Stat 414 - Day 12
Random Slopes (4.6, 5.2)

Last Time

- Interaction terms change slopes
- Interpret "main effects" by "zeroing out" the interaction term.
- Otherwise, coefficient of $x_1 = \hat{\beta}_1 + \hat{\beta}_3 x_2$
- With interactions (and polynomial terms), centering quantitative variables can reduce multicollinearity
- We fit a "random coefficient model" $y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \epsilon_{ij}$ where $\beta_{0j} = \beta_{00} + u_{0j}$ with $u_{0j} \sim N(0, \tau_0^2)$ and $\beta_{1j} = \beta_{10} + u_{1j}$ with $u_{1j} \sim N(0, \tau_1^2)$
- So (assuming) β_{1j} are normally distributed around β_{10}
- This adds a variance component for the slopes (τ_1^2) as well as a covariance between the random slopes and random intercepts (τ_{01})

Example 1: Beaches cont.

Richness varies by Beach, so including Beach in the model (as fixed or random effects) should give us more accurate standard errors.

```
library(lme4)
library(tidyverse)

rikzdata <- read.table("http://www.rossmanchance.com/stat414/data/RIKZ.txt", header
=TRUE)
rikzdata$Beach = factor(rikzdata$Beach)

model0 = lmer(Richness ~ 1 + (1 | Beach), data = rikzdata)
```

(a) Calculate and interpret the ICC value

Code

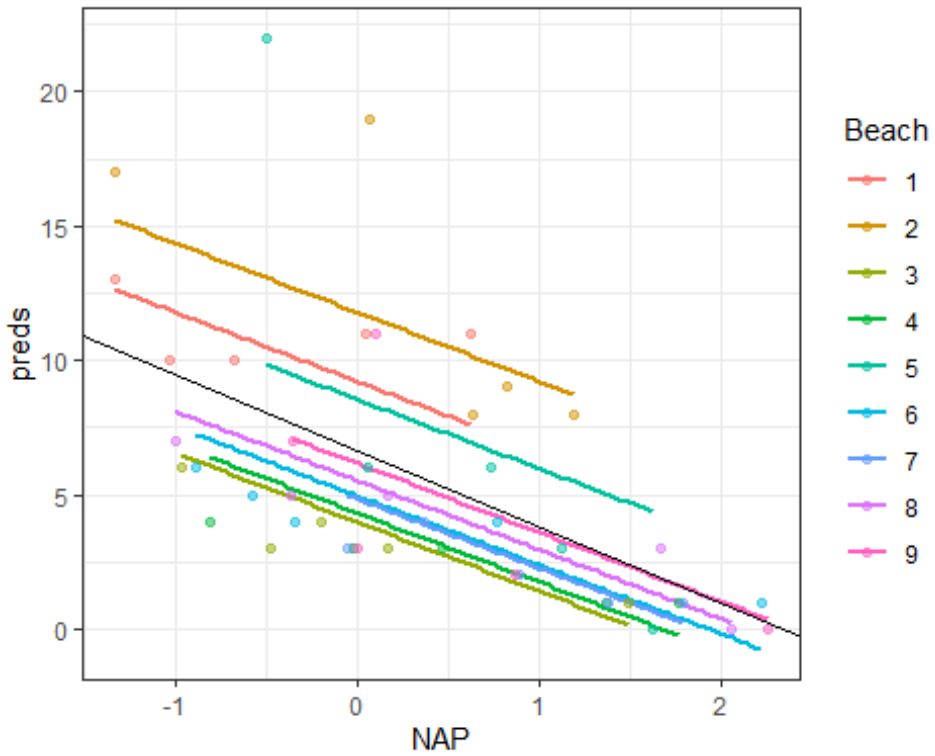
```
performance::icc(model0)
# Intraclass Correlation Coefficient

    Adjusted ICC: 0.403
  Unadjusted ICC: 0.403
```

The expected correlation of two observations in the same beach is 0.403/About 40% of the variation in Richness values is between beaches.

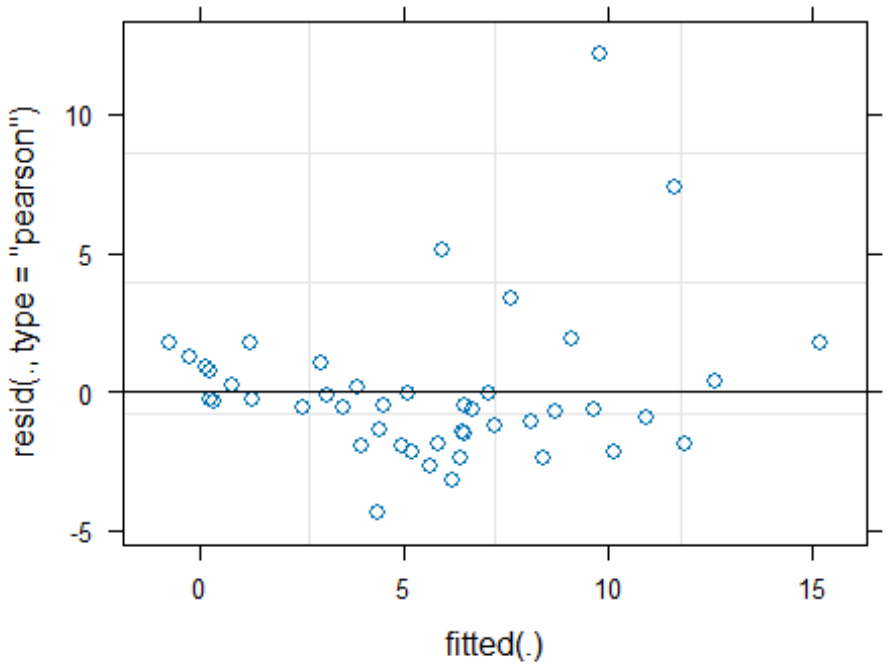
Fit the random intercepts model with NAP

```
summary(model1 <- lmer(Richness ~ NAP + (1 | Beach), data  = rikzdata), 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
#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 = 6.58, slope = -2.83) +
  geom_point(data = rikzdata, aes(y = Richness, color=Beach), alpha = .5) +
  theme_bw()
```




```
plot(model1)
```




```
#performance::check_model(model1)
```

(b) Do the regression model assumptions appear to be met?

We also see evidence of unequal variance.

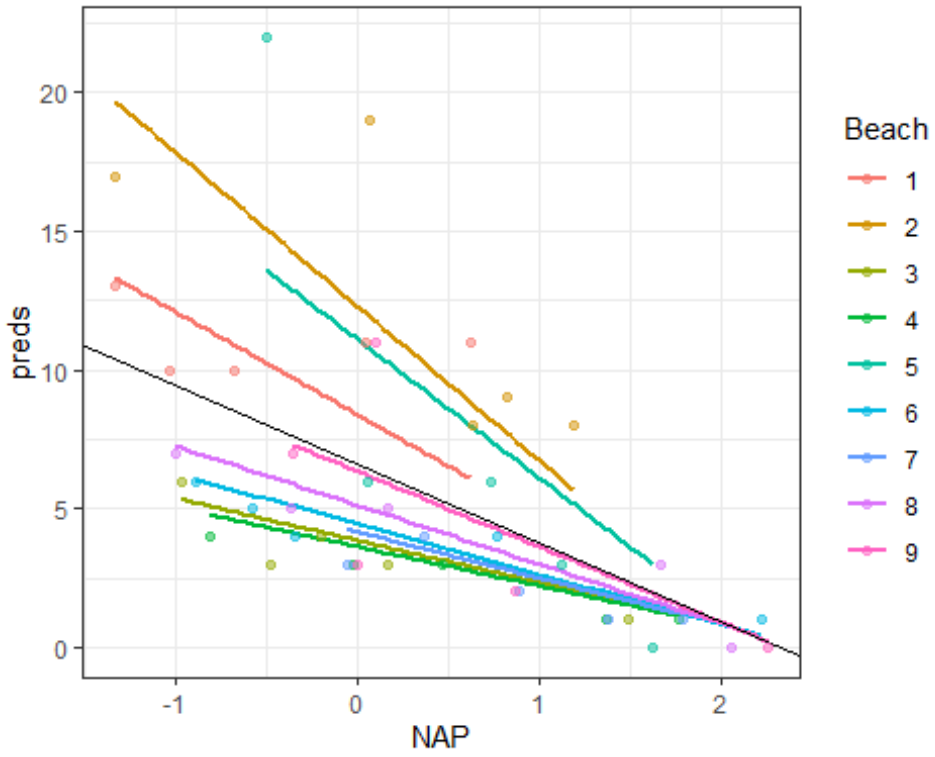
But we see some patterns (including related to the beaches) in the residuals and this tells us that we might be able to improve the fit between the model and the data.

Random Slopes

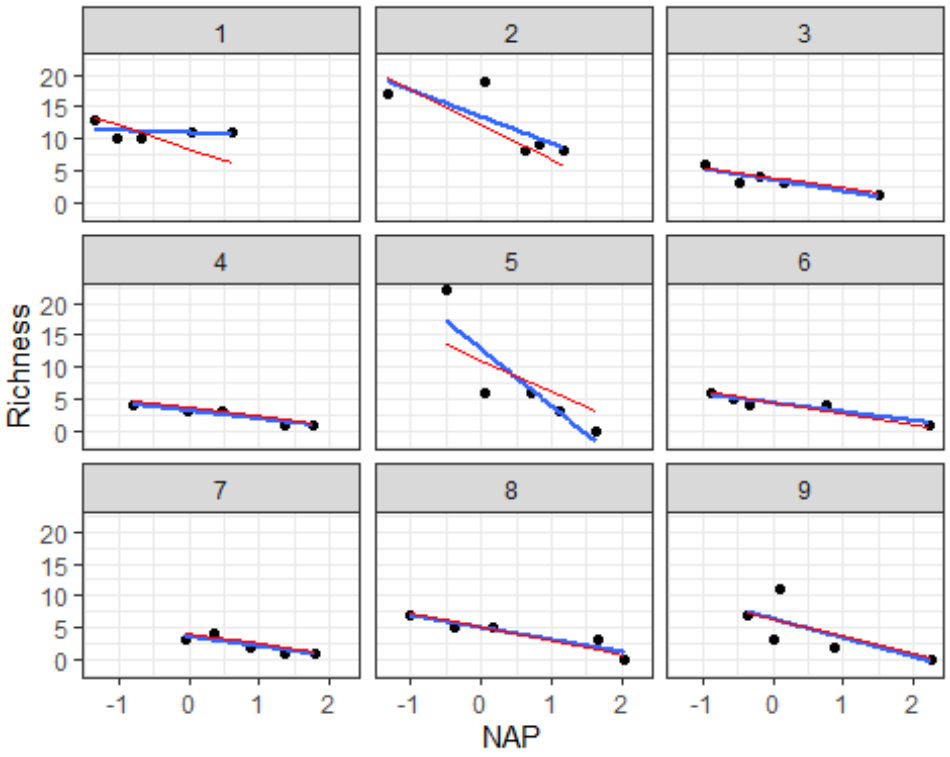
So then we tried a random slopes model (like an interaction between NAP and Beach), and the variation in the slopes ($\hat{\tau}_1^2$) was statistically significant.

```
model2 = lmer(Richness ~ NAP + (1 + NAP | Beach), data = rikzdata, REML = FALSE)
summary(model2, corr=FALSE)
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
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
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"
anova(model1, model2)
Data: rikzdata
Models:
model1: Richness ~ NAP + (1 | Beach)
model2: Richness ~ NAP + (1 + NAP | Beach)
       npar AIC BIC logLik -2*log(L) Chisq Df Pr(>Chisq)
model1    4 250 257   -121       242
model2    6 247 258   -117       235  7.17  2      0.028 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
preds = predict(model2, newdata = rikzdata)
ggplot(rikzdata, aes(x = NAP , y = preds , group = Beach, color = Beach )) +
  geom_smooth(method = "lm", formula = y ~ x, 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()
```




```
ggplot(rikzdata, aes(x = NAP, y = Richness)) +
  geom_point() +
  geom_smooth(method="lm", formula= y ~ x, se=FALSE) +
  geom_line(aes(y= preds), color = "red") +
  facet_wrap(~Beach) +
  theme_bw()
```



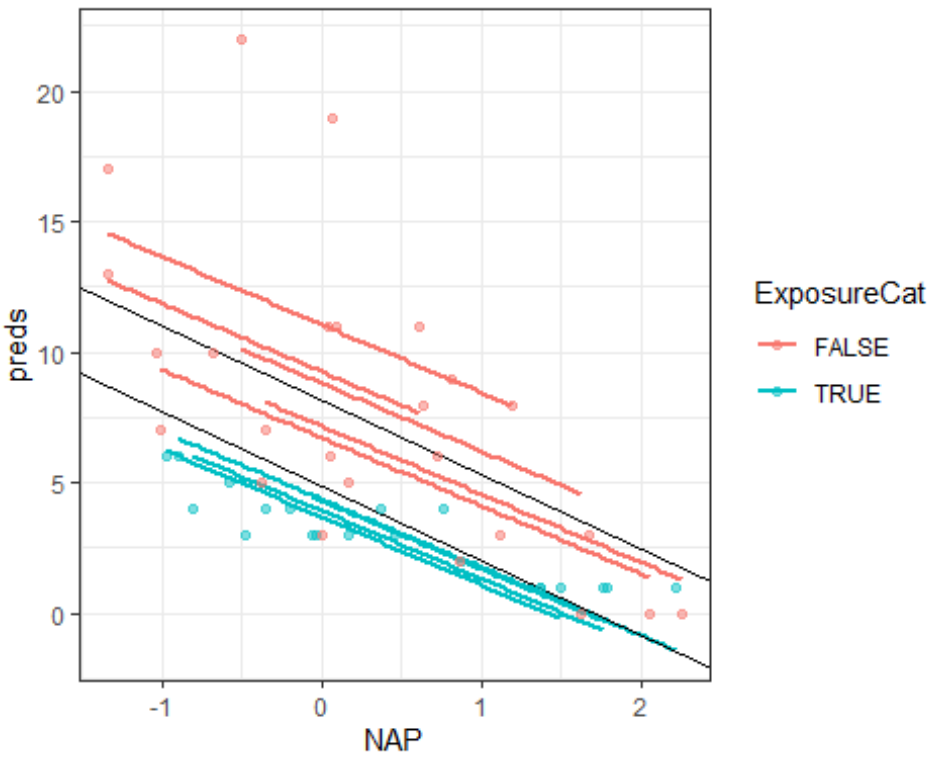
Alternatively

One recommended approach for model selection is to start with all potential fixed effects (including interactions), and then decide on the random effects (e.g., slopes and/or intercepts). Then use that model to pare down the fixed effects.

Fit the random intercepts model including NAP and Exposure.

```
rikzdata$ExposureCat = (rikzdata$Exposure > 10)
model3 = lmer(Richness ~ NAP + ExposureCat + (1  | Beach), data  = rikzdata,  REML=
FALSE)
summary(model3, corr=F)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: Richness ~ NAP + ExposureCat + (1 | Beach)
   Data: rikzdata

      AIC       BIC    logLik -2*log(L)  df.resid
    244.8     253.8    -117.4     234.8        40

Scaled residuals:
   Min     1Q Median     3Q    Max
-1.517 -0.467 -0.085  0.218  3.932

Random effects:
 Groups   Name        Variance Std.Dev.
 Beach    (Intercept) 2.42     1.56
```

```
 Residual             9.12     3.02
Number of obs: 45, groups:  Beach, 9

Fixed effects:
                Estimate Std. Error t value
(Intercept)        8.608      0.932    9.24
NAP               -2.604      0.479   -5.44
ExposureCatTRUE   -4.530      1.383   -3.28
anova(model1, model3)
Data: rikzdata
Models:
model1: Richness ~ NAP + (1 | Beach)
model3: Richness ~ NAP + ExposureCat + (1 | Beach)
       npar AIC BIC logLik -2*log(L) Chisq Df Pr(>Chisq)
model1    4 250 257   -121       242
model3    5 245 254   -117       235  7.07  1     0.0078 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#Our predicted model
preds = predict(model3, newdata = rikzdata)
ggplot(rikzdata, aes(x = NAP , y = preds , group = Beach, color = ExposureCat )) +
geom_smooth(method = "lm", formula = y ~ x, alpha = .5, se = FALSE) +
geom_abline(intercept = 8.1923-3.3238, slope = -2.85) +
geom_abline(intercept = 8.1923, slope = -2.85) +
geom_point(data = rikzdata, aes(y = Richness, color=ExposureCat), alpha = .5) +
  theme_bw()
```



(n) Is the Exposure variable significant? Do you see a substantial improvement in the fit of the model compared to Model 1? How do the variance components change/what has been the main impact?

The ExposureCat variable is significant (The LRT test comparing model 1 and model 3 gives a p-value = .0078). The AIC reduces by about 5 (which is decent). The unexplained variation in the intercepts reduced from 8.668 to 2.419.

Cross-level Interaction

We saw that Exposure was "negatively" related to the intercepts (beaches with high exposure tended to have lower (below average) intercepts (Richness when NAP = 0) than beaches with low exposure) and "positively" related to the slopes (beaches with high exposure tended to have 'above average' or less negative slopes (decrease with each one unit increase in NAP) than beaches with low exposure).

(o) To expand our model to allow for Exposure to explain variation in slopes, write the Level 1 and Level 2 equations, including Exposure in both Level 2 equations.

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

$$\beta_{0j} = \beta_{00} + \beta_{01}Exp_j + u_{0j}$$ Adding the exposure variable to each Level 2 equation

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

(p) Now make the composite equation, what happens?

$$y_{ij} = \beta_{00} + \beta_{01}Exp_j + u_{0j} + (\beta_{10} + \beta_{11}Exp_j + u_{1j})NAP_{ij} + \epsilon_{ij}$$

This creates the 'cross-level interaction' term of $\beta_{11}Exp_j \times NAP_{ij}$ so β_{11} is how much the slope of NAP differs between high and low exposure beaches.

Fit the model

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

      AIC       BIC    logLik -2*log(L)  df.resid
    242.1     253.0    -115.1     230.1        39

Scaled residuals:
   Min     1Q Median     3Q    Max
-1.596 -0.417 -0.085  0.231  3.841

Random effects:
 Groups   Name        Variance Std.Dev.
```

```
 Beach    (Intercept) 2.21     1.49
 Residual             8.21     2.87
Number of obs: 45, groups:  Beach, 9

Fixed effects:
                    Estimate Std. Error t value
(Intercept)            8.870      0.896    9.90
NAP                   -3.492      0.607   -5.76
ExposureCatTRUE       -5.262      1.358   -3.87
NAP:ExposureCatTRUE    2.025      0.915    2.21
texreg::screenreg(list(model3, model4), digits = 3, single.row = TRUE, stars = 0,
custom.model.names = c("exposure",  "interaction"), custom.note = "")

=============================================================
                         exposure          interaction
-------------------------------------------------------------
(Intercept)                8.608 (0.932)     8.870 (0.896)
NAP                       -2.604 (0.479)    -3.492 (0.607)
ExposureCatTRUE           -4.530 (1.383)    -5.262 (1.358)
NAP:ExposureCatTRUE                          2.025 (0.915)
-------------------------------------------------------------
AIC                      244.759           242.114
BIC                      253.792           252.953
Log Likelihood          -117.379          -115.057
Num. obs.                 45                45
Num. groups: Beach         9                 9
Var: Beach (Intercept)     2.419             2.208
Var: Residual              9.117             8.210
=============================================================
anova(model3, model4)
Data: rikzdata
Models:
model3: Richness ~ NAP + ExposureCat + (1 | Beach)
model4: Richness ~ NAP * ExposureCat + (1 | Beach)
       npar AIC BIC logLik -2*log(L) Chisq Df Pr(>Chisq)
model3    5 245 254   -117       235
model4    6 242 253   -115       230  4.65  1      0.031 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(q) How many parameters did we add to the model? What is the estimate for that parameter? Is it statistically significant? How are you deciding?

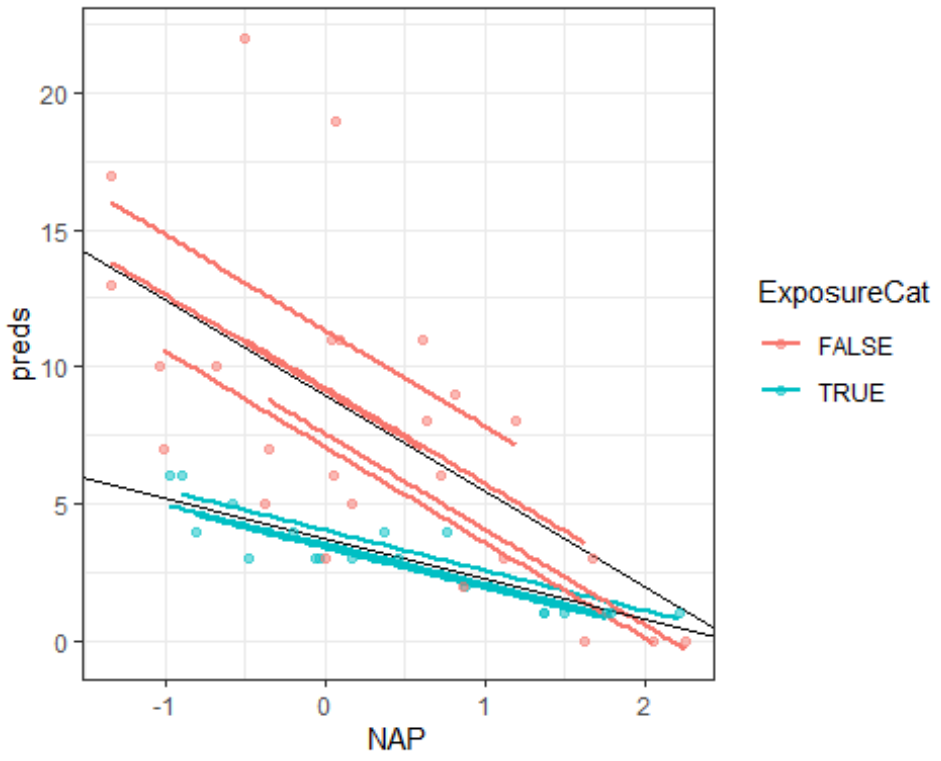
This just adds one parameter to the model β_{10}. The LRT test for this coefficient gives a p-value of .03, so it's moderately significant.

(r) In particular, what is the overall intercept and the overall slope for low exposure beaches, and the overall intercept and the overall slope for high exposure beaches?

Use these values to create a new graph:

Code

```
preds = predict(model4, newdata = rikzdata)
ggplot(rikzdata, aes(x = NAP , y = preds , group = Beach, color = ExposureCat ))
+
geom_smooth(method = "lm", formula = y ~ x, alpha = .5, se = FALSE) +
geom_abline(intercept = 8.9695, slope = -3.49) +
geom_abline(intercept = 8.9695 -5.2625, slope = -3.49 + 2.025) +
geom_point(data = rikzdata, aes(y = Richness, color=ExposureCat), alpha = .5) +
  theme_bw()
```



(s) Do we still have significant random variation in the slopes?

```
model5 = lmer(Richness ~ NAP*ExposureCat + (1 + NAP | Beach), data  = rikzdata,  RE
ML=FALSE)
summary(model5, 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
ExposureCatTRUE       -5.382      1.586   -3.39
NAP:ExposureCatTRUE    2.446      1.099    2.23
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
anova(model4, model5)
Data: rikzdata
Models:
model4: Richness ~ NAP * ExposureCat + (1 | Beach)
model5: Richness ~ NAP * ExposureCat + (1 + NAP | Beach)
       npar AIC BIC logLik -2*log(L) Chisq Df Pr(>Chisq)
model4    6 242 253   -115       230
model5    8 243 258   -114       227  2.89  2       0.24
```

Tthe p-value for the likelihood ratio test is not small (0.24), so we don't need to add random slopes to the model that already has the cross-level interaction.

Notes

- "In cases where the explanation of the random effects works extremely well, one may end up with models with no random effects at level two… random intercepts, slope have zero variance.. Omitted.. The resulting model may be analyzed just as well with OLS regression analysis… within group dependence has been fully explained by the available explanatory variables/interactions (no more dependence in the residuals)."

Properties of random slopes model

But there might be another reason to use random slopes… Rerun model 1 using lme instead of lmer.

```
#install.packages("nlme")
library(nlme)
model1B = lme(Richness ~ NAP, random = ~ 1 | Beach, data = rikzdata)
summary(model1B)  #can't suppress the correlation of fixed effects output
Linear mixed-effects model fit by REML
  Data: rikzdata
    AIC   BIC logLik
  247.5 254.5 -119.7

Random effects:
 Formula: ~1 | Beach
```

```
        (Intercept) Residual
StdDev:       2.944     3.06

Fixed effects:  Richness ~ NAP
             Value Std.Error DF t-value p-value
(Intercept)  6.582    1.0958 35   6.007       0
NAP         -2.568    0.4947 35  -5.192       0
 Correlation:
    (Intr)
NAP -0.157

Standardized Within-Group Residuals:
    Min      Q1     Med      Q3     Max
-1.4227 -0.4848 -0.1576  0.2519  3.9794

Number of Observations: 45
Number of Groups: 9
```

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(model1B, type = "marginal", individual = "1"); vcm
Beach 1
Marginal variance covariance matrix
       1      2      3      4      5
1 18.030  8.668  8.668  8.668  8.668 = τ̂01
2  8.668 18.030  8.668  8.668  8.668
3  8.668  8.668 18.030  8.668  8.668
4  8.668  8.668  8.668 18.030  8.668
5  8.668  8.668  8.668  8.668 18.030
  Standard Deviations: 4.246 4.246 4.246 4.246 4.246
cov2cor(vcm[[1]])
       1      2      3      4      5
1 1.0000 0.4807 0.4807 0.4807 0.4807 = ICC = 8.688/√(18.03 × (18.03))
2 0.4807 1.0000 0.4807 0.4807 0.4807
3 0.4807 0.4807 1.0000 0.4807 0.4807
4 0.4807 0.4807 0.4807 1.0000 0.4807
5 0.4807 0.4807 0.4807 0.4807 1.0000
#What are these again?
getVarCov(model1B, type = "conditional")
Beach 1
Conditional variance covariance matrix
      1     2     3     4     5
1 9.362 0.000 0.000 0.000 0.000
2 0.000 9.362 0.000 0.000 0.000
3 0.000 0.000 9.362 0.000 0.000
4 0.000 0.000 0.000 9.362 0.000
5 0.000 0.000 0.000 0.000 9.362
  Standard Deviations: 3.06 3.06 3.06 3.06 3.06
getVarCov(model1B)
```

```
Random effects variance covariance matrix
            (Intercept)
(Intercept)       8.668  τ̂₀²
  Standard Deviations: 2.944
```

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

The variances for each observation in the beach (if we were to keep measuring that site). We are assuming those measurements are the same for each site and for each beach $= 2.944^2 + 3.06^2 = 18.03$. The off diagonal values are the covariances of two observations in the same beach.

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

These are now the correlations.
corr(sitei, sitej) = cov(sitei, sitej)/(SD(sitei)SD(sitej)
8.6675/18.03 = 0.48, the interclass correlation coefficient after adjusting for NAP (between two observations on the same beach with same NAP)

Now let's look at the random coefficients (add the random slopes) model (with lme):

```
summary(model2B <- lme(Richness ~ NAP, random = ~ 1 + NAP | Beach, data = rikzdata)
)
Linear mixed-effects model fit by REML
  Data: rikzdata
    AIC BIC logLik
  244.4 255 -116.2

Random effects:
 Formula: ~1 + NAP | Beach
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev Corr
(Intercept) 3.549  (Intr)
NAP         1.715  -0.99
Residual    2.703

Fixed effects:  Richness ~ NAP
             Value Std.Error DF t-value p-value
(Intercept)  6.589    1.2648 35   5.209  0.0000
NAP         -2.830    0.7229 35  -3.915  0.0004
 Correlation:
    (Intr)
NAP -0.819

Standardized Within-Group Residuals:
    Min      Q1     Med      Q3     Max
-1.8213 -0.3411 -0.1675  0.1921  3.0397

Number of Observations: 45
Number of Groups: 9
```

```
getVarCov(model2B)
Random effects variance covariance matrix
            (Intercept)    NAP
(Intercept)      12.596 -6.027
NAP              -6.027  2.941
  Standard Deviations: 3.549 1.715
vcm2 <- getVarCov(model2B, type = "marginal", individual = "1")
vcm2
Beach 1
Marginal variance covariance matrix
       1     2     3      4     5
1 19.365 18.43 20.20  8.694 16.36
2 18.431 35.55 30.96 13.250 25.05
3 20.200 30.96 41.25 14.515 27.46
4  8.694 13.25 14.52 13.592 11.77
5 16.356 25.05 27.46 11.766 29.52
  Standard Deviations: 4.401 5.962 6.423 3.687 5.433
```

(f) What changes about the matrix? Good news or bad news?

We are no longer assuming the variances are the same across the sites (or between the beaches). This could model the unequal variance we saw at the very beginning.

(g) According to the model, which site(s) in Beach 1 have larger variance?

sites 2 and 3

Examine the data for the 5 observations for beach 1:

```
head(rikzdata, 5)
  Sample Richness Exposure    NAP Beach ExposureCat
1      1       11       10  0.045     1       FALSE
2      2       10       10 -1.036     1       FALSE
3      3       13       10 -1.336     1       FALSE
4      4       11       10  0.616     1       FALSE
5      5       10       10 -0.684     1       FALSE
```

(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?

Sites 2 and 3 have the most negative NAP values. Sites 1 and 4 have the positive (higher) NAP values and the least variability.

Correlations for random slopes model

```
cov2cor(vcm2[[1]])
       1      2      3      4      5
1 1.0000 0.7025 0.7147 0.5359 0.6841
2 0.7025 1.0000 0.8085 0.6028 0.7732
3 0.7147 0.8085 1.0000 0.6130 0.7868
```

```
4 0.5359 0.6028 0.6130 1.0000 0.5874
5 0.6841 0.7732 0.7868 0.5874 1.0000
```

(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?)

Now the correlation of Richness values between a pair of sites within the same beach depends on which two sites you look at

The point is that a random slopes model also allows us to model heteroscedasticity 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 $\tau_0^2 + x_{ij}^2\tau_1^2 + 2x_{ij}\tau_{01} + \sigma^2$

(j) so is minimized at x_{ij} =

$$(-1)\tau_{01}/\tau_1^2$$

(k) What does τ_{01} represent? What is the estimate for this model?

This was the covariance between the intercepts and the slopes. The lmer output gives us the correlation which we can convert -.99 x 3.549 x 1.715. Or we can use getVarCov(model2B), -6.026

(l) 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?)

(-1) x -6.026/2.9411 = 2.05. It makes sense that this NAP value is 'just off the graph' as the lines are fanning in for the NAP values we have in our data.

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 shown 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?

cov = $\hat{\tau}_0^2$ = 3.549^2 = 12.59; correlation divides this by the total variance: $\tau_0^2/(\tau_0^2 + \sigma^2)$ - the ICC

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 talked about "variability explained."