Stat 414 - Day 11
Interactions

Last Time

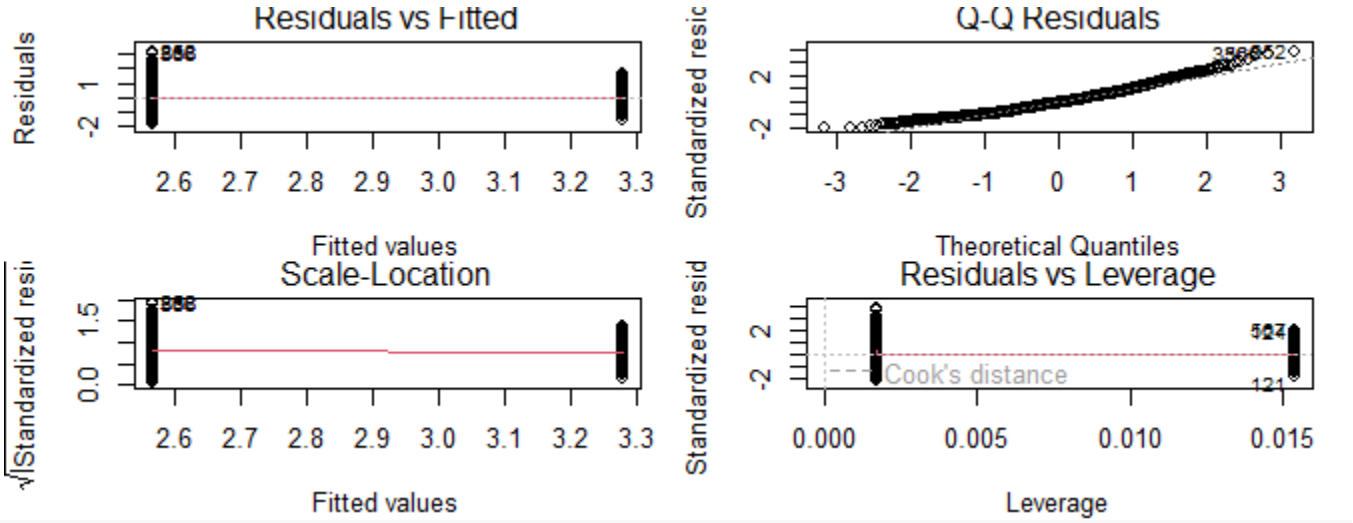
- Splitting the "composite equation" into "level equations"
- Adding Level 1 and Level 2 variables into the multilevel model and seeing what Level 1 and/or Level 2 (and total) variance is explained.
- Visualizing fitted models

Example 1: Forced Expiratory Volume (FEV)

Data were collected on 654 youths in the area of East Boston during the middle to late 1970s. The youth in the study were of ages 3 to 19 years, an age period during which much physical development, such as increase in lung capacity, takes place. The objective was to analyze the relationship between smoking status, and forced expiratory volume (FEV, measured in liters). (FEV is a measure of strength of a person's lungs – the maximum volume of air a person can blow out in the first second; higher numbers are better/healthier lungs)

```
FEVdata = read.table("https://www.rossmanchance.com/stat414F20/data/FEV.txt", heade
r=TRUE)
contrasts(factor(FEVdata$Smoker)) #see how R will code the variable
    yes
no    0
yes   1
#FEVdata$Smoker = factor(FEVdata$Smoker)
#contrasts(FEVdata$Smoker) <- contr.sum(levels(FEVdata$Smoker))
# effect coding
#contrasts(FEVdata$Smoker)
model1 <- lm(FEV ~  Smoker, data = FEVdata)
summary(model1)

Call:
lm(formula = FEV ~ Smoker, data = FEVdata)

Residuals:
   Min     1Q Median     3Q    Max
-1.775 -0.634 -0.102  0.480  3.227

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5661     0.0347   74.04   <2e-16 ***
Smokeryes     0.7107     0.1099    6.46    2e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.841 on 652 degrees of freedom
Multiple R-squared:  0.0602,    Adjusted R-squared:  0.0588
F-statistic: 41.8 on 1 and 652 DF,  p-value: 1.99e-10
```

```
par(mfrow=c(2,2))
plot(model1)
```




```
par(mfrow=c(1,1))
```

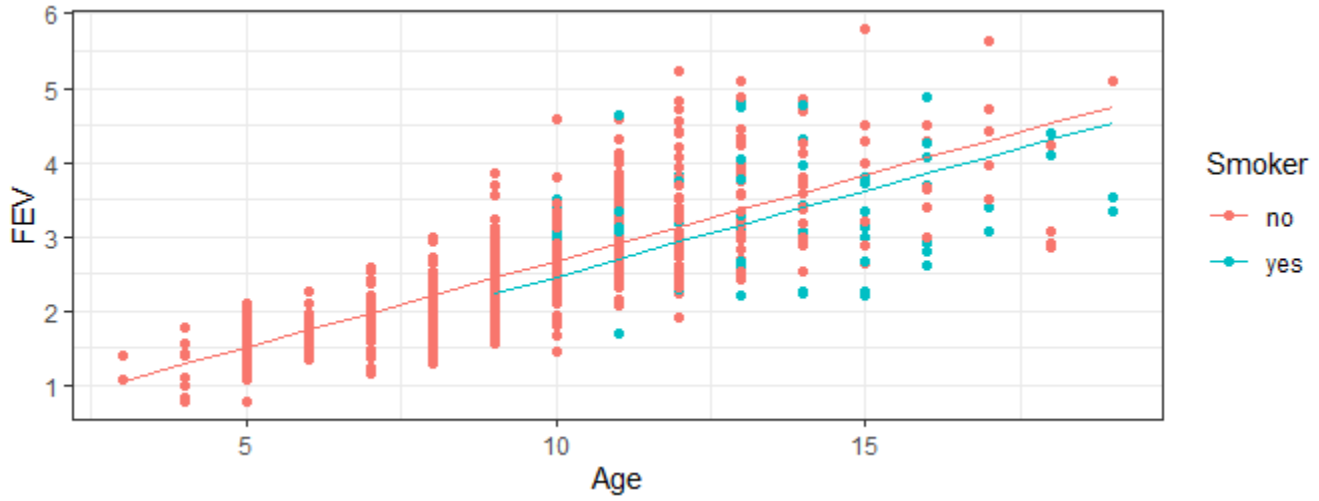
(a) Interpret the coefficient of smoker in this model. Statistically significant? Any problems with the model?

On average, nonsmokers have 0.71 liters lower FEV than smokers, with a very small p-value so statistically significant (t = 6.464), though we do see some evidence of unequal variance in FEV values between the smokers and non-smokers

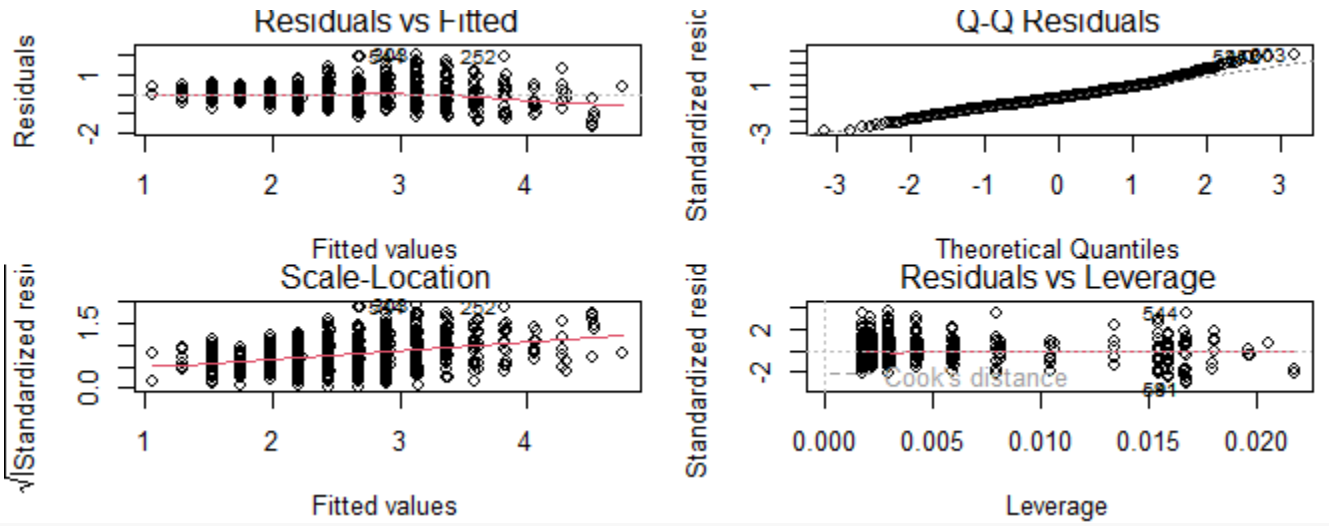
Add Age to the model

```
model2 <- lm(FEV ~ Smoker + Age, data = FEVdata)
summary(model2)

Call:
lm(formula = FEV ~ Smoker + Age, data = FEVdata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6653 -0.3564 -0.0508  0.3495  2.0894

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.36737    0.08144    4.51 0.0000076 ***
Smokeryes   -0.20899    0.08075   -2.59    0.0099 **
Age          0.23060    0.00818   28.18   < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.565 on 651 degrees of freedom
Multiple R-squared:  0.577, Adjusted R-squared:  0.575
F-statistic:  443 on 2 and 651 DF,  p-value: <2e-16
```

```
newx1 = data.frame(Age = FEVdata$Age, Smoker = rep("yes", 654))
newx2 = data.frame(Age = FEVdata$Age, Smoker = rep("no", 654))
fits1=predict(model2, newx1)
fits2=predict(model2, newx2)

#may need to copy these into Session window/run them as a set
#plot(FEVdata$FEV ~ FEVdata$Age, col=as.factor(FEVdata$Smoker))
#lines(FEVdata$Age, fits1, col=1)
#lines(FEVdata$Age, fits2, col=2)

#or, with tidyverse
ggplot(FEVdata,aes(x=Age,y=FEV,color= Smoker))+
  geom_point()+
  geom_line(aes(y=predict(model2))) +
  theme_bw()
```




```
par(mfrow=c(2,2))
plot(model2)
```




```
par(mfrow=c(1,1))
```

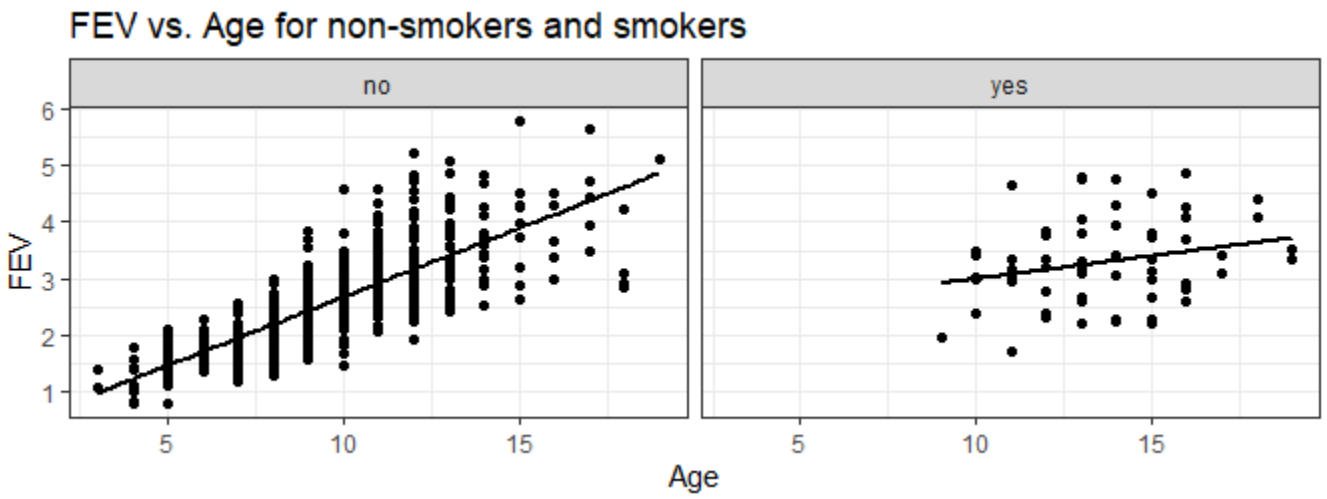
(b) How do we interpret the intercept, coefficient of smoker, and coefficient of age in this model? Any problems with the model?

Intercept: predicted FEV for nonsmoker at 0 years of age; Smoker: comparing smokers and non-smokers of the same age, the smokers have a predicted FEV .209 liters lower than nonsmokers. Comparing smokers to smokers or non-smokers to non-smokers, each one-year increase in age is associated with a 0.231 liter increase in FEV. Variability in residuals appears to increase with age.

Including the binary variable allows the intercepts to differ, but we are still assuming the slopes are the same.

Produce a graph to decide whether there is evidence that the relationship between FEV and age differs for the smokers and nonsmokers.

```
#coplot(FEV ~ Age | Smoker, data = FEVdata,
#       panel = function(x,y,...) {
#         panel.smooth(x,y)
#        abline(lm(y ~ x), col="blue")
#        }
#       )
#or
FEVdata |>
  ggplot(aes(x = Age, y = FEV)) +
  geom_point()+
  facet_wrap(~Smoker) +
  geom_smooth(method = "lm", se=FALSE, color="black", formula = y ~ x) +
  labs(title = "FEV vs. Age for non-smokers and smokers") +
 theme_bw()
```



(c) What do you learn?

The rate of growth with age is larger for the nonsmokers compared to the smokers.

Definition

A quantitative variable and a categorical variable $interact$ if the slopes of the regression lines differ. (After all, it's the slope that tells us about the association between the two variables, so this says the association between the response and the quantitative variable depends on the category of the categorical variable.) To include an interaction between x_1 and x_2 in the model, we literally multiply x_1 and x_2 together and add this variable to the model.

(d) What does it mean to multiply Smoker and Age (one categorical and one quantitative variable) together?

but the categorical variable has been coded numerically so we can literally multiply the columns.

Add the interaction to the model

```
#You can make R do the multiplication for you by including Smoker:Age with a colon
to signify an interaction
model3 = lm(FEV ~ Smoker + Age + Smoker:Age, data = FEVdata)
#or
model3 = lm(FEV ~ Smoker*Age, data = FEVdata)
#notice the possible short-cut here, the * means include all 3 terms
summary(model3)

Call:
lm(formula = FEV ~ Smoker * Age, data = FEVdata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7664 -0.3495 -0.0336  0.3368  2.0599

Coefficients:
             Estimate Std. Error t value   Pr(>|t|)
(Intercept)   0.25340    0.08265    3.07     0.0023 **
Smokeryes     1.94357    0.41428    4.69 0.00000331 ***
Age           0.24256    0.00833   29.11    < 2e-16 ***
Smokeryes:Age -0.16270    0.03074   -5.29 0.00000016 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.554 on 650 degrees of freedom
Multiple R-squared:  0.594, Adjusted R-squared:  0.592
F-statistic:  317 on 3 and 650 DF,  p-value: <2e-16
```

(e) Write out the full equation and then write out the equation (FEV vs. age) for the smokers and the non-smokers.

Predicted FEV = 0.25 + 1.94 smoker.yes + 0.242 age - 0.163 smoker.yes*age;
Nonsmokers: -.243 + 0.243 age;
Smokers = (0.25 + 1.94) + (.243 - .153) age

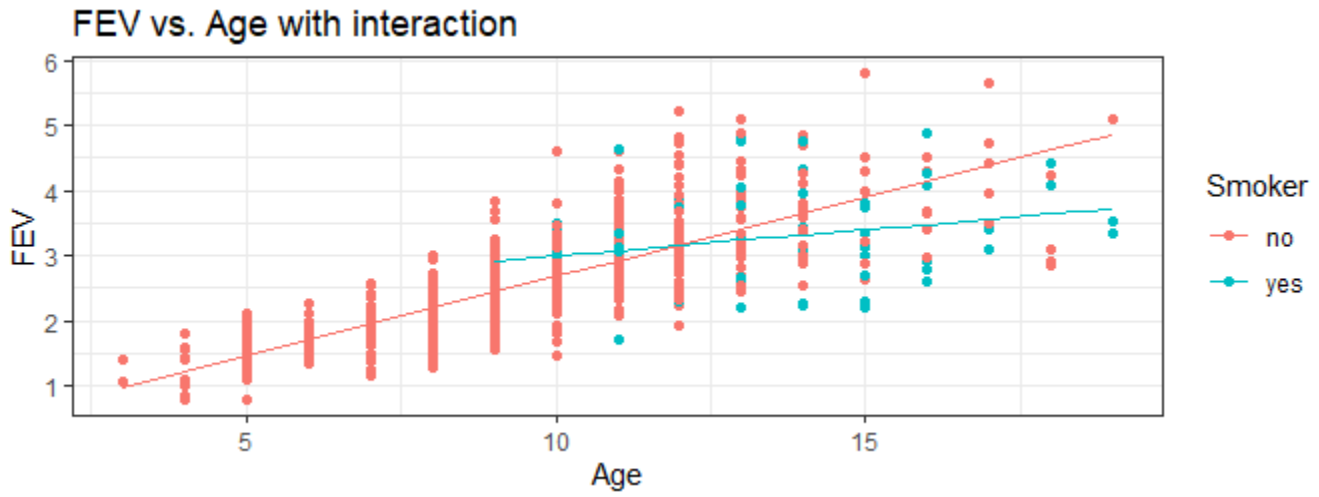
(f) How do we interpret the intercept? How do we interpret the coefficient of Age?

0.25 is predicted FEV for non smoker at age zero. 0.242 is coefficient of age for nonsmokers.

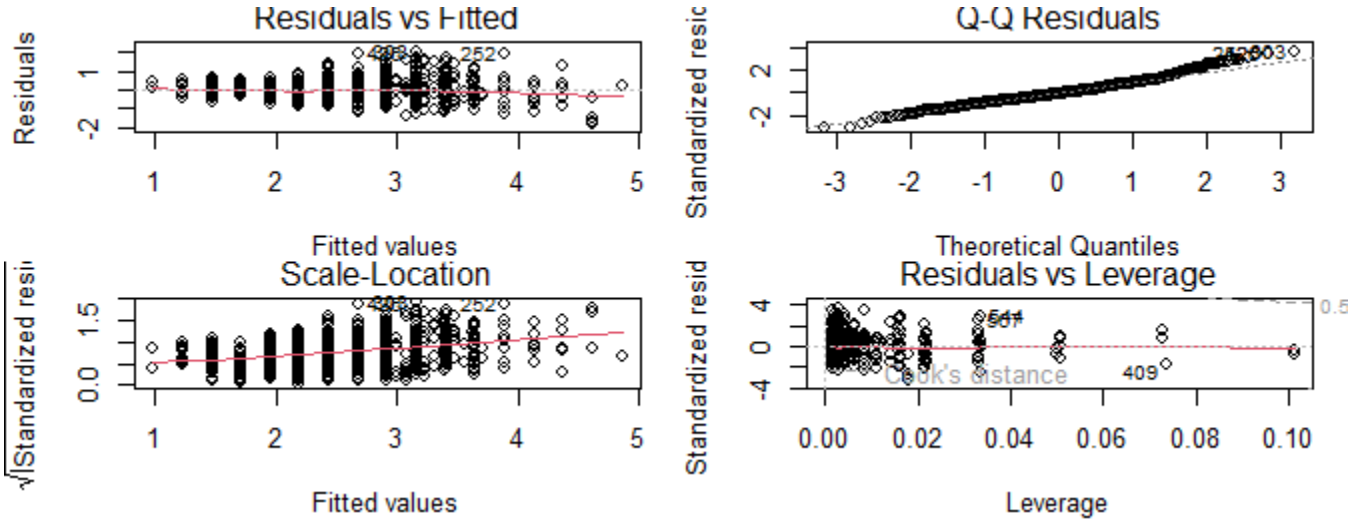
(g) What does the sign of the interaction term tell you? (Note: Another way to interpret this interaction - what is the slope of age in the full equation?)

Negative so the coefficient of age gets lower when smoker changes from 0 (nonsmoker) to 1 (smoker)

```
ggplot(FEVdata,aes(x=Age,y=FEV,color= Smoker))+geom_point()+
  geom_line(aes(y=predict(model3))) +
   labs(title = "FEV vs. Age with interaction") +
  theme_bw()
```



(h) Is this model valid?




```
    Smoker        Age Smoker:Age
     32.77       1.29      34.07
```

(i) Are you surprised there is some multicollinearity? What could we do about it?

We are not surprised because the 'smoker*age' variable is just a subset of the ages in the age variable. Because this is a 'product term' centering the quantitative variable might help!

Because an interaction is a "product," centering the quantitative variable might help with the multicollinearity.

```
#Manually centering the age variable
Age.c = FEVdata$Age - mean(FEVdata$Age)
model4 = lm(FEV ~ Smoker*Age.c, data = FEVdata)

# or
model4 = lm(FEV ~ Smoker*scale(Age, center=TRUE), data = FEVdata)

summary(model4)

Call:
lm(formula = FEV ~ Smoker * scale(Age, center = TRUE), data = FEVdata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7664 -0.3495 -0.0336  0.3368  2.0599

Coefficients:
                                    Estimate Std. Error t value   Pr(>|t|)
(Intercept)                           2.6623     0.0231  115.49    < 2e-16 ***
Smokeryes                             0.3277     0.1286    2.55      0.011 *
scale(Age, center = TRUE)             0.7165     0.0246   29.11    < 2e-16 ***
Smokeryes:scale(Age, center = TRUE)  -0.4806     0.0908   -5.29 0.00000016 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.554 on 650 degrees of freedom
Multiple R-squared:  0.594, Adjusted R-squared:  0.592
F-statistic:  317 on 3 and 650 DF,  p-value: <2e-16
car::vif(model4)
                           Smoker        scale(Age, center = TRUE)
                            3.159                            1.290
Smoker:scale(Age, center = TRUE)
                            3.407
```

(k) Did we improve the multicollinearity?

Yes, the FEV values are smaller

(l) How do we interpret the intercept, coefficient of age, and coefficient of smoker in this model? (Hint: Can you make the interaction go away in order to interpret the main effect?)

intercept: nonsmoker at average age for coefficient; coefficient of age is increase in FEV associated with one-year increase in age for non-smokers; and coefficient of smoker if comparison of smokers and nonsmokers at the average age.

(m) Did adding the interaction help our unequal variance problem? Could it have?

Not a ton, still have some unequal variance across the smokers' line and across the nonsmokers' line…

Notes:

- Indicator variables change intercepts; Interaction terms change slopes.
- *Always* good idea to use graphs to help illustrate an interaction.
- Centering to remove multicollinearity doesn't work on all pairs of variables, just "products" like quadratic and interaction.
- When center with interaction, the interpretation of the "main effect" is about the change in response when the other variable is at its mean (to "zero out" the interaction).

Example 2: Beach data revisited

Recall our Beach data

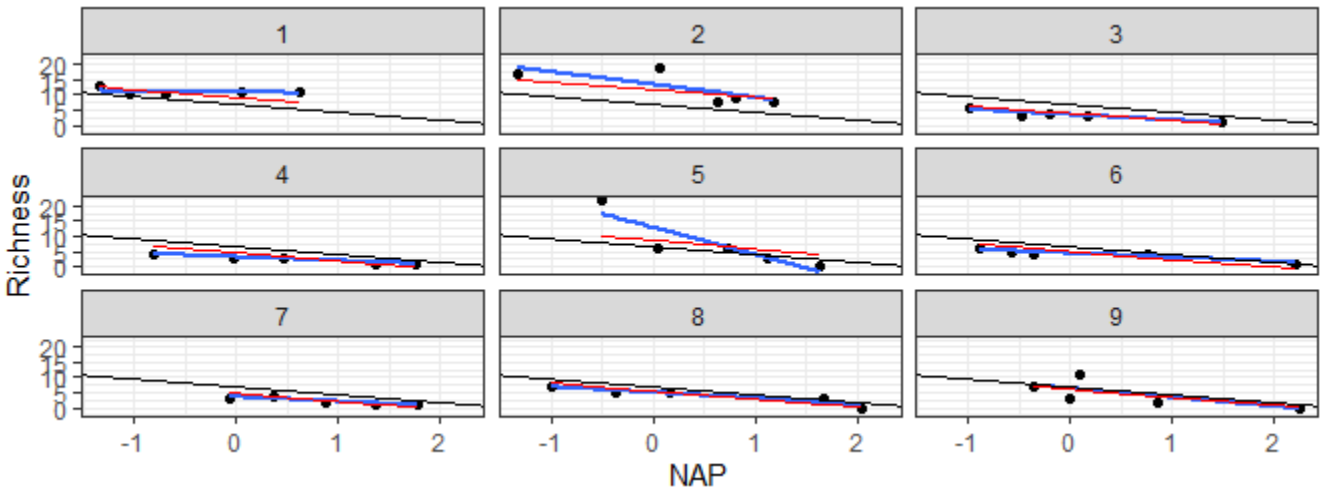
```
rikzdata<-read.table("http://www.rossmanchance.com/stat414/data/RIKZ.txt",header=T)
head(rikzdata)
  Sample Richness Exposure    NAP Beach
1      1       11       10  0.045     1
2      2       10       10 -1.036     1
3      3       13       10 -1.336     1
4      4       11       10  0.616    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

library(lme4)
model1 = lmer(Richness ~ NAP + (1 | Beach), data  = rikzdata)
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
```

We started with a random intercepts model including NAP.

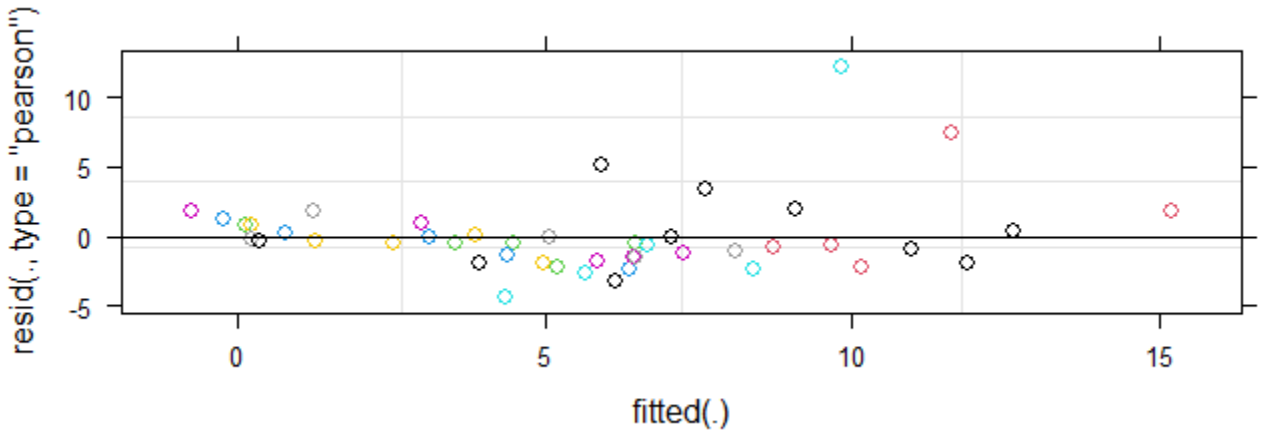
Level 1: $y_{ij} = \beta_{0j} + \beta_{1j}NAP_{ij} + \epsilon_{ij}$

Level 2: $\beta_{0j} = \beta_{00} + u_{0j}$ and $u_{0j} \sim N(0, \tau_0^2)$

```
preds = predict(model1, type = "response")
rikzdata$predictions <- preds
ggplot(rikzdata, aes(x = NAP, y = Richness)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  geom_line(aes(y= predictions), color = "red") +
  facet_wrap(~Beach) +
  theme_bw() +
  geom_abline(intercept = fixef(model1)[[1]], slope=fixef(model1)[[2]])
```




```
plot(model1, col=rikzdata$Beach)
```




```
#plot(residuals(model1) ~ fitted.values(model1), col=rikzdata$Beach)
```

But we saw a possible pattern in the residual plots which suggested we might not want to have the same slope for every beach. In other words, we want an interaction between NAP and Beach…

To allow the slopes to vary across the Level 2 units in the model equation, we add a j index to the slope too.

(a) Write out the level 2 equation for these slopes.

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

(b) What assumption do we want to make about the distribution of the random slopes (effects)?

The u_{1j} are normally distributed with mean 0 and variable $ au_1^2$

(c) Now create the composite equation.

$$y_{ij} = \beta_{00} + \beta_{10}NAP_{ij} + u_{0j} + u_{1j}NAP_{ij} + \epsilon_{ij}$$

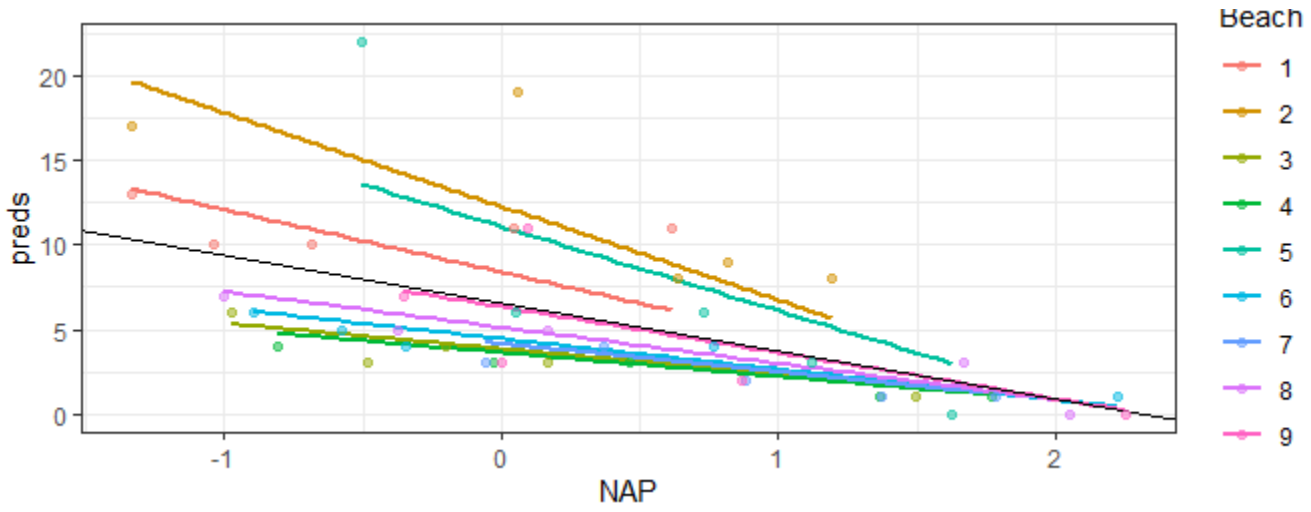
(d) Give the expression for beach j's intercept. Give the expression for beach j's slope.

intercept: $\beta_{0j} = \beta_{00} + u_{0j}$. slope: $\beta_{1j} = \beta_{10} + u_{1j}$

Fit the random slopes (or "random coefficients") model, allowing the slopes to vary across the beaches:

```
model2 = lmer(Richness ~ NAP + (1 + NAP | Beach), data = rikzdata, REML = FALSE)
# you get a warning (not an error) and can ignore it
summary(model2)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: Richness ~ NAP + (1 + NAP | Beach)
   Data: rikzdata

      AIC       BIC    logLik -2*log(L)  df.resid
    246.7     257.5    -117.3     234.7        39

Scaled residuals:
   Min     1Q Median     3Q    Max
-1.798 -0.342 -0.183  0.175  3.139

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Beach    (Intercept) 10.95    3.31
          NAP          2.50    1.58     -1.00
 Residual              7.17    2.68
Number of obs: 45, groups:  Beach, 9

Fixed effects:
            Estimate Std. Error t value
```

```
(Intercept)    6.582      1.188    5.54
NAP           -2.829      0.685   -4.13

Correlation of Fixed Effects:
    (Intr)
NAP -0.810
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
#Our predicted model
preds = predict(model2, newdata = rikzdata)
ggplot(rikzdata, aes(x = NAP , y = preds , group = Beach, color = Beach )) +
geom_smooth(method = "lm", alpha = .5, se = FALSE) +
geom_abline(intercept = 6.58, slope = -2.83) +
geom_point(data = rikzdata, aes(y = Richness, color=Beach), alpha = .5) +
  theme_bw()
```



(e) How many parameters does this add to the model? (What if beach was a fixed effect?) What do these new parameter estimate(s) tell you?

Adds 2 parameters to the model (still better than adding 8 interaction terms). One is the population variance of the slopes, τ_1^2, another is the covariance (or correlation) between the slopes and the intercepts.

```
fixef(model2)
(Intercept)         NAP
      6.582      -2.829
ranef(model2)
$Beach
  (Intercept)     NAP
1      1.7986 -0.8598
2      5.6926 -2.7212
3     -2.7427  1.3111
4     -2.9682  1.4189
5      4.5045 -2.1532
```

```
6     -2.1372  1.0216
7     -2.4399  1.1663
8     -1.4646  0.7001
9     -0.2431  0.1162

with conditional variances for "Beach"
```

(f) Find the estimated equations for Beach 1 and Beach 5. Do they differ as we expected based on our visual inspections of the data?

Beach 1 = 6.58 + 1.80 + (-2.83 - 0.86) NAP and Beach 5 = 6.58 + 4.5 + (-2.83 - 2.15), a much larger intercept and a much steeper slope than Beach 1, consistent with what we saw when we fit the individual regression lines.

(g) Are the differences in the slopes (collectively) statistically significant? (State hypotheses, df, test statistic, p-value, conclusion in context.)

```
#do something here
```

$H_o: \tau_1^2 = \tau_{01} = 0$ vs. $H_a: \tau^2 > 0$, the test statistic is χ^2 = 7.17 with df = 2. The p-value = 0.0277. The variation in the slopes from beach to beach is statistically significant, this would suggest keeping the random slopes in the model.

(h) Does your analysis in the previous question agree with the following output?

```
confint(model2)
              2.5 % 97.5 %
.sig01       1.9594  6.017
.sig02      -1.0000  1.000
.sig03       0.4424  3.399
.sigma       2.0684  3.442
(Intercept)  3.9711  9.179
NAP         -4.4095 -1.356
```

Yes, the confidence interval for sigma3 (0.44, 3.399) does not contain zero (our point estimate was about 1.5)

(i) Using our assumption, about the normal distribution of the population of slopes, between what two values do we expect 95% of the slopes to fall?

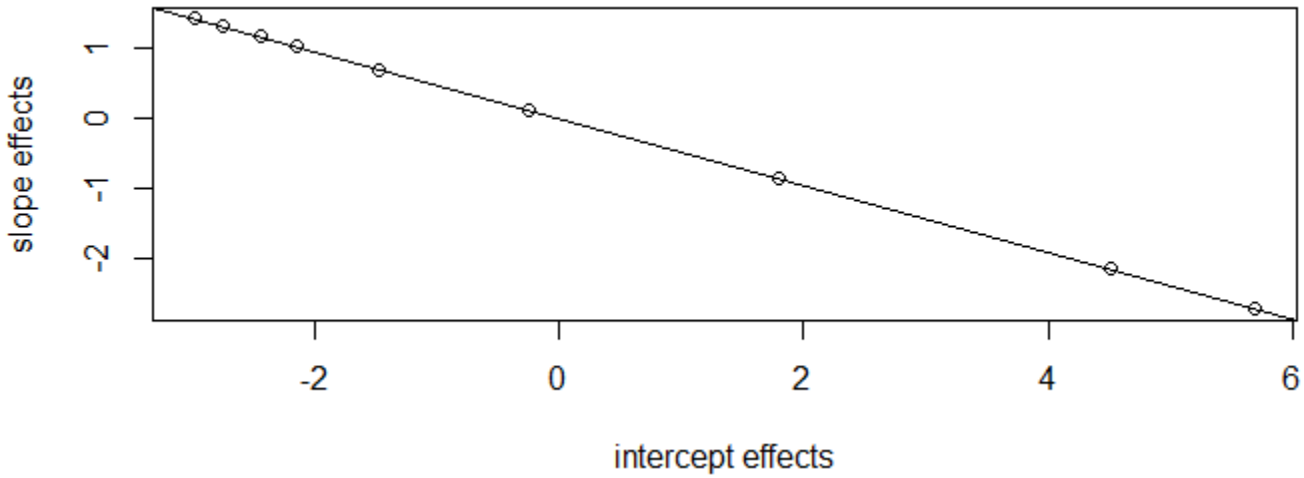
Yes, the confidence interval for sigma3 (0.44, 3.399) does not contain zero (our point estimate was about 1.5)

(j) What's the difference between the intervals in (h) and (i)?

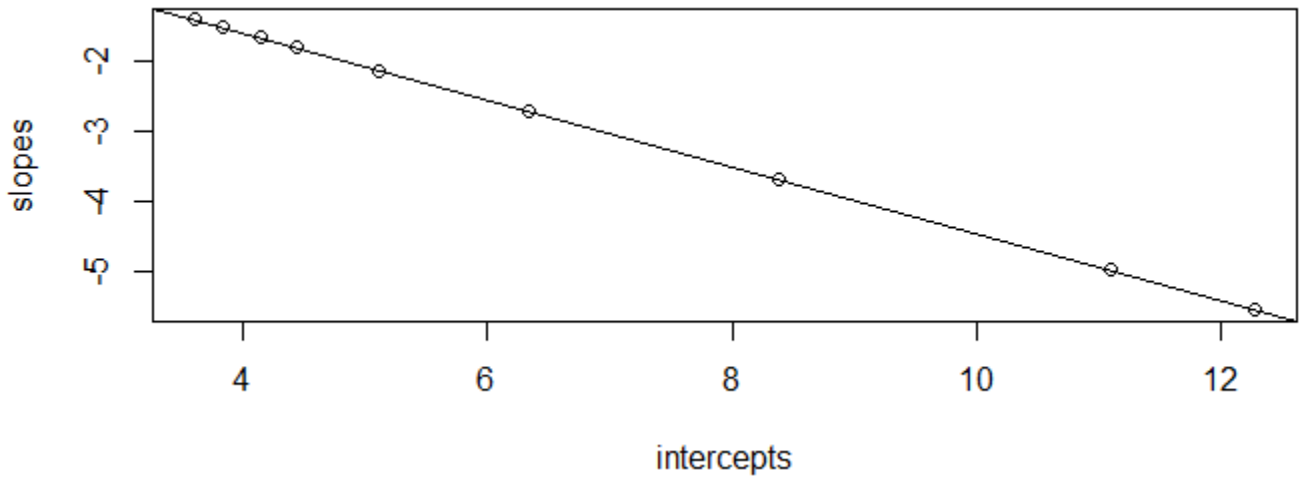
The confidence interval is trying to capture the population value of $ au$ by accounting for the margin of error from random sampling variability. The second interval is genearlizing from the beaches in our sample to the larger population of beaches and what the largest and smallest values of 'beach mean slope + random effect' might plausibly be based on the estimated beach to beach variation in the slopes.

What does it mean for the intercepts and slopes to be correlated?

```
#ranef(model2)$Beach[,1] extracts the estimated random intercepts and ranef(model2)
$Beach[,2] extracts the estimated random slopes

#this plots the "effects" (intercepts vs. slopes)
plot(ranef(model2)$Beach[,2]~ranef(model2)$Beach[,1], ylab = "slope effects", xlab=
"intercept effects")
abline(lm(ranef(model2)$Beach[,2]~ ranef(model2)$Beach[,1]))
```




```
#this plots the resulting slopes and intercepts
slopes= fixef(model2)[[2]]+ranef(model2)$Beach[,2]
intercepts = fixef(model2)[[1]]+ranef(model2)$Beach[,1]
plot(slopes ~ intercepts, xlab = "intercepts", ylab="slopes")
abline(lm(slopes~intercepts))
```



(k) What is the difference between the two graphs we created? How can you tell? What do the graphs tell you (in context).

Yes, the confidence interval for sigma3 (0.44, 3.399) does not contain zero (our point estimate was about 1.5). Bottom line, we learn that beaches with higher intercepts (above average richness for average NAP) tend to have below average = more negative (steeper) slopes (larger impact of NAP on richness). If a beach tends to be lower in Richness overall, the 'effect' of NAP is smaller.

Computer Problem 11: Due noon on Friday

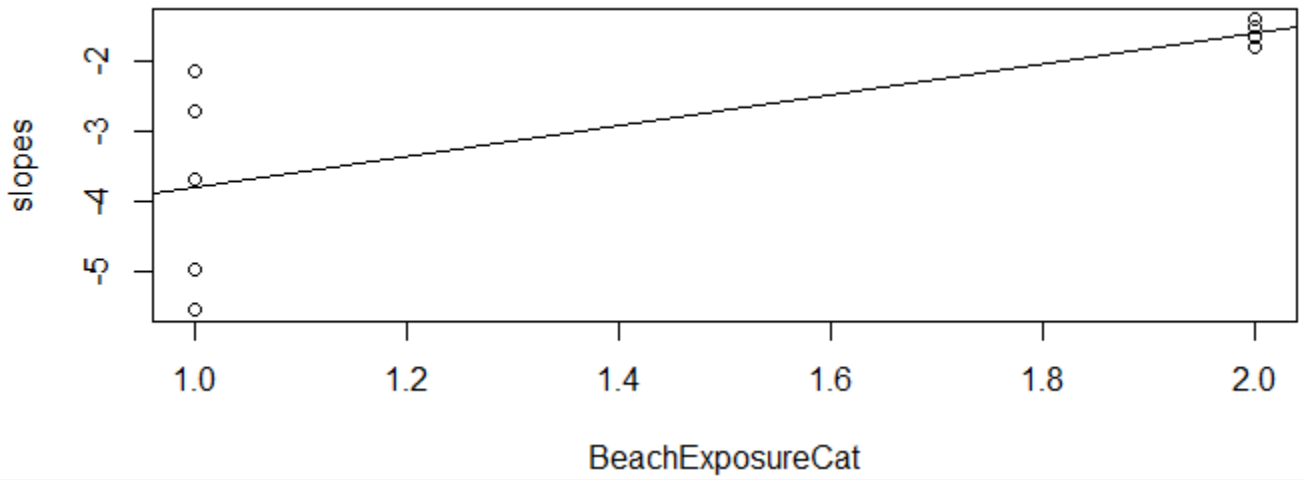
You are encoruaged to work with a partner and turn in one write-up with both names.

Recall we turned Exposure into a binary variable for high vs. low exposure (an index composed of the following elements: wave action, length of the surf zone, slope, grain size, and the depth of the anaerobic layer).

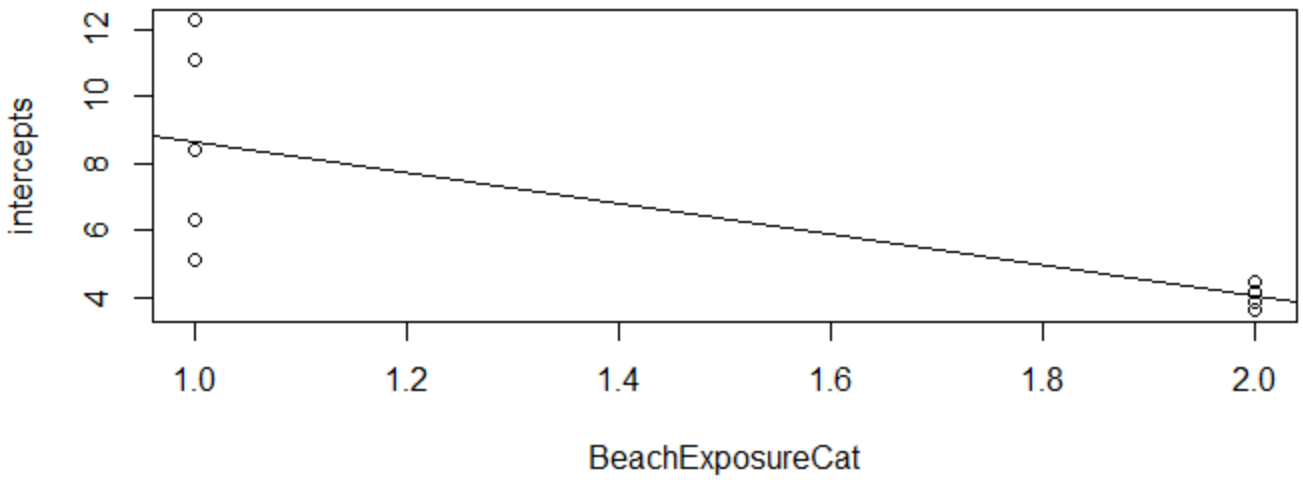
```
rikzdata$ExposureCat = factor(rikzdata$Exposure > 10,
                              levels=c(FALSE, TRUE),
                              labels=c("<=10", ">10")
)
```

Let's consider adding Exposure to our model. Explore whether exposure is related to the intercepts and/or the slopes.

```
BeachExposureCat = tapply(as.numeric(rikzdata$ExposureCat), rikzdata$Beach, mean)
plot(slopes ~ BeachExposureCat); abline(lm(slopes~BeachExposureCat))
```




```
plot(intercepts ~ BeachExposureCat); abline(lm(intercepts~BeachExposureCat))
```

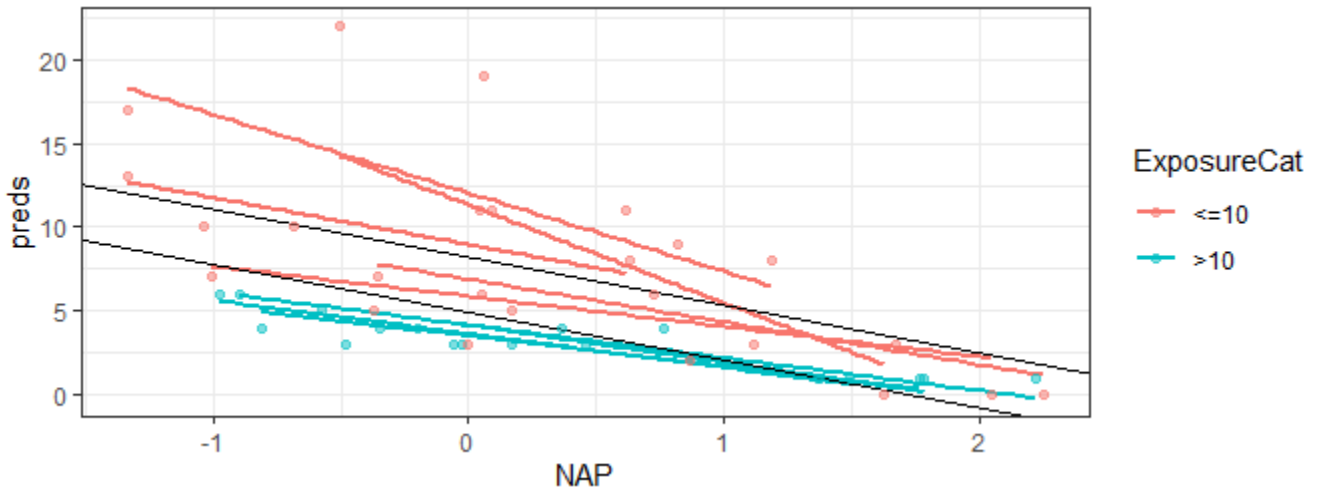


(a) Is Exposure "positively" or "negatively" related to the intercepts? How about the slopes? (Interpret the natures of these associations in context.)

Add ExposureCat to the model

```
par(mfrow=c(1,3))

model3 = lmer(Richness ~ NAP + ExposureCat + (1 + NAP | Beach), data  = rikzdata,
REML=FALSE)
summary(model3)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: Richness ~ NAP + ExposureCat + (1 + NAP | Beach)
   Data: rikzdata

      AIC       BIC    logLik -2*log(L)  df.resid
    245.3     258.0    -115.7     231.3        38

Scaled residuals:
   Min     1Q Median     3Q    Max
-1.928 -0.437 -0.107  0.261  2.974

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Beach    (Intercept) 5.37     2.32
          NAP         2.68     1.64     -0.84
 Residual             6.76     2.60
Number of obs: 45, groups:  Beach, 9

Fixed effects:
               Estimate Std. Error t value
(Intercept)       8.192      1.057    7.75
NAP              -2.852      0.693   -4.11
ExposureCat>10   -3.324      1.280   -2.60

Correlation of Fixed Effects:
```

```
            (Intr) NAP
NAP         -0.571
ExposrCt>10 -0.542 -0.025
preds = predict(model3, newdata = rikzdata)
ggplot(rikzdata, aes(x = NAP , y = preds , group = Beach, color = ExposureCat )) +
geom_smooth(method = "lm", alpha = .5, se = FALSE) +
geom_abline(intercept = 8.19, slope = -2.85) +
geom_abline(intercept = 8.19 - 3.32, slope = -2.85) +
geom_point(data = rikzdata, aes(y = Richness, color=ExposureCat), alpha = .5) +
  theme_bw()
```




```
texreg::screenreg(list(model2, model3), digits = 3, single.row = TRUE, stars = 0,
custom.model.names = c("no exposure", "exposure"), custom.note = "")

=============================================================
                            no exposure       exposure
-------------------------------------------------------------
(Intercept)                   6.582 (1.188)     8.192 (1.057)
NAP                          -2.829 (0.685)    -2.852 (0.693)
ExposureCat>10                                 -3.324 (1.280)
-------------------------------------------------------------
AIC                         246.656           245.335
BIC                         257.496           257.982
Log Likelihood             -117.328          -115.668
Num. obs.                   45                45
Num. groups: Beach           9                 9
Var: Beach (Intercept)      10.949             5.371
Var: Beach NAP               2.502             2.681
Cov: Beach (Intercept) NAP  -5.234            -3.203
Var: Residual                7.174             6.756
=============================================================
anova(model2, model3)
Data: rikzdata
Models:
model2: Richness ~ NAP + (1 + NAP | Beach)
```

```
model3: Richness ~ NAP + ExposureCat + (1 + NAP | Beach)
       npar AIC BIC logLik -2*log(L) Chisq Df Pr(>Chisq)
model2    6 247 258   -117       235
model3    7 245 258   -116       231  3.32  1      0.068 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

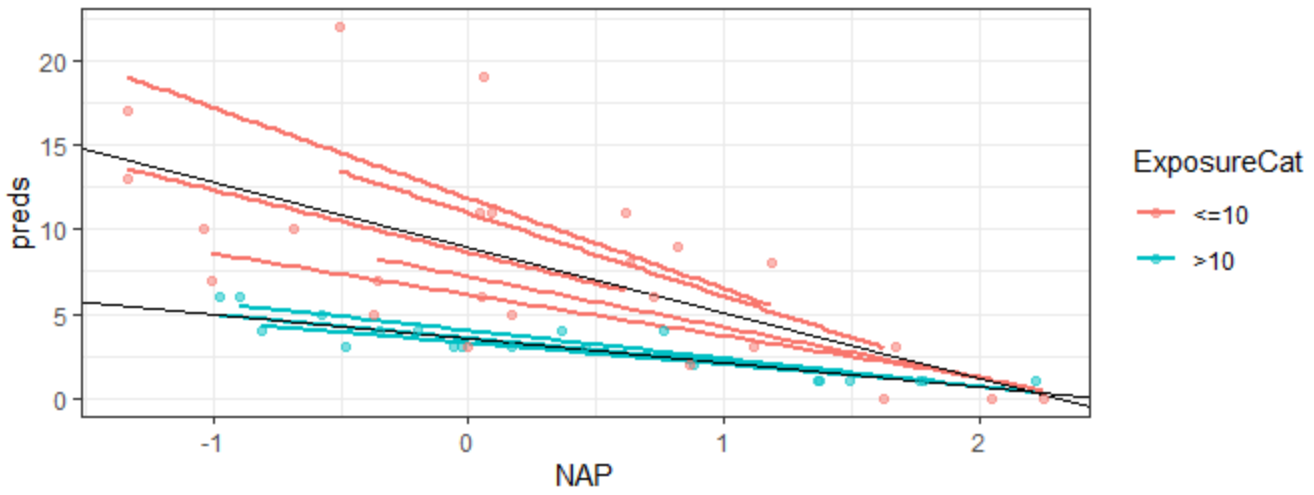
(b) Did adding exposure explain variation in the intercepts? in the slopes? in the within beach residuals?

Add the interaction to the model

```
model4 = lmer(Richness ~ NAP*ExposureCat + (1 + NAP | Beach), data  = rikzdata,  RE
ML=FALSE)  #with interaction
summary(model4, corr = F)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: Richness ~ NAP * ExposureCat + (1 + NAP | Beach)
   Data: rikzdata

      AIC       BIC    logLik -2*log(L)  df.resid
    243.2     257.7    -113.6     227.2        37

Scaled residuals:
   Min     1Q Median     3Q    Max
-1.757 -0.455 -0.158  0.251  3.200

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Beach    (Intercept) 3.83     1.96
          NAP         1.00     1.00     -1.00
 Residual             7.16     2.68
Number of obs: 45, groups:  Beach, 9

Fixed effects:
                   Estimate Std. Error t value
(Intercept)           8.959      1.047    8.55
NAP                  -3.881      0.723   -5.37
ExposureCat>10       -5.382      1.586   -3.39
NAP:ExposureCat>10    2.446      1.099    2.23
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

How do we create the graph now?

```
preds = predict(model4, newdata = rikzdata)
ggplot(rikzdata, aes(x = NAP , y = preds , group = Beach, color = ExposureCat)) +
geom_smooth(method = "lm", alpha = .5, se = FALSE) +
geom_abline(intercept = 8.96, slope = -3.88) +
geom_abline(intercept = 8.96 - 5.38, slope = -3.88 + 2.45) +
geom_point(data = rikzdata, aes(y = Richness, color=ExposureCat), alpha = .5) +
  theme_bw()
```



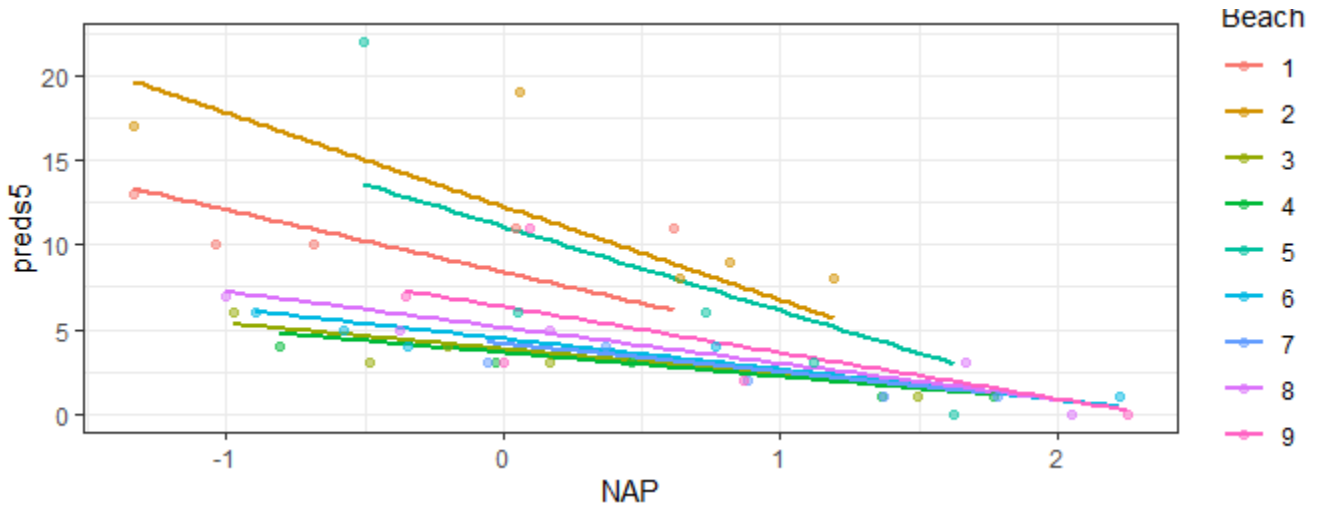

```
texreg::screenreg(list(model3, model4), digits = 3, single.row = TRUE, stars = 0,
custom.model.names = c("exposure", "interaction"), custom.note = "")

=============================================================
                             exposure          interaction
-------------------------------------------------------------
(Intercept)                    8.192 (1.057)     8.959 (1.047)
NAP                           -2.852 (0.693)    -3.881 (0.723)
ExposureCat>10                -3.324 (1.280)    -5.382 (1.586)
NAP:ExposureCat>10                               2.446 (1.099)
-------------------------------------------------------------
AIC                          245.335           243.221
BIC                          257.982           257.674
Log Likelihood              -115.668          -113.611
Num. obs.                     45                45
Num. groups: Beach             9                 9
Var: Beach (Intercept)         5.371             3.832
Var: Beach NAP                 2.681             1.002
Cov: Beach (Intercept) NAP    -3.203            -1.959
Var: Residual                  6.756             7.161
=============================================================
anova(model3, model4)
Data: rikzdata
Models:
model3: Richness ~ NAP + ExposureCat + (1 + NAP | Beach)
model4: Richness ~ NAP * ExposureCat + (1 + NAP | Beach)
       npar AIC BIC logLik -2*log(L) Chisq Df Pr(>Chisq)
model3    7 245 258   -116       231
model4    8 243 258   -114       227  4.11  1      0.043 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c) Describe the nature of the interaction between NAP and Exposure. (whether or not it's significant.)

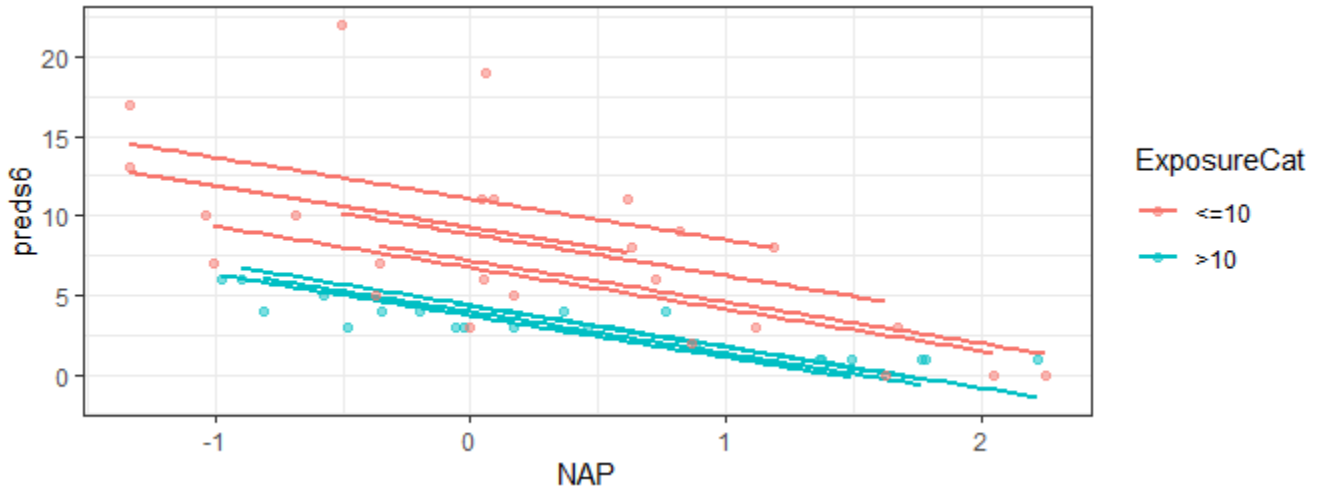
(d) Did adding exposure explain variation in the intercepts? in the slopes? How do you interpret the residual variance increasing?

Compare the following two models

```
model5 = lmer(Richness ~ NAP + (1 + NAP | Beach), data = rikzdata, REML=FALSE)
preds5 = predict(model5, newdata = rikzdata)
ggplot(rikzdata, aes(x = NAP , y = preds5 , group = Beach, color = Beach )) +
geom_smooth(method = "lm", alpha = .5, se = FALSE) +
geom_point(data = rikzdata, aes(y = Richness, color=Beach), alpha = .5) +
theme_bw()
```




```
model6 = lmer(Richness ~ NAP + ExposureCat + (1| Beach), data = rikzdata, REML=FALS
E)
preds6 = predict(model6, newdata = rikzdata)
ggplot(rikzdata, aes(x = NAP , y = preds6 , group = Beach, color = ExposureCat )) +
geom_smooth(method = "lm", alpha = .5, se = FALSE) +
geom_point(data = rikzdata, aes(y = Richness, color=ExposureCat), alpha = .5) +
theme_bw()
```



```
texreg::screenreg(list(model5, model6), digits = 3, single.row = TRUE, stars = 0,
custom.model.names = c("model5", "model6"), custom.note = "")

=============================================================
                            model5            model6
-------------------------------------------------------------
(Intercept)                   6.582 (1.188)     8.608 (0.932)
NAP                          -2.829 (0.685)    -2.604 (0.479)
ExposureCat>10                                 -4.530 (1.383)
-------------------------------------------------------------
AIC                          246.656           244.759
BIC                          257.496           253.792
Log Likelihood              -117.328          -117.379
Num. obs.                     45                45
Num. groups: Beach            9                 9
Var: Beach (Intercept)       10.949             2.419
Var: Beach NAP                2.502
Cov: Beach (Intercept) NAP   -5.234
Var: Residual                 7.174             9.117
=============================================================
anova(model2, model3, model4)
Data: rikzdata
Models:
model2: Richness ~ NAP + (1 + NAP | Beach)
model3: Richness ~ NAP + ExposureCat + (1 + NAP | Beach)
model4: Richness ~ NAP * ExposureCat + (1 + NAP | Beach)
       npar AIC BIC logLik -2*log(L) Chisq Df Pr(>Chisq)
model2    6 247 258   -117       235
model3    7 245 258   -116       231  3.32  1      0.068 .
model4    8 243 258   -114       227  4.11  1      0.043 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(e) Write a few sentences explaining the differences in these two models, what are they assuming/how they are modelling the data. Which model would you recommend and why? (You should consider issues like model fit, parsimony)

Notes

- In a random slopes model, be careful with the interpretation of the intercept variance and the intercept-by-slope covariance, they assume $x = 0$. Another reason why its always good practice to center your explanatory variables so the intercept is meaningful.
- You can use Likelihood-ratio tests to assess whether the random slopes model is “worth pursuing statistically.” As you can see, randomly slopes can improve the complexity of the model pretty quickly, so you should have empirical, or better yet theoretical, reasons for doing so.