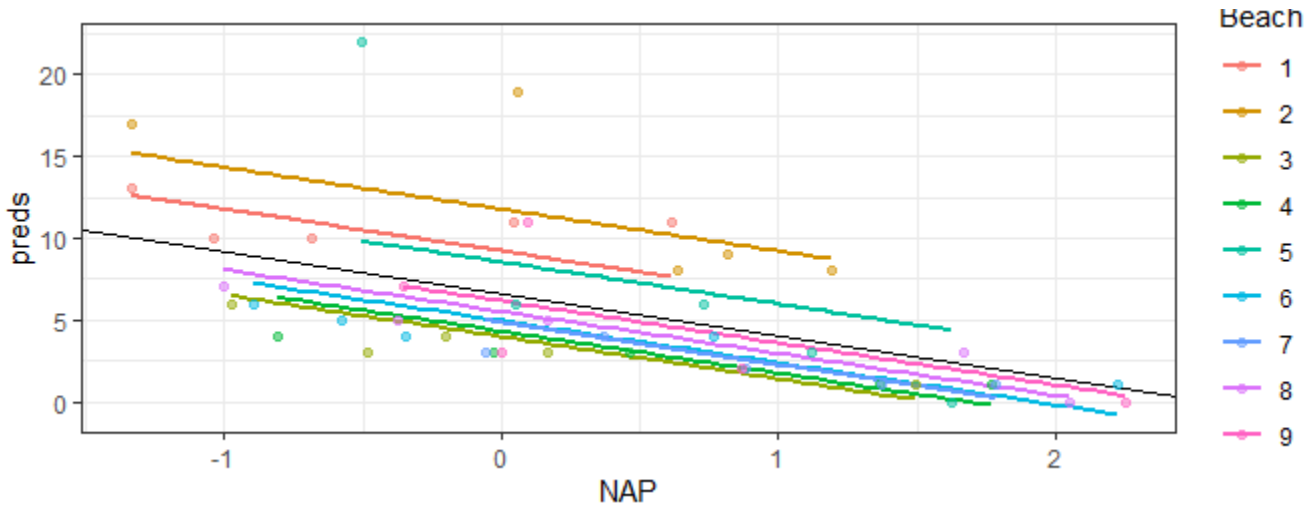
Level 1: $Y_{ij} = \beta_{0j} + \beta_1 NAP_{ij} + \epsilon_{ij}$ with $\epsilon_{ij} \sim N(0, \sigma^2)$

Level 2: $\beta_{0j} = \beta_{00} + u_j$ with $u_j \sim N(0, \tau^2)$

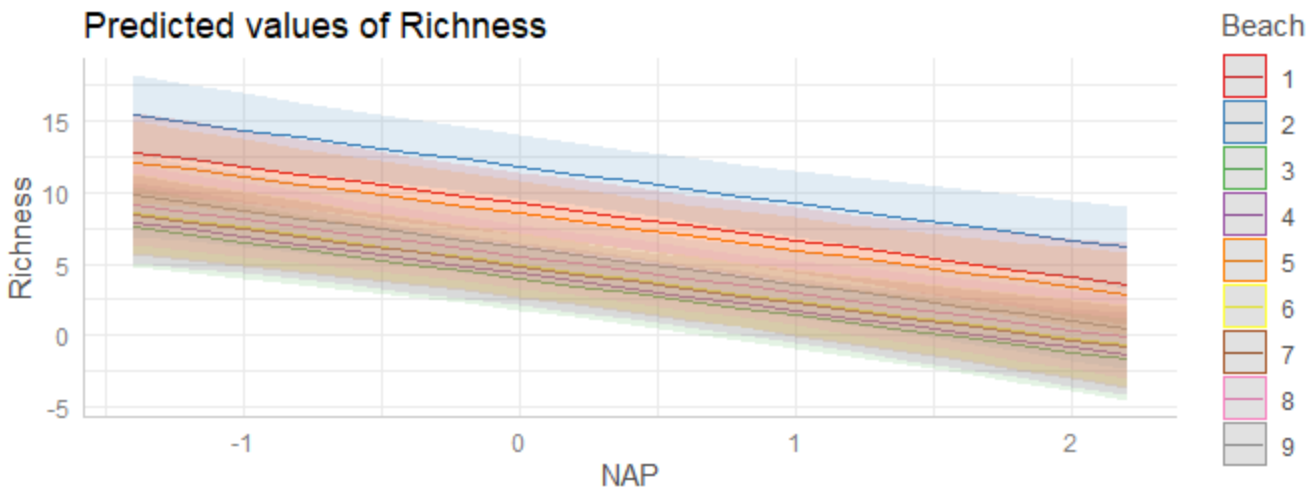
Fit the multilevel model that also allows Richness to vary with NAP, after adjusting for beach:

```
model1 = lmer(Richness ~  NAP + (1 | Beach), data  = rikzdata)  #Note the 1 + for t
he intercept is assumed!
summary(model1, corr=F)
```

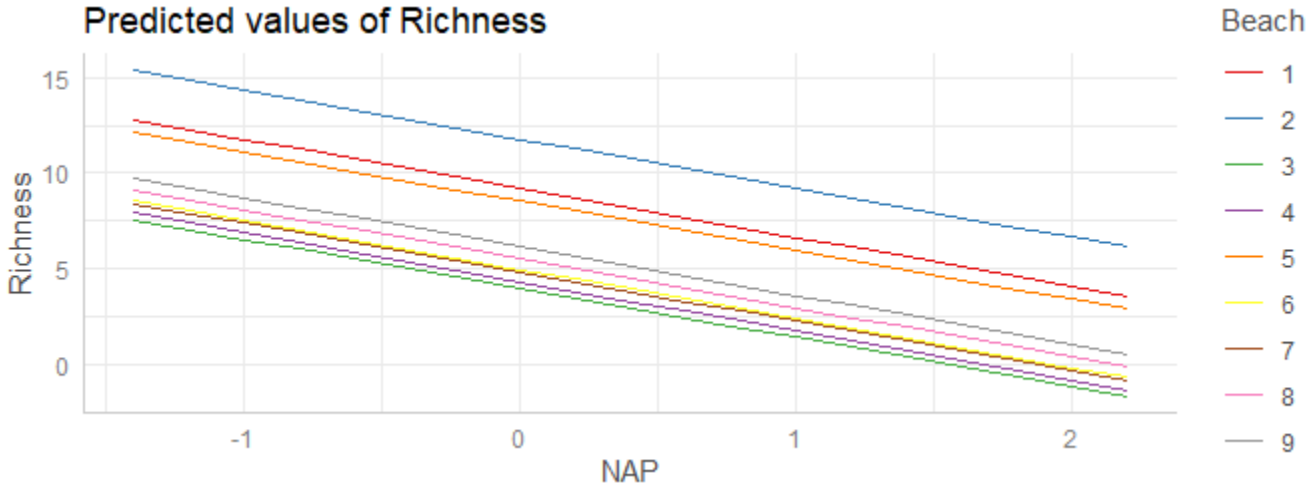
```
Linear mixed model fit by REML ['lmerMod']
Formula: Richness ~ NAP + (1 | Beach)
   Data: rikzdata

REML criterion at convergence: 239.5

Scaled residuals:
   Min     1Q Median     3Q    Max
-1.423 -0.485 -0.158  0.252  3.979

Random effects:
 Groups   Name        Variance Std.Dev.
 Beach    (Intercept) 8.67     2.94
 Residual             9.36     3.06
Number of obs: 45, groups:  Beach, 9

Fixed effects:
            Estimate Std. Error t value
(Intercept)    6.582      1.096    6.01
NAP           -2.568      0.495   -5.19
#Our predicted model
#library(tidyverse)
preds = predict(model1, newdata = rikzdata)
ggplot(rikzdata, aes(x = NAP , y = preds , group = Beach, color = Beach )) +
geom_smooth(method = "lm", alpha = .5, se = FALSE) +
geom_abline(intercept = 6.582, slope = -2.568) +
geom_point(data = rikzdata, aes(y = Richness, color=Beach), alpha = .5) +
  theme_bw()
```



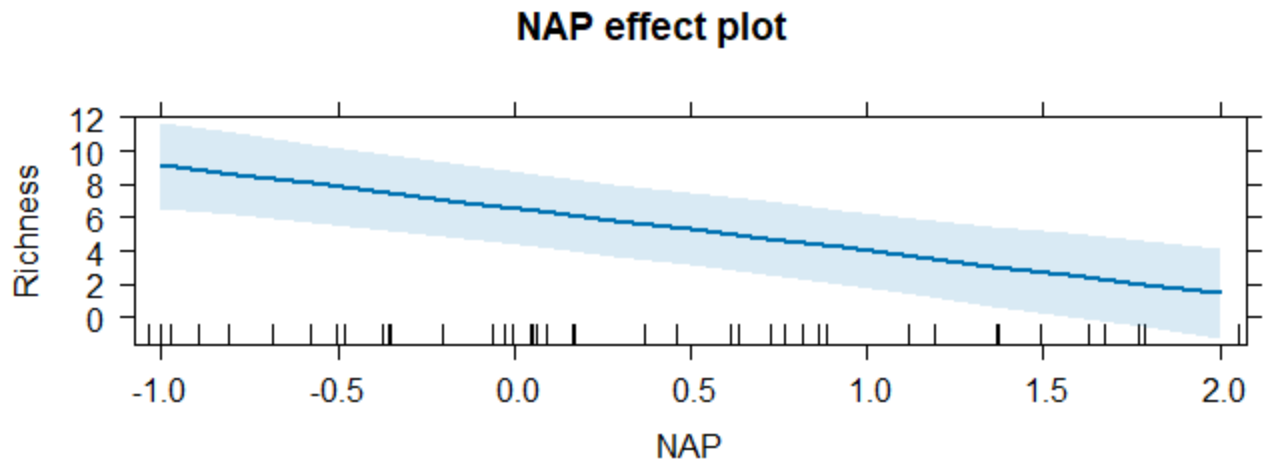

```
#See also
library(ggeffects)
ggpredict(model1, terms = c("NAP", "Beach"), type = "random") |> plot()
```




```
plot(ggpredict(model1, terms = c("NAP", "Beach"), type = "random"), show_ci=FALSE)
```




```
#See also
library(effects)
plot(effects::allEffects(model1))
```



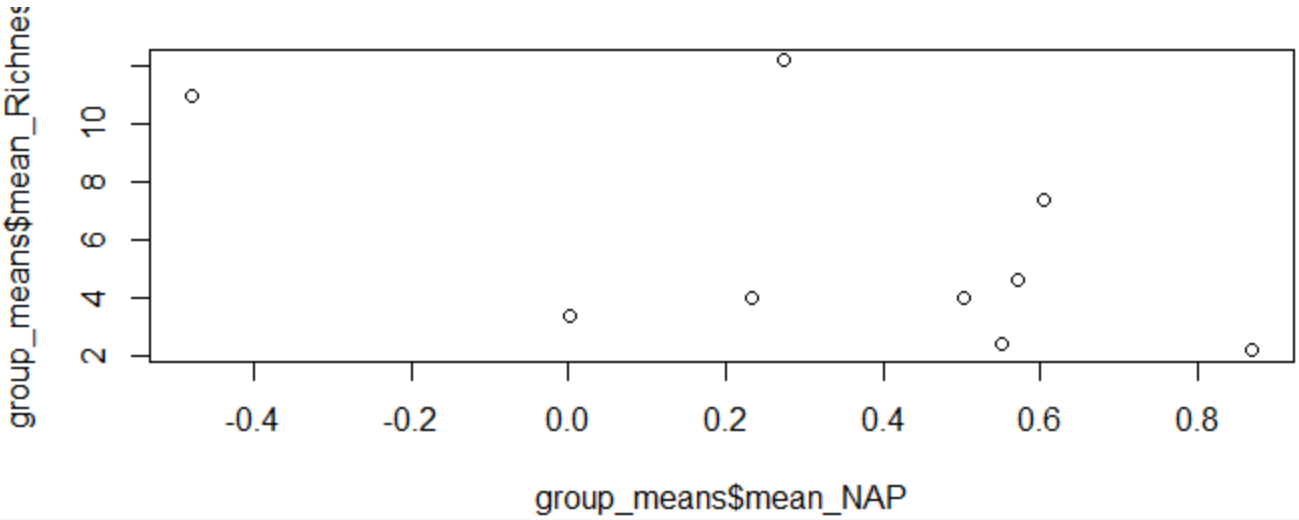

```
performance::r2(model1, by_group = TRUE)
# Explained Variance by Level

Level   |    R2
---------------
Level 1 | 0.396
Beach   | 0.173
```

(e) How do you interpret the ("Pseudo-R^2") variance explained output?

NAP explains 39.6% of the variation in Richness values within beaches and 17.3% of the beach-to-beach variation in average Richness. A Level 1 variable can explain variability at Level 2 if the Level 1 variable varies across the Level 2 units.

At the beach level, there is a weak association between avg Richness and avg NAP for these 9 beaches:

```
group_means <- rikzdata |>
  group_by(Beach) |>
  summarise(
    mean_Richness = mean(Richness, na.rm = TRUE),
    mean_NAP = mean(NAP, na.rm = TRUE)
  )

plot(group_means$mean_Richness~ group_means$mean_NAP)
```




```
summary(lm(group_means$mean_Richness~ group_means$mean_NAP))

Call:
lm(formula = group_means$mean_Richness ~ group_means$mean_NAP)

Residuals:
   Min     1Q Median     3Q    Max
-4.040 -2.263 -0.855  1.143  6.142

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)              7.44       1.50    4.97   0.0016 **
group_means$mean_NAP    -5.04       2.92   -1.73   0.1279
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.3 on 7 degrees of freedom
Multiple R-squared:  0.299, Adjusted R-squared:  0.198
F-statistic: 2.98 on 1 and 7 DF,  p-value: 0.128
```

Probably not worth adding avg NAP to the model. So let's try a different Level 2 variable.

Adding a Level 2 variable

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$ExposureCat = factor(rikzdata$Exposure > 10,
                              levels=c(FALSE, TRUE),
                              labels=c("<=10", ">10")
)
contrasts(rikzdata$ExposureCat)
     >10
<=10   0
>10    1
```

Note: I used the contrasts command to see that >10 (high exposure) will be coded 1 and <= 10 (low exposure) will be coded 0.

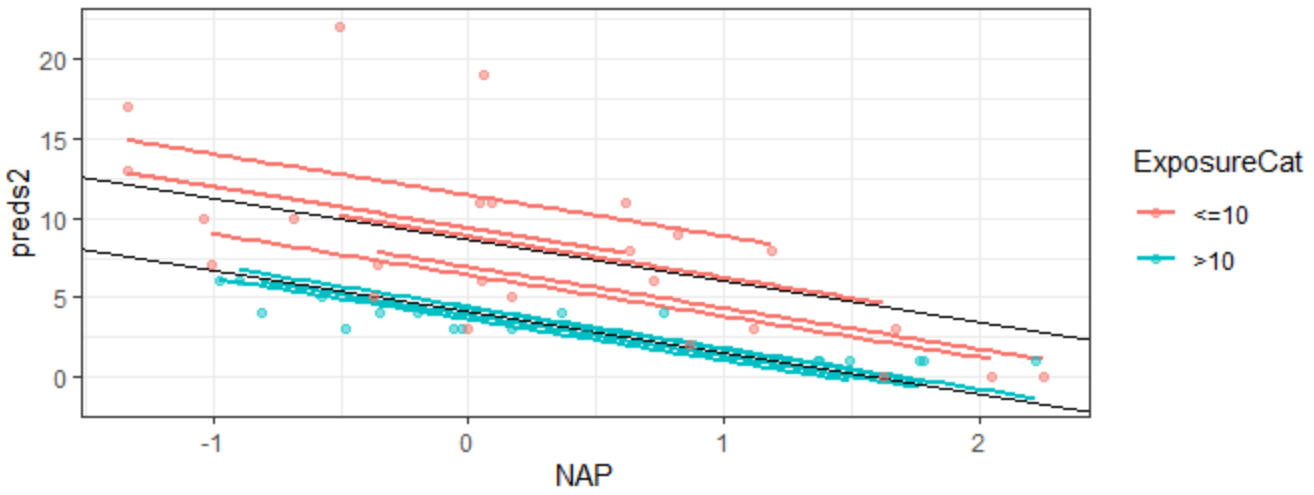
(f) How would you add the Exposure variable to the model equations? (Which level? What indices?)

Level 1: $Y_{ij} = \beta_{0j} + \beta_1 NAP_{ij} + \epsilon_{ij}$ with $\epsilon_{ij} \sim N(0, \sigma^2)$

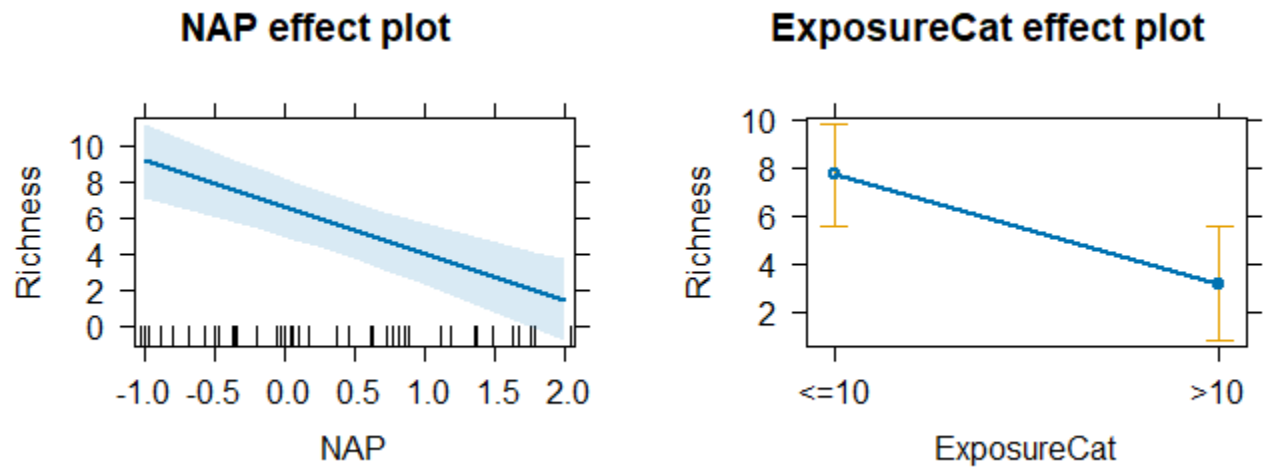
Level 2: $\beta_{0j} = \beta_{00} + \beta_{01} Exp_j + \beta_{01} Exp_j + u_j$ with $u_j \sim N(0, \tau^2)$

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)
summary(model2, corr=F)
Linear mixed model fit by REML ['lmerMod']
Formula: Richness ~ NAP + ExposureCat + (1 | Beach)
   Data: rikzdata

REML criterion at convergence: 230.6

Scaled residuals:
   Min     1Q Median     3Q    Max
-1.516 -0.482 -0.122  0.292  3.878

Random effects:
 Groups   Name        Variance Std.Dev.
 Beach    (Intercept) 3.64     1.91
 Residual             9.36     3.06
Number of obs: 45, groups:  Beach, 9

Fixed effects:
               Estimate Std. Error t value
(Intercept)       8.601      1.059    8.12
NAP              -2.582      0.488   -5.29
ExposureCat>10   -4.533      1.576   -2.88
#Graph the model
preds2 = predict(model2, newdata = rikzdata)
ggplot(rikzdata, aes(x = NAP , y = preds2 , group = Beach, color = ExposureCat )) +
geom_smooth(method = "lm", alpha = .5, se = FALSE) +
geom_abline(intercept = 8.6011, slope = -2.5817) +
  geom_abline(intercept = 8.6011 - 4.532, slope = -2.5817) +
geom_point(data = rikzdata, aes(y = Richness, color=ExposureCat), alpha = .5) +
  theme_bw()
```




```
plot(effects::allEffects(model2))
```



(g) Where did the intercept and slope values come from? How do we interpret the slope of Exposure?

The fitted model is 8.6011 - 2.58 NAP - 4.53 highExp. So for low exposure beaches, the equation is 8.6011 - 2.58 NAP and for high exposure beaches, the equation is 8.6011 - 4.53 (=4.07) - 2.58 NAP. So 4.53 is the predicted decrease in average Richness moving from low exposure to high exposure beaches, after adjusting for NAP and Beach.

(h) What do you learn from the graphs? Does ExposureCat appear meaningful? Is it statistically significant?

So we have one intercept for low exposure beaches and a lower one for high exposure beaches and then random variation of the beaches around these means, all with the same negative association (slope) with NAP. Back to the earlier output, we have a t-value of -2.88 so we would probably consider that to be a statistically significant change in the average Richness between high and low exposure beaches after adjusting for NAP.

(i) How does adding the Exposure variable to the model change the variance components? How much variation is explained at each level?

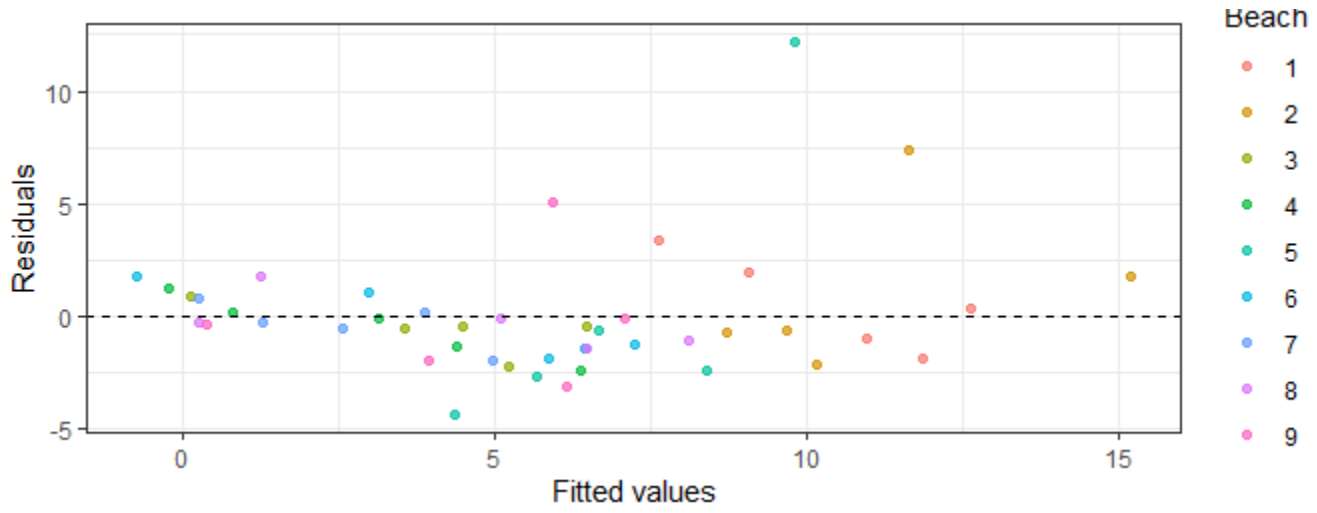
```
VarCorr(model1)
 Groups   Name        Std.Dev.
 Beach    (Intercept) 2.94
 Residual             3.06
VarCorr(model2)
 Groups   Name        Std.Dev.
 Beach    (Intercept) 1.91
 Residual             3.06
```

The Level 1 variance is unchanged, which we would expect because we added a Level 2 variable. The Level 1 variance changed by $2.944^2 - 1.0972^2)/2.944^2 x 100$. So Exposure explained 58% of the beach to beach variation in Richness (that wasn't explained by NAP).

```
#Just remember this is compared to the null model
performance::r2(model2, by_group = TRUE)
# Explained Variance by Level

Level   |    R2
---------------
Level 1 | 0.397
Beach   | 0.653
```

Examine the residuals vs. fits (Model 1):

```
rikzdata$preds <- predict(model1)
ggplot(rikzdata, aes(x = preds, y = residuals(model1), color = Beach)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  theme_bw() +
  labs(x = "Fitted values", y = "Residuals")
```



(j) What do the run of positive and negative residuals tells us?

We are missing an important variable? For example, for the salmon dots, we are consistently underpredicting high NAP values and over predicting low NAP? Allow the slopes to vary by beach? Sounds like an interaction…

Example 2: Netherlands Language Scores

Recall the Netherlands Language dataset with language test scores (langPOST) for Grade 8 students (~ age 11).

Null Model

```
neth = read.table("https://www.rossmanchance.com/stat414F20/data/NetherlandsLanguage.txt", "\t", header=TRUE)
head(neth)
  schoolnr pupilNR_new langPOST    ses IQ_verb sex Minority denomina sch_ses
1        1           3       46  -4.73    3.13   0        0        1  -14.04
2        1           4       45 -17.73    2.63   0        1        1  -14.04
3        1           5       33 -12.73   -2.37   0        0        1  -14.04
4        1           6       46  -4.73   -0.87   0        0        1  -14.04
5        1           7       20 -17.73   -3.87   0        0        1  -14.04
6        1           8       30 -17.73   -2.37   0        1        1  -14.04
  sch_iqv sch_min
1  -1.404    0.63
2  -1.404    0.63
3  -1.404    0.63
4  -1.404    0.63
5  -1.404    0.63
6  -1.404    0.63
load(url("https://www.rossmanchance.com/iscam4/ISCAM.RData"))

#Library(lme4)
nullmodel = lmer(langPOST ~ 1 + (1|schoolnr), data = neth, REML = FALSE)
performance::icc(nullmodel)
# Intraclass Correlation Coefficient

    Adjusted ICC: 0.224
  Unadjusted ICC: 0.224
```

(a) Was IQ_verb a Level 1 or Level 2 variable?

IQ_verb is Level 1

Model 1

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

(b) Write out the by-level model equations

Level 1: $Y_{ij} = \beta_{0j} + \beta_1 verbalIQ_{ij} + \epsilon_{ij}$ with $\epsilon_{ij} \sim N(0, \sigma^2)$

Level 2: $\beta_{0j} = \beta_{00} + u_j$ with $u_j \sim N(0, \tau^2)$

(c) How do you interpret the percentage in variance explained calculations?

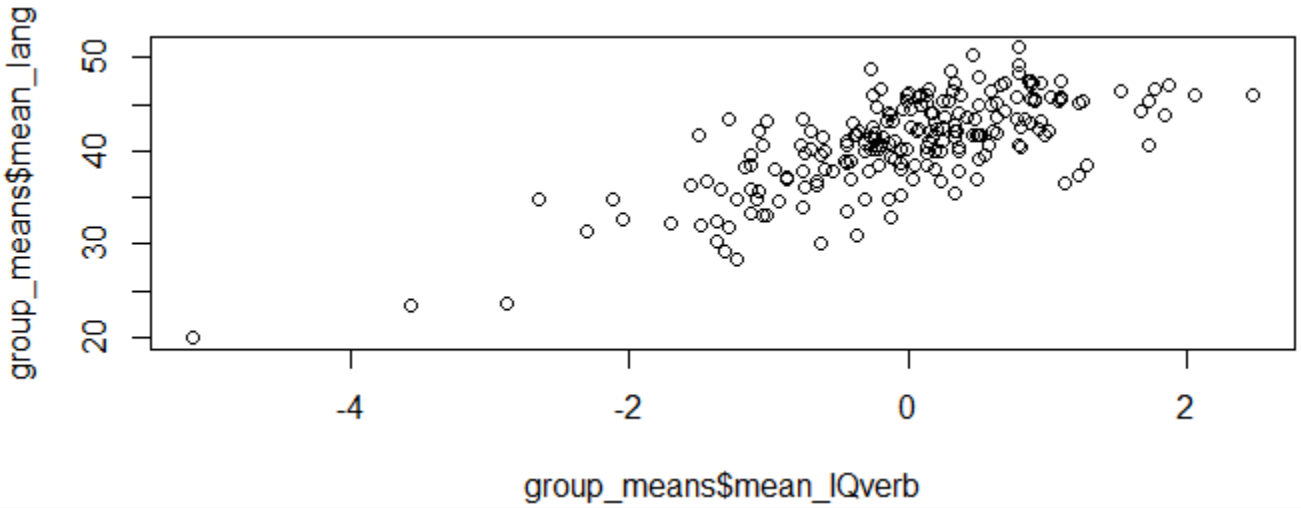
```
performance::r2(model1, by_group = TRUE)
# Explained Variance by Level

Level    |    R2
---------------
Level 1  | 0.356
schoolnr | 0.457
```

Adding IQ_verb to the model explained 35.6% of the variation at Level 1 and 45.7% (!) of the variability at Level 2 (differences among schools in avg language scores. This happens because IQ_verb also varies across the schools. Knowing the average verbal IQ at the school is even more informative than knowing a student's individual IQ in predicting a student's language score.

Level 2 association

We note a very strong positive linear association between average language score and average verbal IQ across the 211 schools:

```
group_means <- neth |>
  group_by(schoolnr) |>
  summarise(
    mean_lang = mean(langPOST, na.rm = TRUE),
    mean_IQverb = mean(IQ_verb, na.rm = TRUE)
  )

plot(group_means$mean_lang~ group_means$mean_IQverb)
```




```
summary(group_means$mean_lang~ group_means$mean_IQverb)
 Length   Class    Mode
      3 formula    call
```

An important "contextual variable" in this study is the average school IQ.

(d) In R, how we could create a variable I could use in my model? Ask ChatGPT?

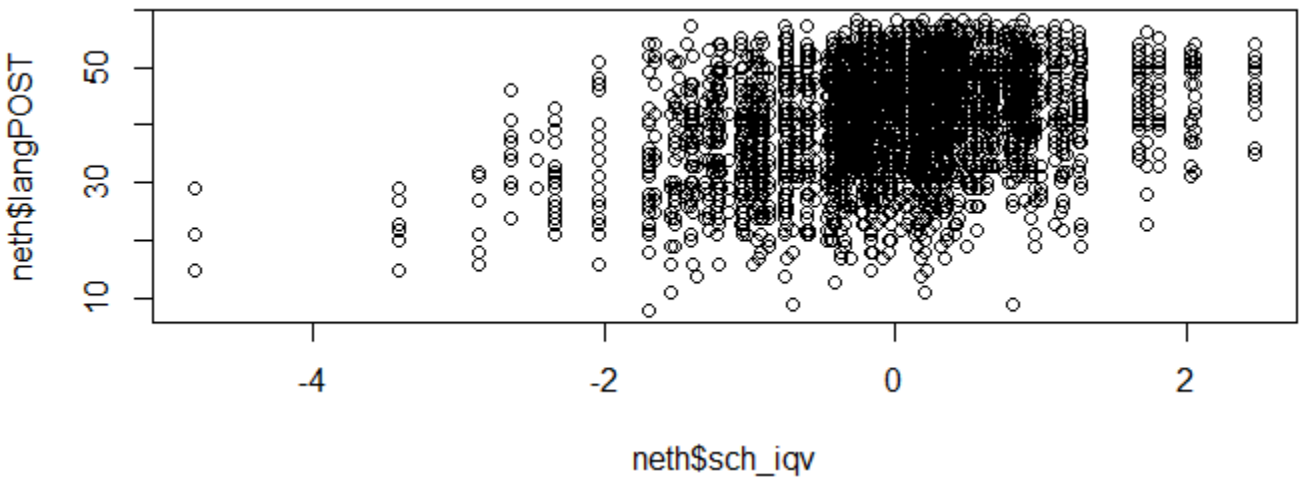
One quick way is to use the 'ave' function to add associated group mean in each row

Code

```
neth$mean_IQ  <- with(neth, ave(IQ_verb, schoolnr, FUN = mean))
head(neth)
  schoolnr pupilNR_new langPOST    ses IQ_verb sex Minority denomina sch_ses
1        1           3       46  -4.73    3.13   0        0        1  -14.04
2        1           4       45 -17.73    2.63   0        1        1  -14.04
3        1           5       33 -12.73   -2.37   0        0        1  -14.04
4        1           6       46  -4.73   -0.87   0        0        1  -14.04
5        1           7       20 -17.73   -3.87   0        0        1  -14.04
6        1           8       30 -17.73   -2.37   0        1        1  -14.04
  sch_iqv sch_min mean_IQ
1  -1.404    0.63   -1.55
2  -1.404    0.63   -1.55
3  -1.404    0.63   -1.55
4  -1.404    0.63   -1.55
5  -1.404    0.63   -1.55
6  -1.404    0.63   -1.55
```

(e) Is the school mean verbal IQ related to (average) performance score?

```
#There was already a school level variable for verbal IQ (created before missing ob
servations removed)
plot(neth$langPOST ~ neth$sch_iqv)
```



There does appear to be a strong positive linear association between student language score and school mean IQ verbal.

(f) How would you add sch_iqv to the Level 1 and Level 2 model equations?

Level 1: $Y_{ij} = \beta_{0j} + \beta_1 \times verbal.IQ_{ij} + \epsilon_{ij}$ with $\epsilon_{ij} \sim N(0, \sigma^2)$

Level 2: $\beta_{0j} = \beta_{00} + \beta_{01} mean.verbal.IQ_j + u_j$ with $u_j \sim N(0, \tau^2)$

Run the model

```
model2 = lmer(langPOST ~ 1 + IQ_verb + sch_iqv + (1 | schoolnr), data = neth, REML
= F)
summary(model2, corr = FALSE)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: langPOST ~ 1 + IQ_verb + sch_iqv + (1 | schoolnr)
   Data: neth

      AIC       BIC    logLik -2*log(L)  df.resid
    24898     24929    -12444     24888      3753

Scaled residuals:
   Min     1Q Median     3Q    Max
-4.222 -0.641  0.063  0.706  3.219

Random effects:
 Groups   Name        Variance Std.Dev.
 schoolnr (Intercept)  8.68    2.95
 Residual             40.43    6.36
Number of obs: 3758, groups:  schoolnr, 211

Fixed effects:
            Estimate Std. Error t value
(Intercept)  41.1138     0.2318  177.36
IQ_verb       2.4536     0.0555   44.22
sch_iqv       1.3124     0.2616    5.02
anova(model1, model2)
Data: neth
Models:
model1: langPOST ~ 1 + IQ_verb + (1 | schoolnr)
model2: langPOST ~ 1 + IQ_verb + sch_iqv + (1 | schoolnr)
       npar   AIC   BIC logLik -2*log(L) Chisq Df Pr(>Chisq)
model1    4 24920 24945 -12456     24912
model2    5 24898 24929 -12444     24888  24.1  1 0.00000089 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
performance::r2(model2, by_group=TRUE)
# Explained Variance by Level

Level    |    R2
----------------
Level 1  | 0.357
schoolnr | 0.521
```

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

A one unit increase in school IQ verbal is associated with an *additional* 1.31 increase in averge language score, above and beyond the individual effect. Or, it's the difference between the group effect and the individual effect.

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

The t-statistic is 5.017 > 2 (and the likelihood ratio test is significant) so yes. So we learn that the level 2 slope (in the sample) is significantly larger than the level 1 slope.

Definitions

A comparison of the within group and between group associations is essentially the Hausman specification test in econometrics.

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

Compared to the null model, we have explained 35.7% of the variation within schools and 52.1% of the variation between schools.

What if we change to the "deviation" variable, verbal_IQ - sch_iqv? (aka Group Mean Centering)

```
devIQ = neth$IQ_verb - neth$sch_iqv

model3 = lmer(langPOST ~ 1 + devIQ + sch_iqv + (1 | schoolnr), data = neth, REML =
F)
summary(model3, corr = FALSE)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: langPOST ~ 1 + devIQ + sch_iqv + (1 | schoolnr)
   Data: neth

      AIC       BIC    logLik -2*log(L)  df.resid
    24898     24929    -12444     24888      3753

Scaled residuals:
   Min     1Q Median     3Q    Max
-4.222 -0.641  0.063  0.706  3.219

Random effects:
 Groups   Name        Variance Std.Dev.
 schoolnr (Intercept)  8.68    2.95
 Residual             40.43    6.36
Number of obs: 3758, groups:  schoolnr, 211

Fixed effects:
            Estimate Std. Error t value
(Intercept)  41.1138     0.2318   177.4
devIQ         2.4536     0.0555    44.2
sch_iqv       3.7660     0.2558    14.7
anova(model1, model3)
Data: neth
Models:
model1: langPOST ~ 1 + IQ_verb + (1 | schoolnr)
```

```
model3: langPOST ~ 1 + devIQ + sch_iqv + (1 | schoolnr)
       npar   AIC   BIC logLik -2*log(L) Chisq Df Pr(>Chisq)
model1    4 24920 24945 -12456     24912
model3    5 24898 24929 -12444     24888  24.1  1 0.00000089 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
performance::r2(model3, by_group = TRUE)
# Explained Variance by Level

Level    |    R2
----------------
Level 1  | 0.357
schoolnr | 0.521
```

(j) Interpret the slope coefficients for this model.

The slope coefficient of the school mean IQ is the predicted increase (3.766) in mean language IQ, holding everyone's relative position within the school constant. So the within group effect is 2.45 (predicted increase in language score comparing two students in the same school but one has 1 pt higher on verbal IQ) and the between group effect is 3.766.

(k) Does the deviation variable explain variation at Level 1 and/or Level 2? (Compare back to the null model)

```
model5 = lmer(langPOST ~ 1 + sch_iqv + (1 | schoolnr), data = neth, REML = F)
performance::r2(model5, by_group = TRUE)
# Explained Variance by Level

Level    |    R2
----------------
Level 1  | 0.003
schoolnr | 0.593
model6 = lmer(langPOST ~ 1 + devIQ + (1 | schoolnr), data = neth, REML = F)
performance::r2(model6, by_group = TRUE)
# Explained Variance by Level

Level    |     R2
-----------------
Level 1  |  0.356
schoolnr | -0.122
library(stargazer)
stargazer( model1, model2, model3, type="text")

=================================================
                          Dependent variable:
                     ------------------------------
                                langPOST
                        (1)         (2)         (3)
-------------------------------------------------
IQ_verb              2.507***    2.454***
```

```
                       (0.054)     (0.055)

devIQ                                          2.454***
                                                (0.055)

sch_iqv                            1.312***    3.766***
                                    (0.262)     (0.256)

Constant              41.050***   41.110***   41.110***
                       (0.244)     (0.232)     (0.232)

-------------------------------------------------------
Observations            3,758       3,758       3,758
Log Likelihood      -12,459.000 -12,444.000 -12,444.000
Akaike Inf. Crit.   24,925.000  24,898.000  24,898.000
Bayesian Inf. Crit. 24,950.000  24,929.000  24,929.000
=======================================================
Note:                          *p<0.1; **p<0.05; ***p<0.01
```

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?

Computer problem 10

Recall our fake salary data

```
saldata <-read.table("https://www.rossmanchance.com/stat414/data/saldata.txt", head
er=T)
#In the fixed effects model, adjusting for major changed the sign of the semesters
coefficient
lm(salary ~ semesters, data = saldata)

Call:
lm(formula = salary ~ semesters, data = saldata)

Coefficients:
(Intercept)    semesters
      34.04         1.16
lm(salary ~ semesters + major, data = saldata)

Call:
```

```
lm(formula = salary ~ semesters + major, data = saldata)

Coefficients:
   (Intercept)       semesters  majorchemistry    majorphysics
         54.76           -2.17           38.75           19.17
```

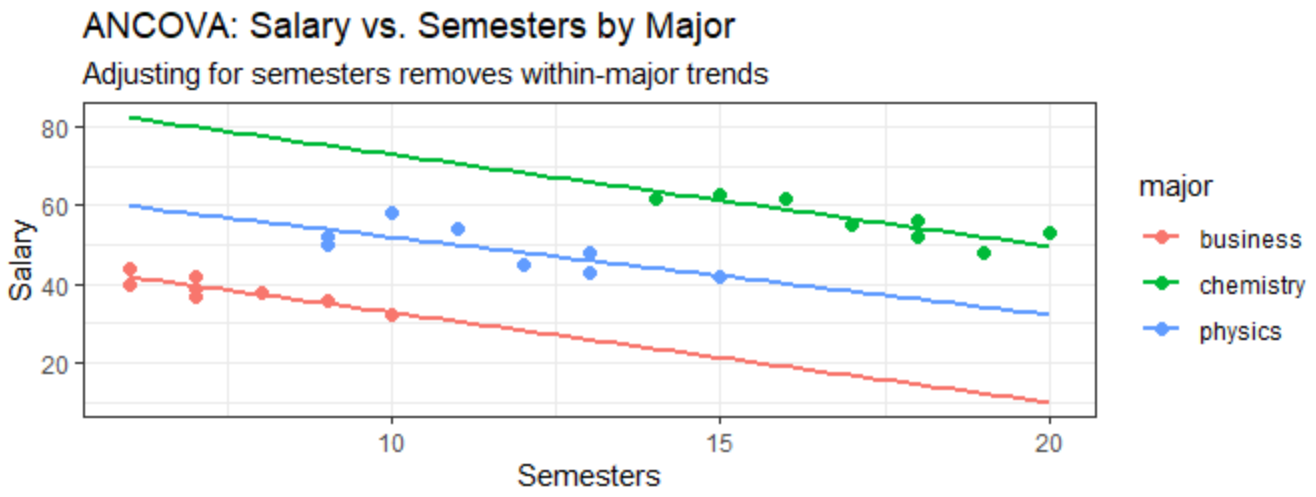
So let's think more about what it means to adjust for another variable, but in terms of adjusting the relationship between salary and major by number of semesters.

```
#Just making sure R uses effect coding and displays the group labels
saldata$majorC <- factor(saldata$major,
                    levels = c("business", "chemistry", "physics"),
                   labels = c("business", "chemistry", "physics"))
C <- contr.sum(nlevels(saldata$majorC))
colnames(C) <- levels(saldata$majorC)[1:ncol(C)]

lm(salary ~ majorC, data = saldata,
   contrasts = list(majorC = C))

Call:
lm(formula = salary ~ majorC, data = saldata, contrasts = list(majorC = C))

Coefficients:
    (Intercept)   majorCbusiness  majorCchemistry
          47.96            -9.46             8.42
```

(a) Based on the effect-coded coefficients, roughly how are apart are the mean salary for chemistry and business majors?

Now we will bring semesters into the model. Consider this graph.

```
ggplot(saldata, aes(x = semesters, y = salary, color = major)) +
     geom_point(size = 2) +
     geom_smooth(method = "lm", se = FALSE, fullrange=TRUE) +
     labs(title = "ANCOVA: Salary vs. Semesters by Major",
          subtitle = "Adjusting for semesters removes within-major trends",
          y = "Salary", x = "Semesters") +
  theme_bw()
```



Verify the difference in means that you just talked about.

(b) Now pick a number of semesters, say 12, would you say the predicted salary for a chemistry major is (closer/further) than the predicted salary for a business major to what you found in (a)? Does it matter what value you pick for number of semesters?

(c) Fit the model with both semesters and majors and discuss whether this supports your answer to (b).

Note: The unadjusted mean for chemistry is 56.375 and the semester-effect for chemistry is 17.12 - 12.04 = 5.08. The adjusted coefficient for semesters is -2.17. So the semesters-adjusted mean for chemistry is then 56.375 - (-2.17)*(5.085) = 67.41 (effect = 67.41 - 47.96 = 19.45).

Now let's see how this translates to a multilevel model. For illustration, we will treat major as a random effect.

```
model0 <- lmer(salary ~ 1 + (1 | major), data = saldata)
model1 <- lmer(salary ~ semesters + (1 | major), data = saldata)

VarCorr(model0)
 Groups   Name        Std.Dev.
 major    (Intercept) 8.81
 Residual             4.99
VarCorr(model1)
 Groups   Name        Std.Dev.
 major    (Intercept) 19.14
 Residual              2.93
```

(d) How much Level 1 variability is explained by adding the *semesters* variable?

(e) How much Level 2 variability is explained by adding the *semesters* variable?

(f) Apply what learned in (a)-(c) to explain what is happening in (e).

Bottom Line:

- A Level 1 variable can explain variation at Level 1 or Level 2, but can also increase variation at Level 2.
  - If distribution of Level 1 variable is the same across the Level 2 units, Level 2 variability won't change ($\bar{x}_j$ not changing, a deviation variable, a percentile variable)
  - If associations agree, then the Level 1 variable will also explain some of the Level 2 variability
  - If positive association at one level and negative at the other, then can increase Level 2 variability
- Level 2 variables can only explain variation at Level 1 (apart from rounding)