# Stat 3301 Final Project

Jessica Li.12986

2024-11-22

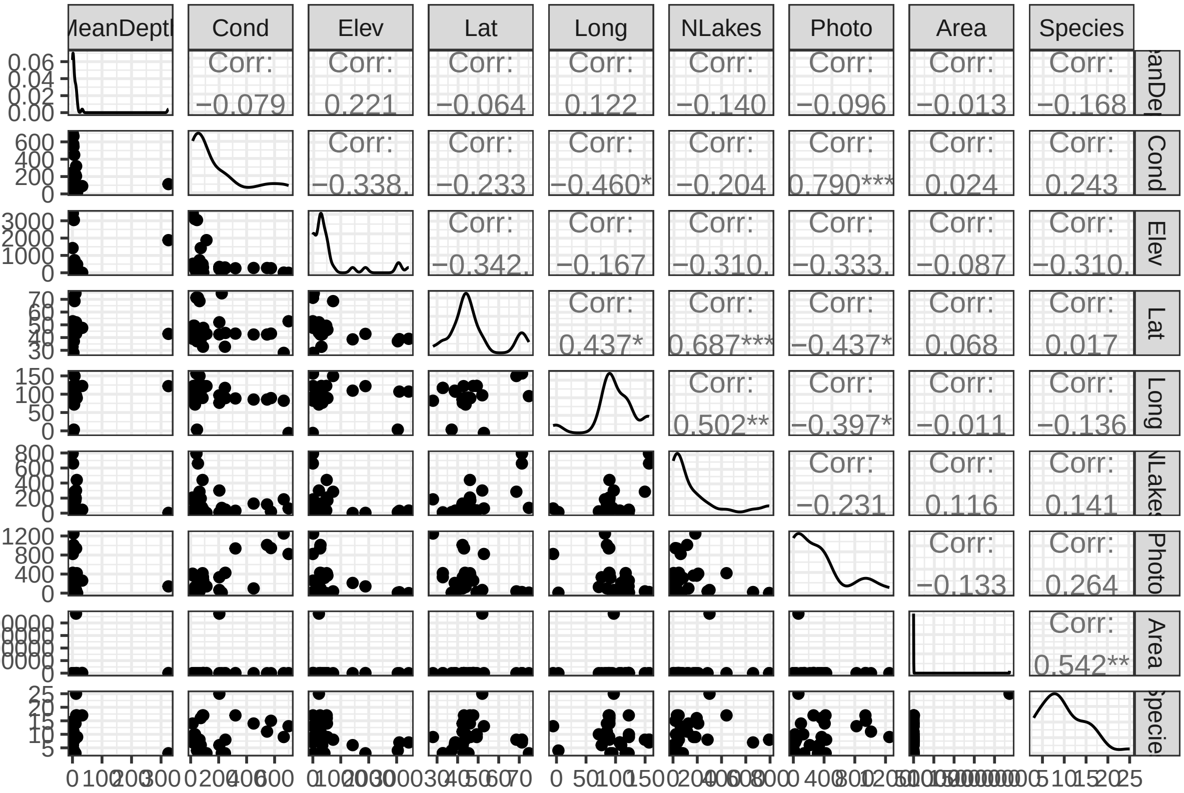
## Introduction

Scientists are worried that we may be experiencing a mass extinction event due to growing concerns about the misuse of the planet's resources. Particularly, they are worried about declining biodiversity. Scientists are interested in the biodiversity of lake crustaceans, a species that is small and difficult to access, making its biodiversity especially difficult to measure. This report summarizes a statistical model that uses lake conditions to predict the number of crustacean species in new lakes.

## Exploratory Data Analysis

The data consists of 30 rows of different lakes, each with the number of crustacean species in them. Each lake also has data on its mean depth in meters, the conductance (measure of mineral content), the lake elevation in meters, the latitude, longitude, number of lakes within 20 kilometers, rate of photosynthesis measured using $C^{14}$, and the surface area of the lake in hectares. The variable names are Species, MeanDepth, Elevation, Cond, NLakes, Area, and Photo, respectively. Since we are interested in predicting the number of crustaceans in a new lake, the response variable is Species. The rest of the variables are predictors and will be used to predict Species. A linear model was chosen to investigate the relationship between Species and the predictor variables. Figure 1 shows a scatterplot matrix of the raw data.

Figure 1: Scatterplot Matrix of Raw Lake Data



From the scatterplots we can see that many predictors are not strongly correlated related to Species and do not represent a linear relationship. Histograms shown in Figure 2 were also created to investigate the normality of individual variables. As such, the variables Species, MeanDepth, NLakes, and Area were transformed using logarithms to make them more normal and linearly related to Species.

## Figure 2: Histograms Before and After Log Transformations

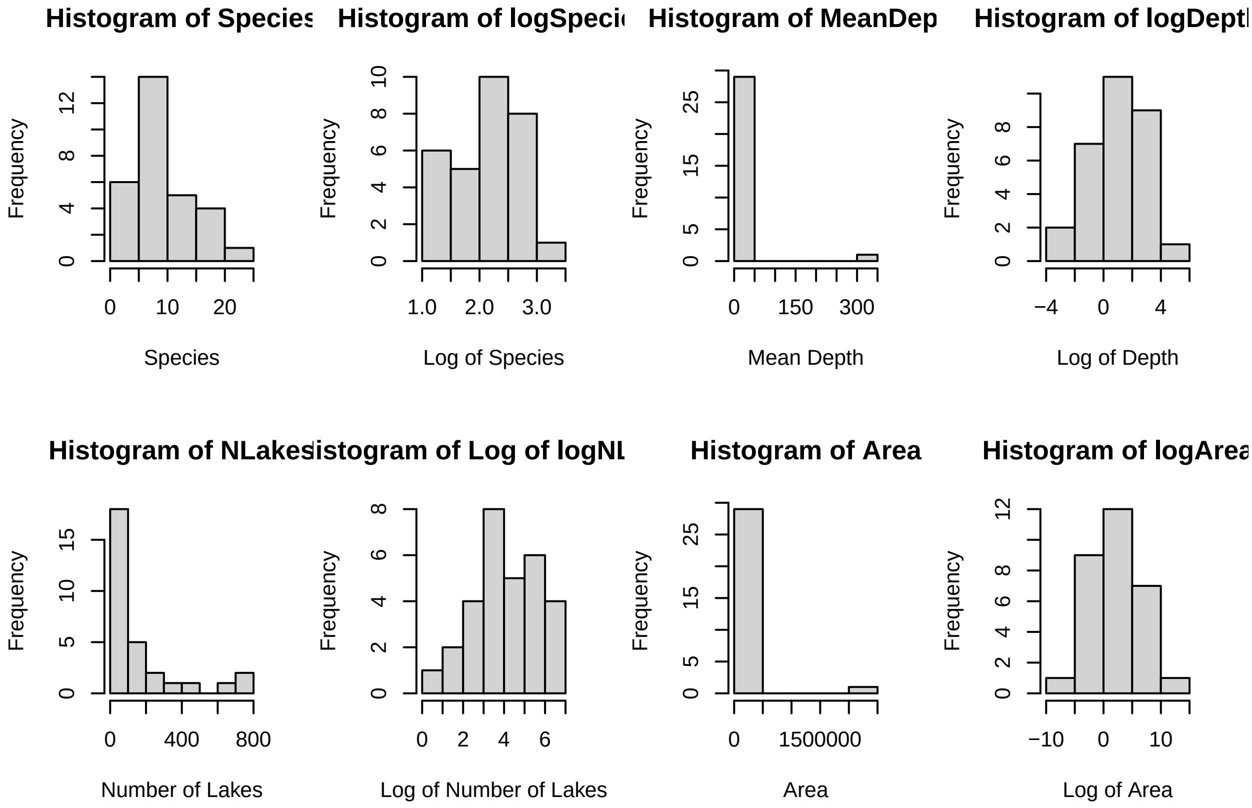


Figure 3: Scatterplot Matrix of Transformed Variables

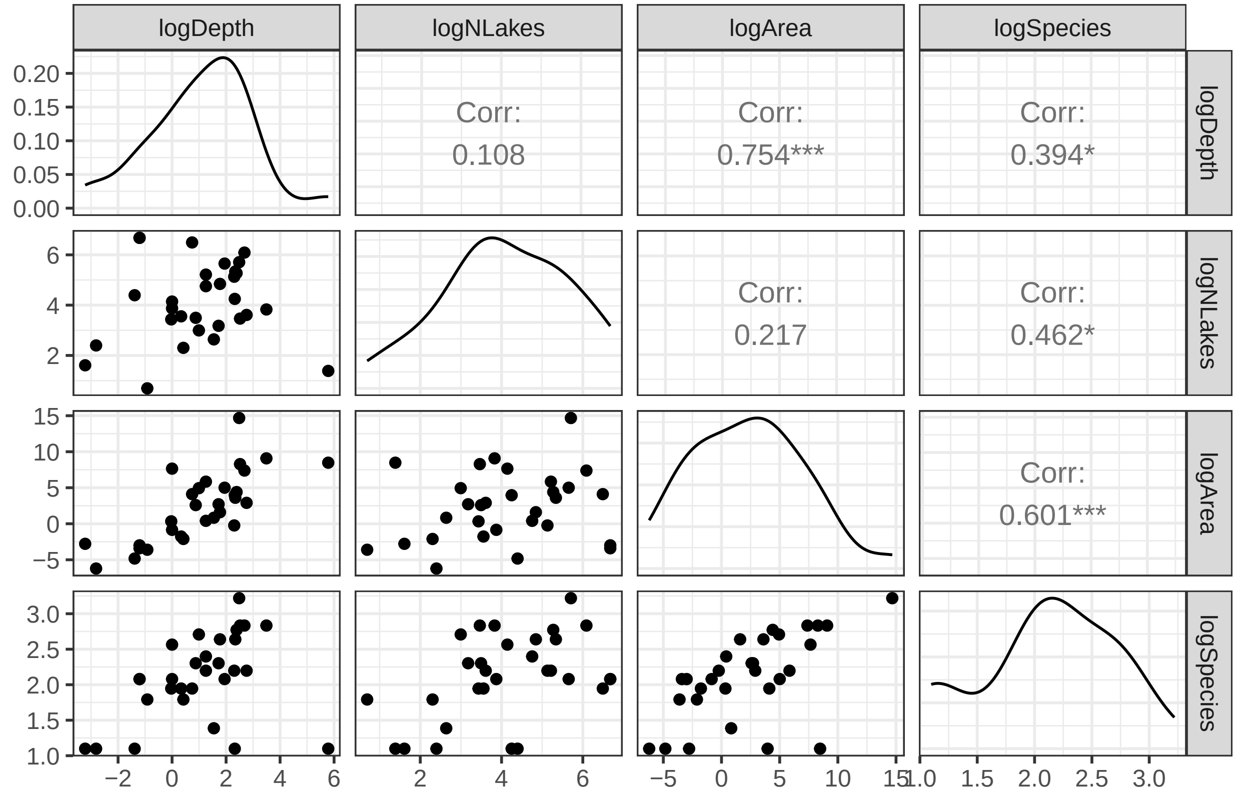


Figure 3 shows a scatterplot matrix with the transformed variables. The scatterplots look much more linear.

## Model Selection Methods

With the appropriate variables transformed, they could now be placed into different models. Many models were tested to identify the simplest and most accurate one. The first method used was model space enumeration. This method is best suited for data with a small number of predictors, generally less than 30. Since the lakes data has 7 predictors, this was an appropriate method to begin with. The method compared BIC values and returned a model that used Lat, logNLakes, and logArea to predict logSpecies.

The next method used was the guided search algorithm because it efficiently identifies a good model among many options using a greedy algorithm. First, stepwise selection was used and compared both AIC and BIC values. Both regressions returned a model that used Lat, logNLakes, and logArea to predict logSpecies. Next, forward and backward selection was used, comparing both AIC and BIC values. All of the regressions returned a model that used Lat, logNLakes, and logArea to predict logSpecies. Since all of these methods consistently returned the same model, we will proceed with using Lat, logNLakes, and logArea, in our main effects model.

The main effects model is $logSpecies_i = \beta_0 + \beta_1 Lat_i + \beta_2 logNLakes_i + \beta_3 logArea_i + e_i$.
$e_i \overset{iid}{\sim} N(0, \sigma^2)$

$logSpecies_i$ is the log of the number of species in the lake.
$Lat_i$ is the latitude of the lake in degrees North.
$logNLakes_i$ is the log of the number of lakes within 20km of the given lake.
$logArea_i$ is the log of the surface area of the lake.
$e_i$ is the error term.
$\beta_0$ = 1.91 and is the mean log number of species when all the predictors are equal to 0.
$\beta_1$ = -0.0176 and is the slope of the mean of Y vs. Lat if logNLakes and logArea are 0.
$\beta_2$ = 0.216 and is the slope of E(Y) vs. logNLakes if Lat and logArea are 0.
$\beta_3$ = 0.0650 and is the slope of E(Y) vs. logArea if Lat and logNLakes are 0.

With the main effects model chosen, interactions between predictors were then tested. First, the main effects model was compared with a model that included an interaction between Lat and logNLakes. This interaction model is $logSpecies_i = \beta_0 + \beta_1 Lat_i + \beta_2 logNLakes_i + \beta_3 logArea_i + \beta_4 Lat_i logNLakes_i$. $\beta_4$ is the change in the slope of E(Y) vs. Lat per unit of logNLakes. The null hypothesis is $H_0$: $\beta_4 = 0$; $\beta_1 \beta_2 \beta_3$ arbitrary. The alternative hypothesis is $H_1$: $\beta_4 \neq 0$; $\beta_1 \beta_2 \beta_3$ arbitrary. The F test returned a p-value of 0.83 which is above the significance level of $alpha = 0.01$. By the parsimony principle, we fail to reject the null hypothesis and conclude that the addition of the interaction effect between latitude and logNLakes to the main effects model is not useful.

Next, the main effects model was compared with including an interaction between Lat and logArea. This interaction model is $logSpecies_i = \beta_0 + \beta_1 Lat_i + \beta_2 logNLakes_i + \beta_3 logArea_i + \beta_5 Lat_i logArea_i$. $\beta_5$ is the change in the slope of E(Y) vs. Lat per unit of logArea. The null hypothesis is $H_0$: $\beta_5 = 0$; $\beta_1 \beta_2 \beta_3$ arbitrary. The alternative hypothesis is $H_1$: $\beta_5 \neq 0$; $\beta_1 \beta_2 \beta_3$ arbitrary. The F test returned a p-value of 0.43 which is above the significance level of $alpha = 0.01$. We fail to reject the null hypothesis and conclude that the addition of the interaction effect between latitude and logArea to the main effects model is not useful.

Lastly, the main effects model was compared with including an interaction between logNLakes and logNLakes. This interaction model is $logSpecies_i = \beta_0 + \beta_1 Lat_i + \beta_2 logNLakes_i + \beta_3 logArea_i + \beta_6 logNLakes_i logArea_i$. $\beta_6$ is the change in the slope of E(Y) vs. logArea per unit of logNLakes. The null hypothesis is $H_0$: $\beta_6 = 0$; $\beta_1 \beta_2 \beta_3$ arbitrary. The alternative hypothesis is $H_1$: $\beta_6 \neq 0$; $\beta_1 \beta_2 \beta_3$ arbitrary. The F test returned a p-value of 0.73 which is above the significance level of $alpha = 0.01$. We fail to reject the null hypothesis and conclude that the addition of the interaction effect between logNLakes and logArea to the main effects model is not useful.

## Final Model Selected

Once all two-way interaction terms were tested and determined to be insignificant, a three-way interaction test was decided to be unnecessary. The final model selected is $logSpecies_i = \beta_0 + \beta_1 Lat_i + \beta_2 logNLakes_i$ $+ \beta_3 logArea_i$. This makes logical sense because $\beta_1$ is negative, so as latitude increases and the lakes become closer to the north (colder), the number of species decreases. $\beta_2$ is positive so a lake with many other lakes around it, which is likely to be more biodiverse due to being closer to other crustaceans, has more species. $\beta_3$ is positive, so a lake that is larger in area would also likely have more species than a small lake. One assumption of this model is that the mean of Species is a linear function of the predictors. We also assume that the errors are independent, normally distributed, and have constant variance.

## Model Fit and Diagnostics

Diagnostics were run on the selected model to assess fit and prove assumptions. A residual vs. fitted values plot, standardized residual vs. fitted values plot, and Q-Q plot were created, shown in Figure 4. Since the residuals are randomly scattered around zero, $e_i$ is assumed to be independently distributed and the mean function is a linear function of the predictors. Since the standardized residuals are randomly scattered around zero, $e_i$ is assumed to have constant variance. Since the Q-Q plot is roughly linear, $e_i$ is assumed to be normal.

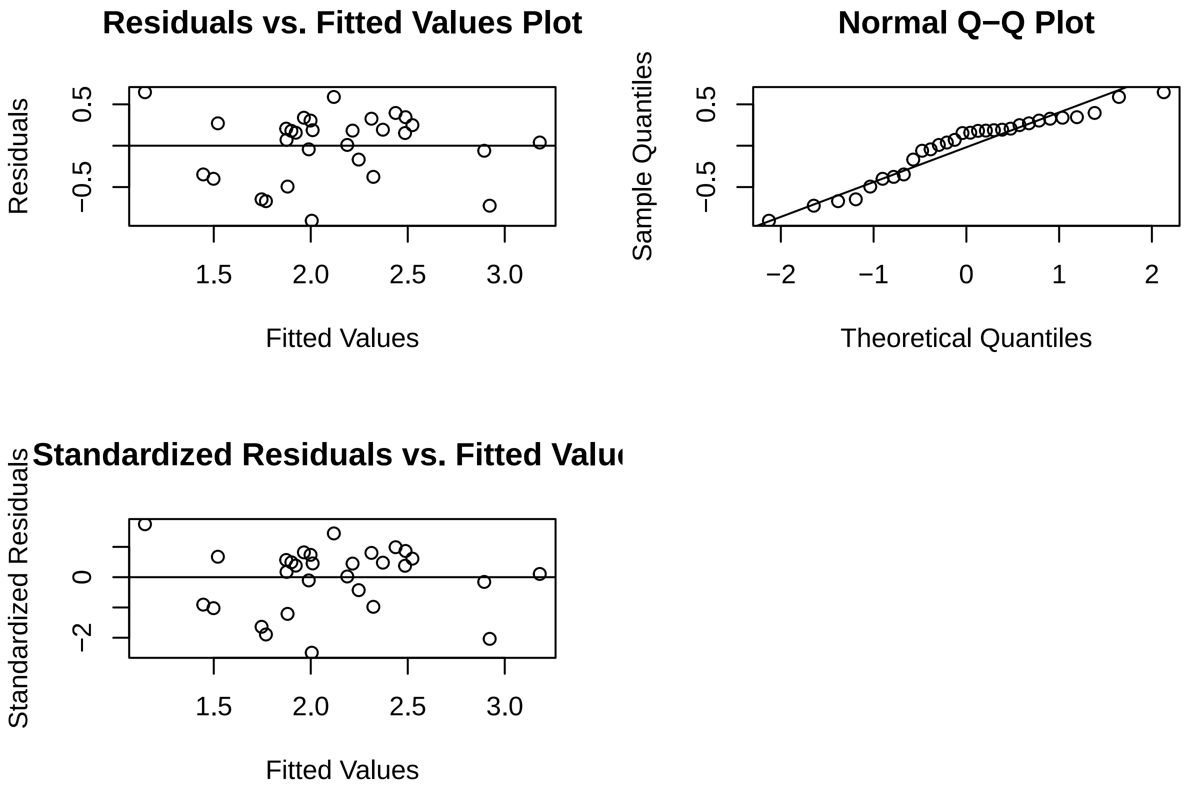
## Interpretations

$\beta_1$ means that a lake with a 5% increase in latitude is expected to have about a 0.037% decrease in the number of species.

$\beta_2$ means that a lake with 5% more nearby lakes is expected to have about a 0.46% increase in the number of species.

$\beta_3$ means that a lake with 5% more surface area is expected to have about a 0.14% increase in the number of species.

## Figure 4: Residual and Q-Q Plots



## Predicting Species in New Lake

A 95% prediction interval for the number of crustacean species in a new lake that has a latitude of 46 degrees, 44 other lakes within 20 kilometers, and a surface area of 58,000 hectares is (49.66, 3663.53).

## Appendix

```r
library(tidyverse)

library(patchwork)

library(broom)

library(scatterplot3d)

library(GGally)
library(plotly)
library(leaps)
library(MASS)

#import data
lakes <-
  read.csv("/Users/jessicali/Documents/STAT 3301/datasets/projectdata67.csv")
```

## Exploratory Data Analysis

### Scatterplot matrix of raw data

```
colsx = which(colnames(lakes) %in% c("MeanDepth", "Cond","Elev", "Lat", "Long",
                                     "NLakes", "Photo", "Area"))
colsy = which(colnames(lakes) %in% c("Species"))
ggpairs(lakes, columns = c(colsx, colsy)) + theme_bw()
```

### Histograms and two-way scatterplots of variables

```
hist(lakes$Species)
```

```
logSpecies <- log(lakes$Species)
hist(logSpecies)
```

```
hist(lakes$MeanDepth)
```

```
plot(lakes$MeanDepth, logSpecies)
```

```
logDepth <- log(lakes$MeanDepth)
hist(logDepth)
```

```r
plot(logDepth, logSpecies)
```

```r
hist(lakes$Cond)
```

```r
plot(lakes$Cond, logSpecies)
```

```r
logCond <- log(lakes$Cond)
hist(logCond)
```

```r
plot(logCond, logSpecies)
```

```r
hist(lakes$Elev)
```

```r
plot(lakes$Elev, logSpecies)
```

```r
logElev <- log(lakes$Elev)
```

```
## Warning in log(lakes$Elev): NaNs produced
```

```r
hist(logElev)
```

```r
plot(logElev, logSpecies)
```

```r
hist(lakes$Lat)
```

```r
plot(lakes$Lat, logSpecies)
```

```r
logLat <- log(lakes$Lat)
hist(logLat)
```

```r
plot(logLat, logSpecies)
```

```r
hist(lakes$NLakes)
```

```r
plot(lakes$NLakes, logSpecies)
```

```r
logNLakes <- log(lakes$NLakes)
hist(logNLakes)
```

```r
plot(logNLakes, logSpecies)
```

```r
hist(lakes$Area)
```

```r
plot(lakes$Photo, logSpecies)
```

```r
logPhoto <- log(lakes$Photo)


plot(logPhoto, logSpecies)
```

```r
plot(lakes$Area, logSpecies)
```

```r
logArea <- log(lakes$Area)
hist(logArea)
```

```r
plot(logArea, logSpecies)
```

```r
#log-transform variables
lakes <- read_csv("/Users/jessicali/Documents/STAT 3301/datasets/projectdata67.csv") %>%
  mutate(logSpecies = log(Species),
         logDepth = log(MeanDepth),
         logNLakes = log(NLakes),
         logArea = log(Area))
```

```
## New names:
## Rows: 30 Columns: 10
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): ...1 dbl (9): Species, MeanDepth, Cond, Elev, Lat, Long, NLakes, Photo,
## Area
## i Use ‘spec()‘ to retrieve the full column specification for this data. i
## Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
## * ‘‘ -> ‘...1‘
```

```
#create subset of relevant data, including log-transformed variables
lakes <- subset(lakes, select = c(logSpecies, logDepth, Cond, Elev, Lat, Long, logNLakes, Photo, logArea

#create new scatterplot
colsx = which(colnames(lakes) %in% c("logDepth", "Cond", "logNLakes", "logArea","Photo","Lat","Elev"))
colsy = which(colnames(lakes) %in% c("logSpecies"))
ggpairs(lakes, columns = c(colsx, colsy)) + theme_bw()
```

## Model Selection Methods

### Model Space Enumeration using BIC

```
regfit_full = regsubsets(logSpecies~.,data = lakes, nvmax = 8)
reg.summary=(summary(regfit_full))
reg.summary$bic
```

```
## [1]  -6.656045  -9.229911 -10.798951  -8.444126  -5.074101  -1.714949   1.659431
## [8]   5.036659
```

### Guided Search algorithms

```


#start from null model
null = lm(logSpecies ~ 1, data = lakes)
full = lm(logSpecies ~., data = lakes)
n = dim(lakes)[1]


# Stepwise regression by AIC
```

```r
stepAIC(null, scope = list(upper = full),
        direction="both", k=2)
```

### Stepwise selection

```
## Start:  AIC=-29.59
## logSpecies ~ 1
##
##               Df Sum of Sq     RSS     AIC
## + logArea     1    3.7838  6.6835 -41.047
## + logNLakes   1    2.2337  8.2337 -34.789
## + logDepth    1    1.6279  8.8395 -32.659
## + Elev        1    0.8989  9.5684 -30.282
## + Photo       1    0.8386  9.6288 -30.093
## <none>                    10.4673 -29.588
## + Cond        1    0.6220  9.8453 -29.426
## + Long        1    0.1210 10.3463 -27.937
## + Lat         1    0.0264 10.4409 -27.664
##
## Step:  AIC=-41.05
## logSpecies ~ logArea
##
##               Df Sum of Sq     RSS     AIC
## + logNLakes   1    1.2070  5.4766 -45.022
## <none>                     6.6835 -41.047
## + Elev        1    0.4074  6.2761 -40.933
## + Photo       1    0.3046  6.3790 -40.446
## + logDepth    1    0.0838  6.5997 -39.425
## + Cond        1    0.0748  6.6087 -39.384
## + Lat         1    0.0055  6.6780 -39.071
## + Long        1    0.0002  6.6833 -39.047
## - logArea     1    3.7838 10.4673 -29.588
##
```

```
## Step:  AIC=-45.02
## logSpecies ~ logArea + logNLakes
##
##             Df Sum of Sq    RSS     AIC
## + Lat        1   0.83615 4.6404 -47.992
## + Photo      1   0.42560 5.0510 -45.449
## <none>                   5.4766 -45.022
## + Cond       1   0.13474 5.3418 -43.769
## + Long       1   0.13257 5.3440 -43.757
## + Elev       1   0.06621 5.4103 -43.387
## + logDepth   1   0.03783 5.4387 -43.230
## - logNLakes  1   1.20697 6.6835 -41.047
## - logArea    1   2.75713 8.2337 -34.789
##
## Step:  AIC=-47.99
## logSpecies ~ logArea + logNLakes + Lat
##
##             Df Sum of Sq    RSS     AIC
## <none>                   4.6404 -47.992
## + Elev       1   0.15906 4.4813 -47.038
## + Photo      1   0.04086 4.5996 -46.257
## + logDepth   1   0.02229 4.6181 -46.136
## + Cond       1   0.01224 4.6282 -46.071
## + Long       1   0.00050 4.6399 -45.995
## - Lat        1   0.83615 5.4766 -45.022
## - logNLakes  1   2.03757 6.6780 -39.071
## - logArea    1   2.71740 7.3578 -36.163
```

```
##
## Call:
## lm(formula = logSpecies ~ logArea + logNLakes + Lat, data = lakes)
##
## Coefficients:
```

```
## (Intercept)     logArea    logNLakes          Lat
##     1.91038     0.06504      0.21574     -0.01761
```

```
# Stepwise regression by BIC
stepAIC(null, scope = list(upper = full),
        direction = "both", k = log(n))
```

```
## Start:  AIC=-28.19
## logSpecies ~ 1
##
##             Df Sum of Sq     RSS     AIC
## + logArea    1    3.7838  6.6835 -38.244
## + logNLakes  1    2.2337  8.2337 -31.986
## + logDepth   1    1.6279  8.8395 -29.857
## <none>                   10.4673 -28.187
## + Elev       1    0.8989  9.5684 -27.479
## + Photo      1    0.8386  9.6288 -27.291
## + Cond       1    0.6220  9.8453 -26.624
## + Long       1    0.1210 10.3463 -25.135
## + Lat        1    0.0264 10.4409 -24.861
##
## Step:  AIC=-38.24
## logSpecies ~ logArea
##
##             Df Sum of Sq     RSS     AIC
## + logNLakes  1    1.2070  5.4766 -40.818
## <none>                    6.6835 -38.244
## + Elev       1    0.4074  6.2761 -36.730
## + Photo      1    0.3046  6.3790 -36.242
## + logDepth   1    0.0838  6.5997 -35.221
## + Cond       1    0.0748  6.6087 -35.181
## + Lat        1    0.0055  6.6780 -34.868
## + Long       1    0.0002  6.6833 -34.844
## - logArea    1    3.7838 10.4673 -28.187
##
## Step:  AIC=-40.82
## logSpecies ~ logArea + logNLakes
##
##             Df Sum of Sq    RSS     AIC
## + Lat        1   0.83615 4.6404 -42.387
## <none>                   5.4766 -40.818
## + Photo      1   0.42560 5.0510 -39.844
## - logNLakes  1   1.20697 6.6835 -38.244
## + Cond       1   0.13474 5.3418 -38.164
## + Long       1   0.13257 5.3440 -38.152
## + Elev       1   0.06621 5.4103 -37.782
## + logDepth   1   0.03783 5.4387 -37.625
## - logArea    1   2.75713 8.2337 -31.986
##
## Step:  AIC=-42.39
## logSpecies ~ logArea + logNLakes + Lat
##
##             Df Sum of Sq    RSS     AIC
## <none>                   4.6404 -42.387
## - Lat        1   0.83615 5.4766 -40.818
## + Elev       1   0.15906 4.4813 -40.032
## + Photo      1   0.04086 4.5996 -39.251
## + logDepth   1   0.02229 4.6181 -39.130
## + Cond       1   0.01224 4.6282 -39.065
## + Long       1   0.00050 4.6399 -38.989
## - logNLakes  1   2.03757 6.6780 -34.868
## - logArea    1   2.71740 7.3578 -31.959
```

```
##
## Call:
## lm(formula = logSpecies ~ logArea + logNLakes + Lat, data = lakes)
```

```
##
## Coefficients:
## (Intercept)      logArea    logNLakes          Lat
##     1.91038      0.06504      0.21574     -0.01761
```

```
#start from full model
# Stepwise regression by AIC
stepAIC(full, scope = list(lower = null, upper = full),
        direction = "both", k = 2)

# Stepwise regression by BIC
stepAIC(full, scope = list(lower = null, upper = full),
        direction = "both", k = log(n))
```

```


# Forward Selection by AIC
stepAIC(null, scope = list(upper = full),
        direction = "forward", k = 2)

# Forward Selection by BIC
stepAIC(null, scope = list(upper = full),
        direction = "forward", k = log(n))
```

**Forward selection**

```


# Backward Elimination by AIC
stepAIC(full, direction = "backward", k = 2)

# Backward Elimination by BIC
stepAIC(full, direction = "backward", k = log(n))
```

**Backwards selection**

## Testing Interactions

### No interaction vs. Lat*NLakes

```
#main effects model (no interactions)
main.model = lm(formula = logSpecies ~ Lat + logNLakes + logArea, data = lakes)
#interaction between Lat and logNLakes
latNLakesInt.model = lm(logSpecies ~ Lat + logNLakes + logArea + Lat:logNLakes, data = lakes)

#anova test between models
anova(main.model, latNLakesInt.model)
```

```
## Analysis of Variance Table
##
## Model 1: logSpecies ~ Lat + logNLakes + logArea
## Model 2: logSpecies ~ Lat + logNLakes + logArea + Lat:logNLakes
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 4.6404
## 2     25 4.6312  1  0.009219 0.0498 0.8253
```

### No interaction vs. Lat*logArea

```
#main effects model (no interactions)
main.model = lm(formula = logSpecies ~ Lat + logNLakes + logArea, data = lakes)
#interaction between Lat and logArea
latAreaInt.model = lm(logSpecies ~ Lat + logNLakes + logArea+Lat:logArea, data = lakes)

anova(main.model, latAreaInt.model)
```

### No interaction vs. logNLakes*logArea

```
#main effects model (no interactions)
main.model = lm(formula = logSpecies ~ Lat + logNLakes + logArea, data = lakes)
#interaction between logNLakes and logArea
NLakesAreaInt.model = lm(logSpecies ~ Lat + logNLakes + logArea+ logNLakes:logArea, data = lakes)

anova(main.model, NLakesAreaInt.model)
```

## Check model fit and diagnostics

```
#find betas
summary(main.model)
```

```
##
## Call:
## lm(formula = logSpecies ~ Lat + logNLakes + logArea, data = lakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9065 -0.3023  0.1550  0.2651  0.6464
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.910381   0.318410   6.000 2.46e-06 ***
## Lat         -0.017614   0.008138  -2.164 0.039794 *
## logNLakes    0.215740   0.063851   3.379 0.002305 **
## logArea      0.065040   0.016668   3.902 0.000603 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4225 on 26 degrees of freedom
## Multiple R-squared:  0.5567, Adjusted R-squared:  0.5055
## F-statistic: 10.88 on 3 and 26 DF,  p-value: 8.198e-05
```

```
resid.lakes = resid(main.model)
fitted.lakes = fitted(main.model)
#residual plot
plot(fitted.lakes, resid.lakes)
```

```
##qqplot of residuals
qqnorm(resid.lakes); qqline(resid.lakes)
```

```
#standardized residuals
plot(fitted(main.model), rstandard(main.model), xlab = "Fitted Values",
     ylab = "Standardized Residuals", main = "Standardized Residuals vs. Fitted Values"); abline(h = 0)
```

## Predicting

```
predict(main.model, newdata = data.frame(Lat = 46, logNLakes = log(44), logArea = log(58000)),se.fit=TRU
```

```
## $fit
##        fit      lwr      upr
## 1 2.629928 1.696013 3.563844
##
## $se.fit
## [1] 0.1671834
##
## $df
## [1] 26
##
## $residual.scale
## [1] 0.4224658
```