ENV 710: Lecture 12

linear models 3

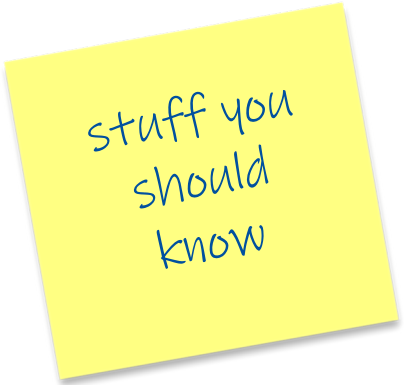
Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

linear models

multiple explanatory variables

learning goals

- conduct linear models with multiple explanatory variables, both continuous and categorical
- what are partial regression coefficients?
- how to interpret the intercept and coefficients?
- how to predict from multiple linear regression?
- what are the assumptions of multiple linear regression?
  - multicollinearity and variance inflation factors
  - ratio of 1:10, predictors to data



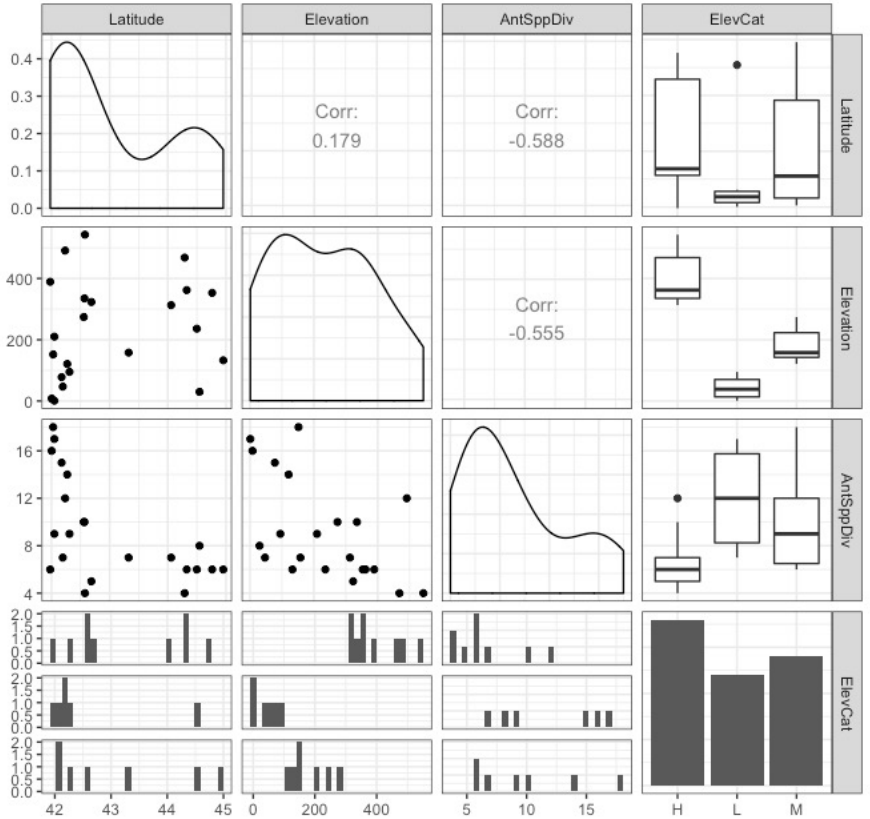
assumptions

- independence of error terms
- homoscedasticity (constant variance) of the errors
- normality of the distribution of errors
- explanatory variables are fixed
- no perfect multicollinearity

example ants and bogs

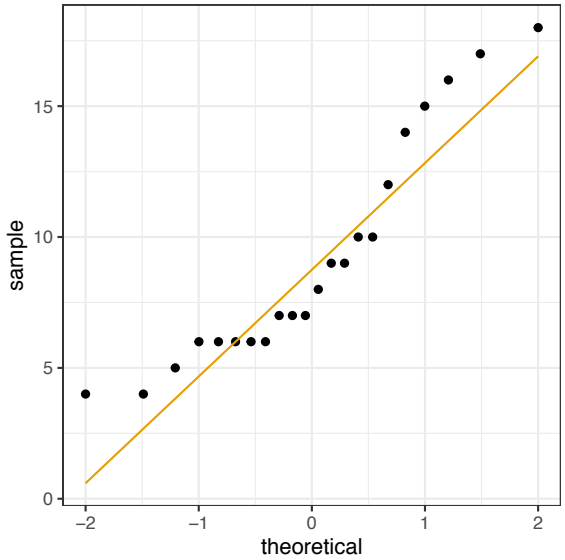
In a study of variation in species richness of ants in New England bogs and forests, Gotelli and Ellison (2012) measured the latitude and elevation of each sampling site.

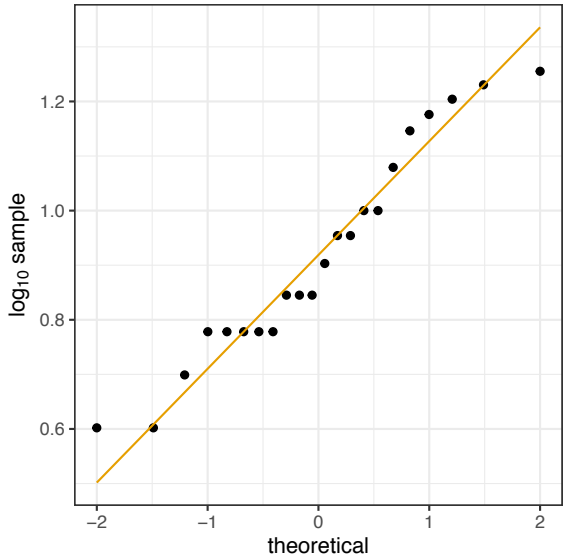
Do latitude and/or elevation influence ant species richness? What is the relationship between each explanatory variable and species richness?



example

ants and bogs





example ants and bogs

1. download the data.

```
dat <- read.csv("AntsBogsForest.csv", header = T)
```

2. use $\log_{10}$ to transform AntSppDiv for both analyses

3. run a simple linear model to assess the effect of elevation (Elevation) on ant species diversity (AntSppDiv)

4. repeat #3 for the effect of latitude (Latitude) on ant species diversity

5. write two statements that describe the effects of latitude and elevation on ant species diversity

lm - latitude

```
bog.lm <- lm(log10(AntSppDiv) ~ Latitude, data = dat)
 summary(bog.lm)
```

```
Call:lm(formula = log10(AntSppDiv) ~ Latitude, data = dat)

Coefficients:
             Estimate    Std. Error t value   Pr(>|t|)
(Intercept)  5.44734    1.43077     3.807    0.00110 **
Latitude    -0.10525    0.03325    -3.165    0.00486 **

Residual standard error: 0.1648 on 20 degrees of freedom
Multiple R-squared:  0.3338,Adjusted R-squared:  0.3005
F-statistic: 10.02 on 1 and 20 DF,  p-value: 0.004864
```

$$log10(ants) = 5.447 - 0.105 \cdot latitude$$

$$ants = 10^{(5.447 - 0.105 \cdot latitude)}$$

$$10^{-0.105} = 0.785$$

Latitude is statistically significant with a -0.105 decrease in $\log_{10}$(diversity), or a 21.5% decrease in diversity, with each degree of latitude.

linear models in R

do both Latitude and Elevation affect ant species diversity?

- continuous response variable and two or more continuous explanatory variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

observed predictor
explanatory variable
covariate
independent variable

estimate $\beta_0, \beta_1, \beta_2, \cdots$ by:

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k$$

linear models in R

$$log10(ants) = b_0 + b_1 \cdot latitude + b_2 \cdot elevation$$

```
bog.lm3 <- lm(log10(AntSppDiv) ~ Latitude + Elevation,
              data = dat)
```

```
Call:lm(formula = log10(AntSppDiv) ~ Latitude + Elevation)

Coefficients:
             Estimate  Std. Error  t value    Pr(>|t|)
(Intercept)  4.8795119  1.1738285    4.157     0.000535 ***
Latitude    -0.0887320  0.0274364   -3.234     0.004367 **
Elevation   -0.0006129  0.0001820   -3.367     0.003234 **
---
Residual standard error: 0.1338 on 19 degrees of freedom
Multiple R-squared: 0.5828,   Adjusted R-squared: 0.5389
F-statistic: 13.27 on 2 and 19 DF,  p-value: 0.0002474
```

$$H_0 : \beta_1 = \beta_2 = ... = 0$$

$$H_a : \text{Not } \beta_j \neq 0, \ j = 1, 2, ...k$$

	model coefficients		
	Elevation	Latitude	Elevation & Latitude
Intercept	1.087	5.447	4.879
Elevation	-0.0007		-0.0006
Latitude		-0.105	-0.089
Adj. R^2	0.321	0.301	0.539

partial regression coefficient

- MLR coefficients are called partial regression coefficients
- amount by which the dependent variable (DV) changes when one independent variable (IV) is changed by one unit, holding all the other independent variables constant

- coefficient is called partial because its value depends, in general, upon the other independent variables

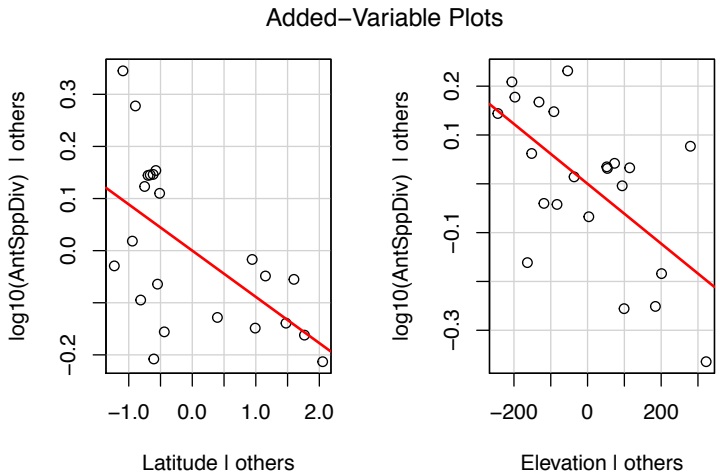
	model coefficients		
	Elevation	Latitude	Elevation & Latitude
Intercept	1.087	5.447	4.879
Elevation	-0.0007		-0.0006
Latitude		-0.105	-0.089
Adj. R^2	0.321	0.301	0.539

partial regression plot

- also called an added-variable plot
- demonstrates the contributions of one variable, while holding all other variables constant
- plots a certain type of residual on the *Y*-axis and a certain type of residual on the *X*-axis

`car package`

```
avPlots(bog.lm3, id.cex = 0.7, id.n=2)
```

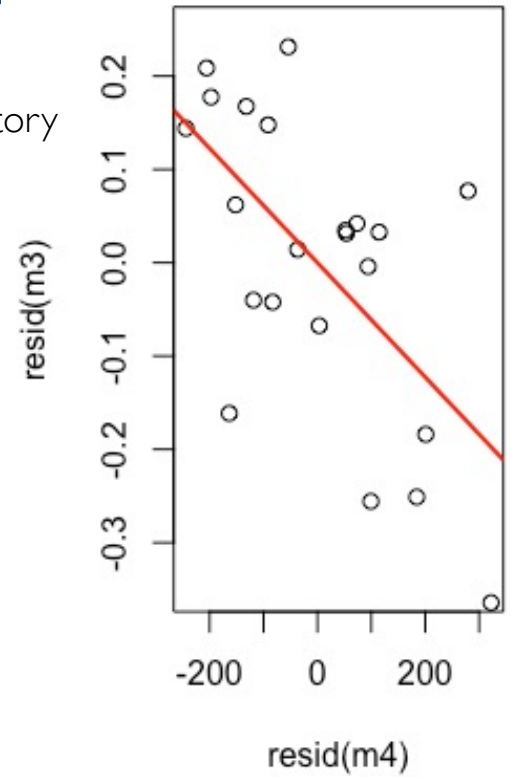


partial regression plot - latitude

- in a two-predictor model, with dependent variable, *Y*, and explanatory variables, X_1 and X_2, the partial regression plot for X_2 gives us the information on X_2 after X_1 has already been included

```
m1 <- lm(log10(AntSppDiv) ~ Elevation)
m2 <- lm(Latitude ~ Elevation)

plot(x=resid(m1), y=resid(m2), ylim = c(-0.35, 0.35))
abline(lm(resid(m1) ~ resid(m2)), lwd = 2)
```

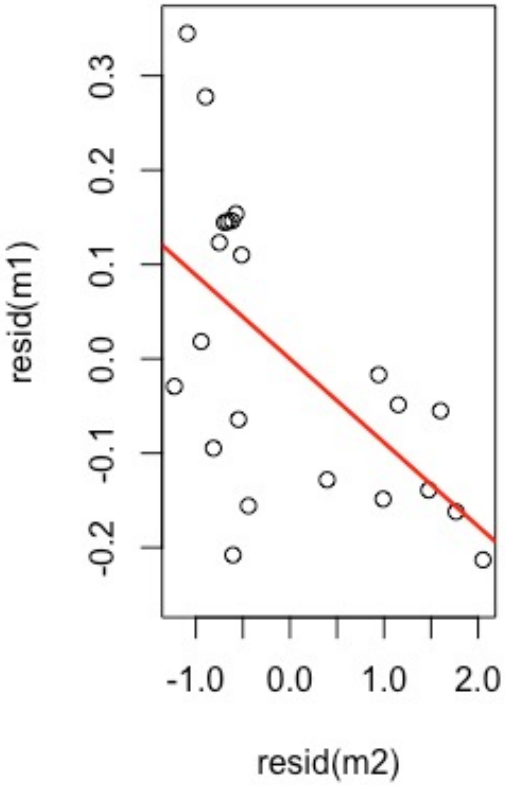


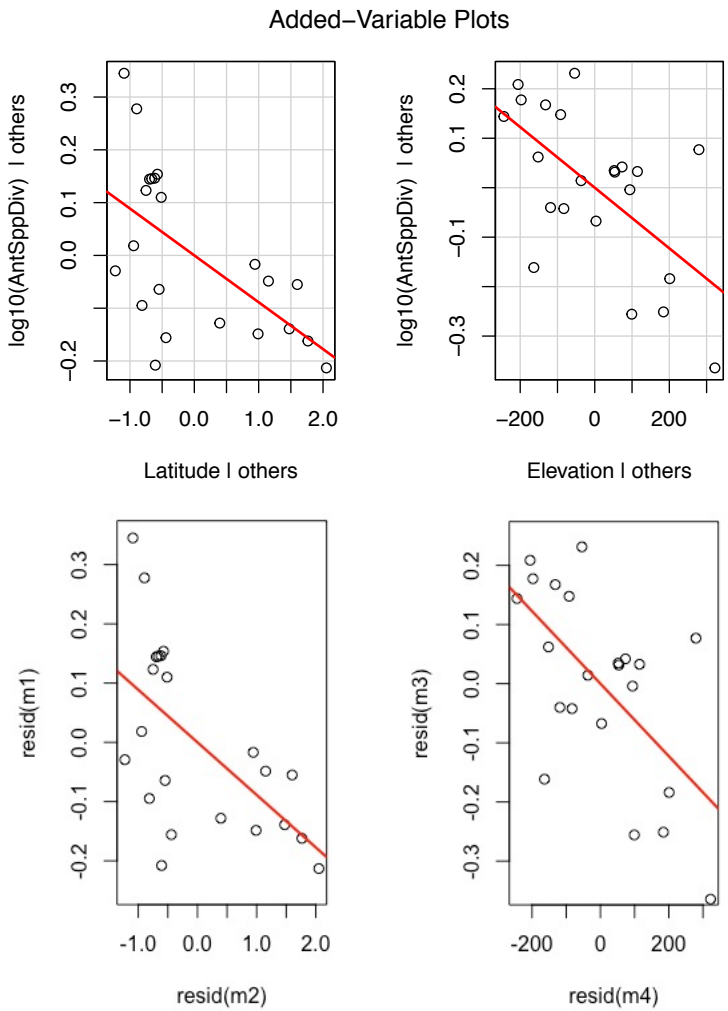
partial regression plot - elev.

- in a two-predictor model, with dependent variable, *Y*, and explanato variables, X_1 and X_2, the partial regression plot for X_2 gives us the information on X_2 after X_1 has already been included

```
m3 <- lm(log10(AntSppDiv) ~ Latitude)
m4 <- lm(Elevation ~ Latitude)

plot(x=resid(m4), y=resid(m3), ylim=c(-0.35, 0.35))
abline(lm(resid(m3) ~ resid(m4)), lwd = 2)
```




Added−Variable Plots
log10(AntSppDiv) | others
Latitude | others
log10(AntSppDiv) | others
Elevation | others
resid(m1)
resid(m2)
resid(m3)
resid(m4)

partial regression coefficient – latitude

variation explained by latitude after taking elevation into account

```
summary(lm(resid(m1) ~ resid(m2)))

Call:
lm(formula = resid(m1) ~ resid(m2))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.717e-18  2.781e-02   0.000  1.00000
resid(m2)   -8.873e-02  2.674e-02  -3.318  0.00343 **
---
Residual standard error: 0.1304 on 20 degrees of freedom
Multiple R-squared:  0.355,   Adjusted R-squared:  0.3228
F-statistic: 11.01 on 1 and 20 DF,  p-value: 0.003431
```

"given all other variables in the model, is this still a useful predictor?"

partial regression coefficient – elev.

variation explained by elevation after taking latitude into account

```
summary(lm(resid(m3) ~ resid(m4)))
```

```
Call:
lm(formula = resid(m3) ~ resid(m4))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.714e-18  2.781e-02   0.000   1.0000
resid(m4)   -6.129e-04  1.774e-04  -3.455   0.0025 **
---
Residual standard error: 0.1304 on 20 degrees of freedom
Multiple R-squared:  0.3738,  Adjusted R-squared:  0.3425
F-statistic: 11.94 on 1 and 20 DF,  p-value: 0.002503
```

"given all other variables in the model, is this still a useful predictor?"

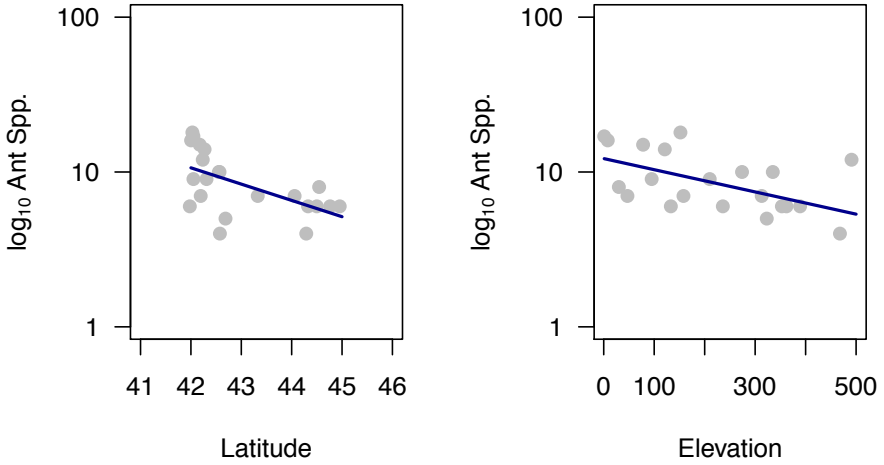
multiple linear regression

```
lm3 <- lm(log10(AntSppDiv) ~ Latitude + Elevation, data = dat)
```

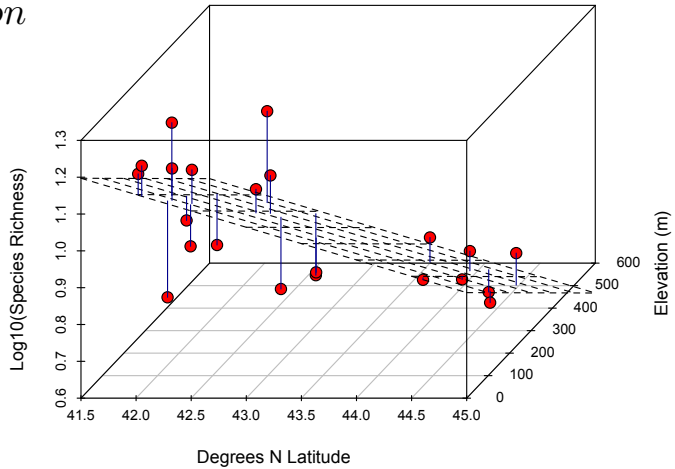
```
Coefficients:
             Estimate  Std. Error  t value    Pr(>|t|)
(Intercept)  4.8795119  1.1738285    4.157    0.000535 ***
Latitude    -0.0887320  0.0274364   -3.234    0.004367 **
Elevation   -0.0006129  0.0001820   -3.367    0.003234 **
---
Residual standard error: 0.1338 on 19 degrees of freedom
Multiple R-squared: 0.5828,   Adjusted R-squared: 0.5389
F-statistic: 13.27 on 2 and 19 DF,  p-value: 0.0002474
```

regression coefficients in MLR are partial regression coefficients

multiple linear regression



$$log10(ants) = b_0 + b_1 \cdot latitude + b_2 \cdot elevation$$



multicollinearity

multicollinearity is strong correlation among independent variables

- model fits data (high *F*-statistic & R^2), but none of the *X*'s have a statistically significant effect on *Y*
- inconsistency in "significant" parameters
- inflates standard errors, making some variables statistically non-significant when they would be significant

how much correlation is too much?

- lots of different opinions…
- more than 0.70 is one rule of thumb
- for correlations > 0.70, you should:
  - remove one of the correlated predictors (usually the one with the lowest *r* with the DV)
  - combine the correlated predictors (average, sum, multivariate combination, etc.…)

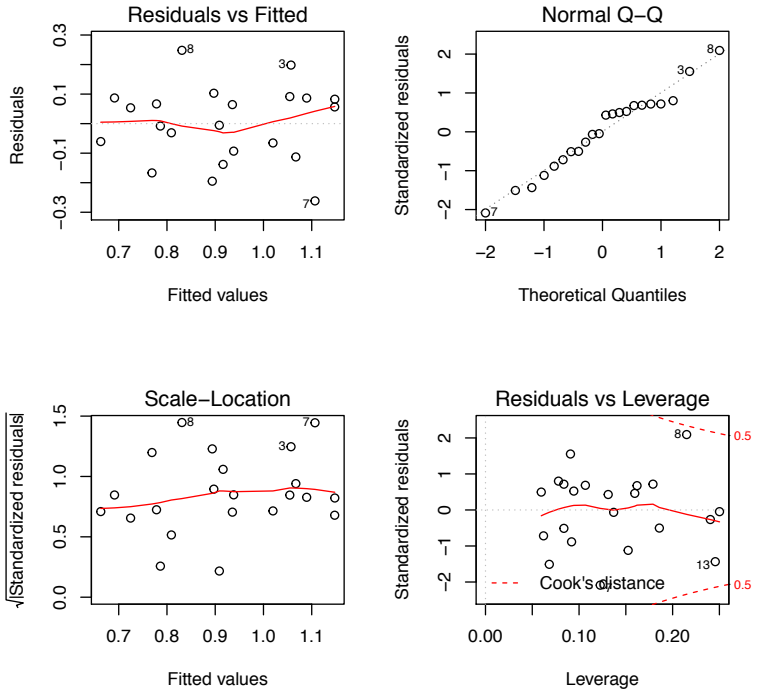
variance inflation factor (VIF)

- calculates how much the variance of an estimated regression coefficient, *j*, is increased because of collinearity

$$VIF_j = \frac{1}{1 - R_j^2}$$

- the higher the VIF, the greater the multicollinearity
  - >10 is too much! high VIF ranges from 5-10
  - indicates that the variance of the jth regression coefficient is 10 times greater than it would have been if the jth independent variable had been independent
  - can use `vif()` from the `car` package, e.g. `vif(lm3)`

```
vif(lm3)
Latitude Elevation
1.033004  1.033004
```



other assumptions

- independence of error terms
- homoscedasticity (constant variance) of the errors
- normality of the distribution of errors
- explanatory variables are fixed
- no perfect multicollinearity

how many observations do we need?

- subject to predictor ratio of 10:1 or 15:1 (even better)
- e.g. with 100 data points, you should test 10 parameters or less…

prediction

how many ant species are found at 41°N and 50 m elevation?

$$log10(y) = 4.8795 - 0.0887 \cdot latitude - 0.0006 \cdot elevation$$

$$y = 10^{(4.8795 - 0.0887 \cdot 41 - 0.0006 \cdot 50)}$$

```
pred <- predict(bog.lm3, list(Latitude = 41, Elevation = 50),
                         interval = "conf")
10^pred
                         fit       lwr       upr
                       16.2501 11.8329 22.3162
```

```
L <- 41
E <- 50

10^(coef(bog.lm3)[1]+coef(bog.lm3)[2]*L + coef(bog.lm3)[3]*E)

                                (Intercept)
                                  16.2501
```

nominal variables

let's modify the ant diversity example, this time treating Elevation as a three-level factor, low, medium, and high, rather than a continuous variable

the variable, ElevCat, includes categories of L, M, and H which correspond to low, medium and high elevations

Low = 1-100 m
Mid = 101 – 300 m
High > 301 m

continuous & categorical IVs

- let's test both *Latitude* and ElevCat at the same time, with elevation coded as a categorical variable

```
lm10 <- lm(log10(AntSppDiv) ~ Latitude + factor(ElevCat), data = dat3)
```

```
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  4.72294    1.31740   3.585    0.00212 **
Latitude    -0.08628    0.03093   -2.789   0.01211 *
ElevH       -0.19351    0.08071   -2.398   0.02756 *
ElevM       -0.03928    0.08356   -0.470   0.64393

Residual standard error: 0.147 on 18 degrees of freedom
Multiple R-squared: 0.5228, Adjusted R-squared: 0.4433
F-statistic: 6.573 on 3 and 18 DF,  p-value: 0.003414
```

- intercept represents the mean for the response when continuous explanatory variables take on the value of 0
  - ant species diversity at a Latitude of 0 and "low" Elevation

continuous & categorical IVs

what is ant species diversity at high elevation and 43 degrees?

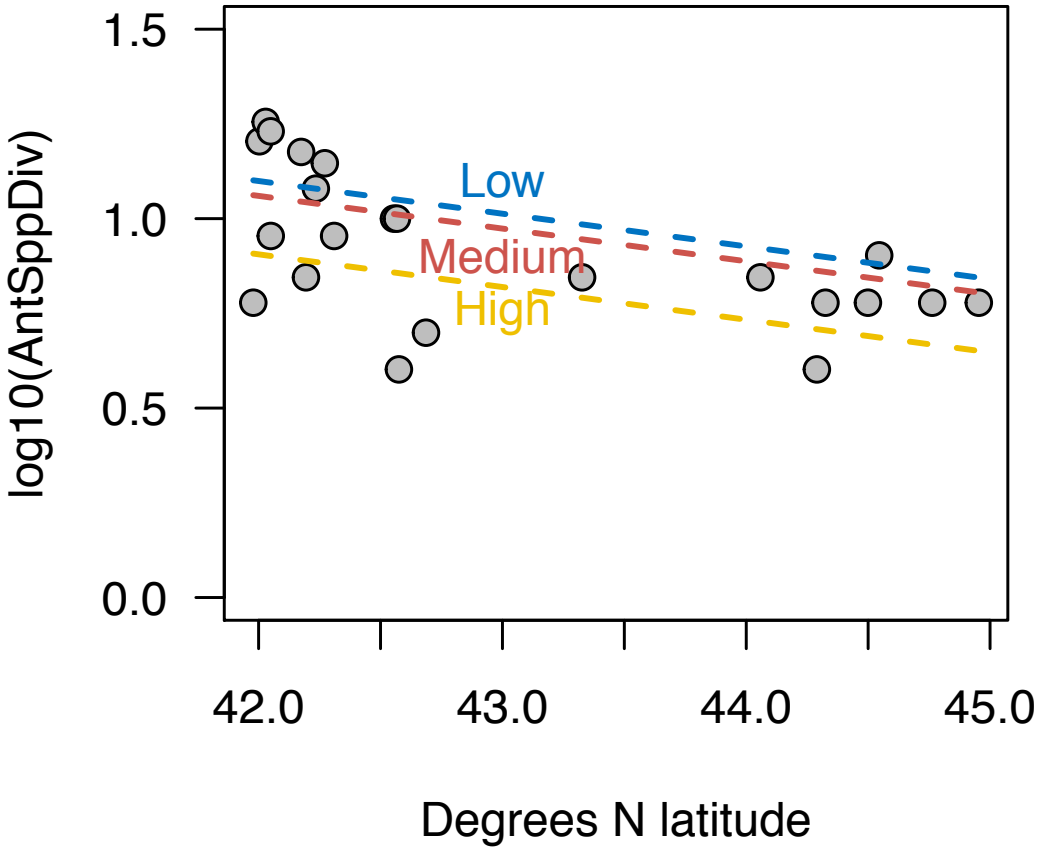
$$y = a + b1 \cdot x_{lat} + b2 \cdot x_{med} + b3 \cdot x_{high}$$

$$\log_{10}(y) = 4.72294 - 0.08628 \cdot 43 - 0.19351 \cdot 1 - 0.03928 \cdot 0 = 0.819$$

```
Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.72294    1.31740   3.585  0.00212 **
Latitude    -0.08628    0.03093  -2.789  0.01211 *
ElevH       -0.19351    0.08071  -2.398  0.02756 *
ElevM       -0.03928    0.08356  -0.470  0.64393

Residual standard error: 0.147 on 18 degrees of freedom
Multiple R-squared: 0.5228, Adjusted R-squared: 0.4433
F-statistic: 6.573 on 3 and 18 DF,  p-value: 0.003414
```

continuous & categorical IVs



Questions?

model comparison

```
anova(bog.lm, bog.lm3)
```

```
Analysis of Variance Table

Model 1: log10(AntSppDiv) ~ Latitude
Model 2: log10(AntSppDiv) ~ Latitude + Elevation
    Res.Df  RSS    Df  Sum of Sq    F    Pr(>F)
 1    20   0.54327
 2    19   0.34022  1   0.20306   11.34  0.003234 **
```

- ANOVA can be used to compare a model with fewer predictors (reduced model) to a model with more predictors (full model)
- comparison of different models can be used to infer whether a predictor variable should be included in a model
- input order of variables doesn't matter because regression coefficients estimated after including all other variables

Code slide 34

```
jcoPalette <- c("#0073C2FF", "#EFC000FF", "#868686FF","#CD534CFF”)
par(mfrow=c(1,1))

x<-seq(min(mydat$Latitude, max(mydat$Latitude), length=30))
coef10<-coefficients(bog.lm5)

with(mydat, plot(Latitude, log10(AntSppDiv), ylim=c(0, 1.5),
     las=1, pch=21, cex=1.2, xlab="Degrees N latitude",
     bg="grey"))
 curve(coef10[1]+coef10[2]*x+coef10[3], add=T, lty=2, lwd=2,
       col=jcoPalette[2])
 curve(coef10[1]+coef10[2]*x+coef10[4], add=T, lty=2,
       col=jcoPalette[4], lwd=2)
 curve(coef10[1]+coef10[2]*x+coef10[3]*0+coef10[4]*0, add=T,
       lty=2, col=jcoPalette[1], lwd=2)

text(43, 1.1, "Low", col = jcoPalette[1])
text(43, 0.9, "Medium", col = jcoPalette[4])
text(43, 0.75, "High", col = jcoPalette[2])
```