Project Report

Hwanho Kim

2024-12-04

Introduction:

The goal of this project is to create a simple model that accurately and properly predict the number of crustacean species in a new lake based on the a given data set. The data contains 30 different lakes' data in terms of 9 variables: number of crustacean species, mean lake depth (in meters), specific conductance (a measure of mineral content) (in micro Siemens), lake elevation (in meters), latitude, longitude, number of lakes within 20km, rate of photosynthesis (in C^{14}) and surface area of lake (in hectares).

Exploratory Data Analysis (EDA):

1. Uni-variate distribution

The first step will be Exploratory Data Analysis. It helps to detect outliers and anomalies of data and transform them if needed. Species column which is represent the number of crustacean species living in lake should be reviewed first because this is response variable. One of the best way for EDA is creating a histogram with statistical measures to see the distribution of the data. On the [Figure1], it is a histogram of species data with mean, median, quartile, min, and max. It shows that the distribution of Species is skewed right which is not normally distributed, and there are two outliers at 30 and 32. To fix this distribution and handle the outliers, the log transformation to Species is performed. On the [Figure2], it is a histogram of log transformed species, and the distribution is pretty normal and there is not outlier.

2.Bivariate distribution

The Species column is already transformed, and other 8 predictors should be considered as well for better linear relation with response variable. [Figure3] is the matrix of scatter plot between each predictor and LogSpecies, and those scatter plots between LogSpecies and predictors helped to decide whether transformation is required on each variable or not. For example,

For the MeanDepth variable, it is hard to see linear relation between variables because of the three outliers and scale of predictor. MeanDepth should be log transformed for better linear relation. For the Cond variable, there is one outlier and transformation improve the linearity between variables even though the scale is not that severe like MeanDepth variable. For the Elev, Photo variables, these variables contain negative value, and it has empty value if negative value is log transformed. Therefore, those are not transformed. For the Lat variable, it is transformed for the better linear relation with response variable. For the long variable, transformation does not affect to the linearity with response variable. Through those process, the list of predictors after the transformation will be [LogMeanDepth, LogCond, Elev, LogLat, Long, LogNLakes, Photo, LogArea]. For the prediction model building, those variables will be used other than original scale variables. [Figure4] is the matrix of scatter plot after the application of log transformation for required variables, and can confirm that the linear relation between response variable is well improved.

Model Selection:

I want to build simple and accurate model to predict Species. For this model, AIC and BIC is useful to decide how many columns is the relevant for modeling. The formula for AIC and BIC is

$AIC_{M_k} = nlog(RSS_{M_k}/2) + 2p_{M_k}$

$BIC_{M_k} = nlog(RSS_{M_k}/2) + log(n)p_{M_k}$

This is a combination of model fit $nlog(RSS_{M_k}/2)$ and model complexity $2p_{M_k}$ or $log(n)p_{M_k}$, and smaller values of AIC or BIC are preferred as relevant model. Adding a variable to a model will improve the model fit term but incur a penalty, and those measures will only decrease if the improvement in model fit outweighs the penalty.

1.Best subset

The regsubsets() R function performs best subset selection by identifying the best model that contains a given number of predictors, where best is qualified using RSS or BIC. [Figure 5] is line graph of RSS or BIC value depends on the number of variables of the model. The model that use all 8 predictors has the lowest RSS which means that it returns the most accurate prediction, and the model that use two variables has lowest BIC which means that it is the most balanced model. The two variables used was “Elev” and “LogArea”. The goal of this modeling was create a simple and accurate model, so BIC is more relevant measure to select model.

2.Forward Selection

Forward selection is one of the methods for model selection. It start with a model with a small number of predictors, typically the null model, and consider add a predictor that not already in the model depends on AIC or BIC values. [Appendix1] shows the process of Forward Selection based on the AIC. It starts from the null model, and LogArea has lowest AIC and RSS at the first step, which means that this variable will be added to the model. At the next step, Elev variable also be added to the model and it stops because adding a new predictor does not reduce the AIC further.

[Appendix2] shows the process of Forward Selection based on the BIC. It also starts from the null model and add variable if adding it reduce the AIC of the current model for each step. As a final model, it also has LogArea and Elev variables as predictors.

3.Backward Selection

Backward Selection is opposite method with Forward Selection. It start with a model with a large number of predictors and delete a predictor for each step. [Appendix3] shows the process of Backward Selection based on the AIC value. Start from the model that has all 8 predictors, it delete one predictor that has largest AIC for each step, and stop when current model has lower AIC than the model that delete one more column. The final model from Backward Selection based on AIC value is model that has Elev and LogArea as predictors. [Appendix4] shows the process of Backward Selection based on the BIC value, and the selected model is the same model with AIC value model.

4.Stepwise regression

Stepwise regression is another model selecting method, and it start with null or full model, and add or delete a predictor from current model at each step depends on the AIC or BIC value. [Appendix5] is a process of stepwise regression starting from null with AIC value. [Appendix6] is a process of stepwise regression starting from null with BIC value. [Appendix7] is a process of stepwise regression starting from full with AIC value. [Appendix8] is a process of stepwise regression starting from full with BIC value. No matter what model start from and what value used for selection, the selected model has Log Area and Elev variables as predictors.

5.Interaction and F test

Hypotheses:

$H_0 : \beta_{LogArea,Elev} = 0$

$H_1 : \beta_{LogArea,Elev} \neq 0$

this is equivalent with

$H_0 : E(LogSpecise) = \hat{\beta}_0 + LogArea * \hat{\beta}_1 + Elev\hat{\beta}_2$

$H_1 : E(LogSpecise) = \hat{\beta}_0 + LogArea * \hat{\beta}_1 + Elev * \hat{\beta}_2 + (LogArea * Elev) * \hat{\beta}_3$

Null hypothesis should be reduced model and alternative hypothesis should be full model. [Appendix9] is the process that do F test using R code. p value(0.6765) is not less than 0.05, which means that adding interaction cannot improve the model because it fail to reject null hypothesis.

Final Model Selection and Assumptions:

1.Final Model Selection

Based on the Forward, Backward Selection, Stepwise regression, Best subset method and F test, the final selected model is

$log(Species)_i = \hat{\beta}_0 + LogArea_i * \hat{\beta}_1 + Elev_i * \hat{\beta}_1 + e_i$

when residuals are independently and identically distributed with $E(e_i) = 0$ and $var(e_i) = \sigma^2$

2.Prove Assumptions

[Figure6] shows the Residuals VS Fitted plot. Residuals are relatively randomly scattered around the horizontal line and the clear patterns are not shown in here. Therefore, the residuals are relatively normally distributed and linear. [Figure6] shows the Q-Q plot and dots are lied on the linear line, which means that residuals are normally distributed. [Figure6] shows the scale-Location plot and the most of the dots are spread evenly across fitted value except one outlier. We can think the residual has quite constant variance. [Figure6] shows the Residual VS Leverage plot, and residual are quite independent each other because the scale of standardized residuals are not that big depends on the leverage level. Through those plots, the linear model is assumed to be normally and independently distributed with constant variance.

Diagnostics:

The normality, linearity and constant variance are already confirmed from the previous step. [Table6] is the histogram of residual of this model. It also shows that the distribution of residuals are pretty normal.

Interpretation of the model

Our model is: $log(Species)_i = \hat{\beta}_0 + LogArea_i * \hat{\beta}_1 + Elev_i * \hat{\beta}_1 + e_i$

Interpretation:

[Appendix10] has summary of my model.

$\hat{\beta}_0$: This represent the expected value of log(Species) when both LogArea and Elev are 0.

$\hat{\beta}_1$: For every 1-unit increase in LogArea, the expected value of log(Species) increase by $\hat{\beta}_1$ with fixed Elev. For the original scale, the species count multiplies by $e^{\hat{\beta}_1}$. $e^{0.06829} \approx 1.071$ and Percentage Change = (1.071 - 1) x 100 = 7.09%, which means that 7.1% increase in species count per 1 unit increase in LogArea.

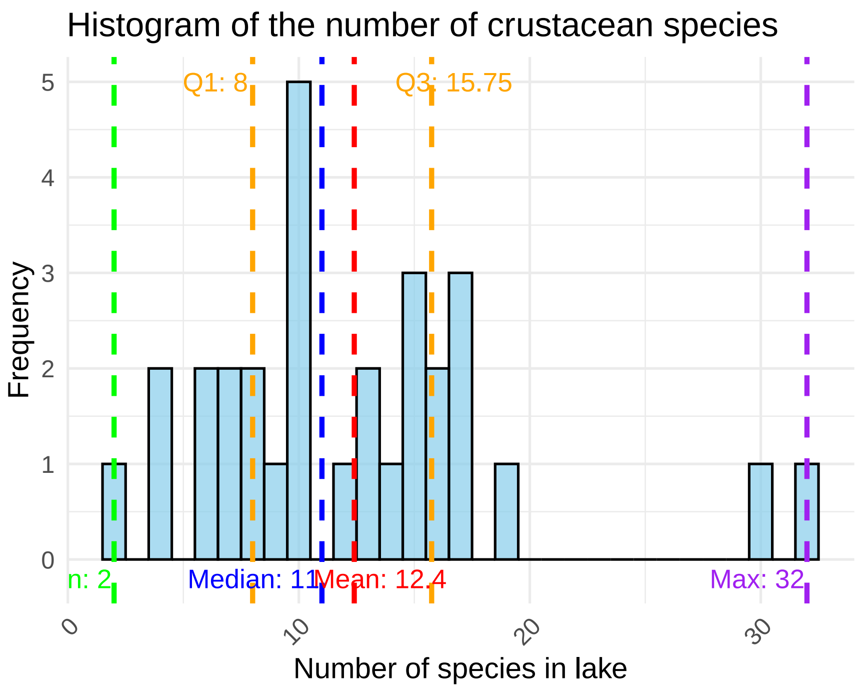
$\hat{\beta}_2$: For every 1-unit increase in Elev, the expected value of log(Species) increase by $\hat{\beta}_2$ with fixed LogArea. $e^{-0.0001702} \approx 0.99983$ and Percentage Change = (0.99983 - 1) x 100 = -0.017%. Therefore, 0.017% decrease in Species count per 1 unit increase in Elev.

Prediction

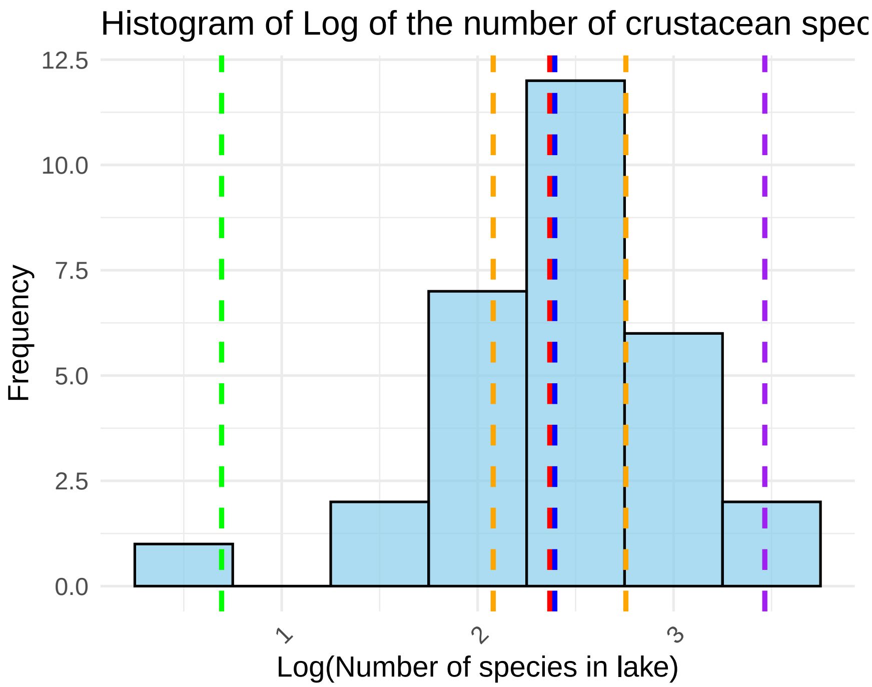
The characteristics given in the project for my model is Area = 58,000 and Elev = 372. [Figure8] shows the process to get 95% confidence prediction interval using R code. As a result, predicted number of crustacean species for a new lake with 58,000 hectares of area and 372 meter elevation is approximately 18. With 95% confidence, the number of crustacean species for a new lake is expected to be between approximately 9 and 37.

Figure

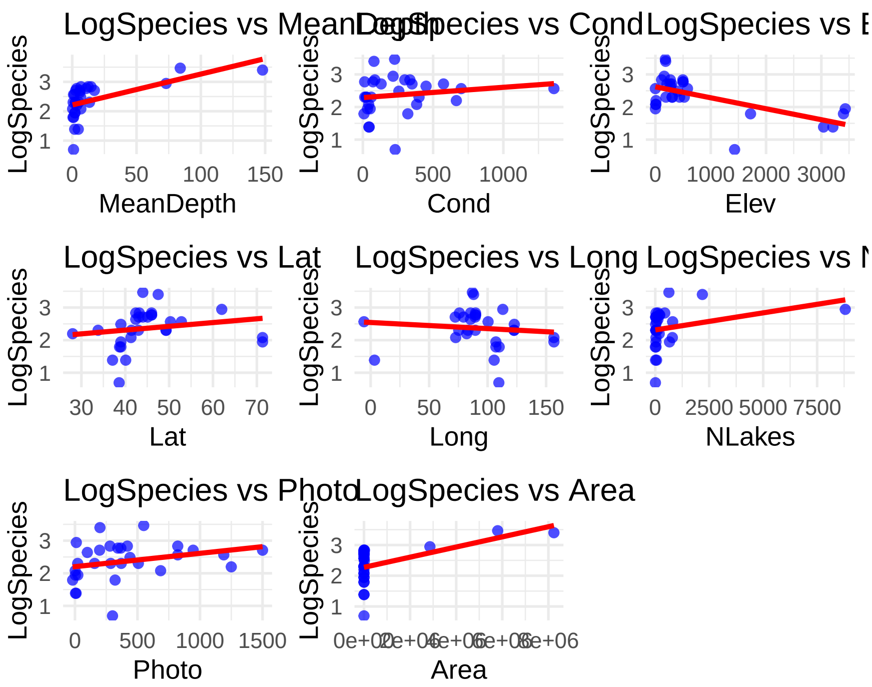
[Figure1]



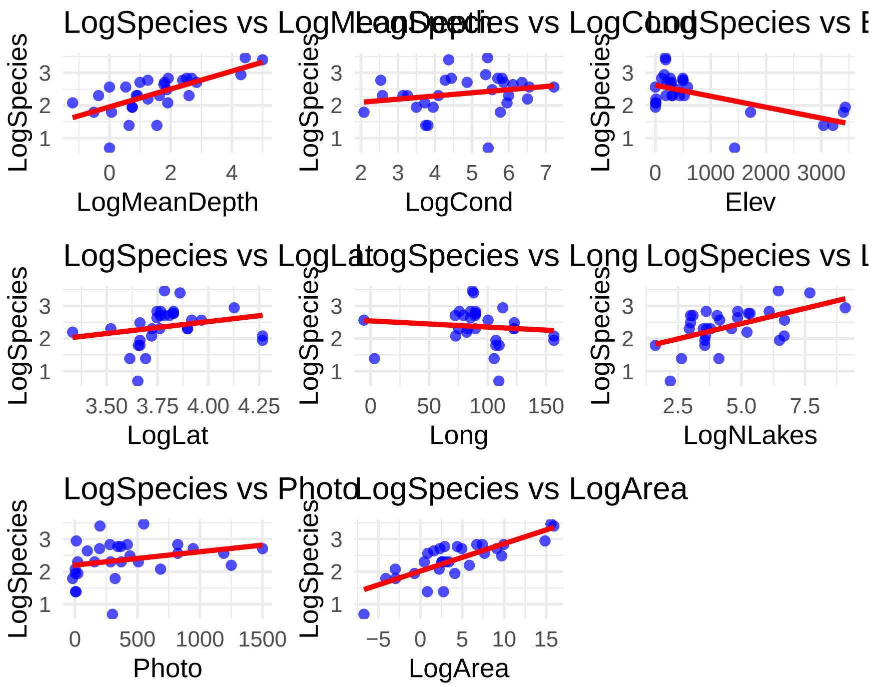
[Figure2]



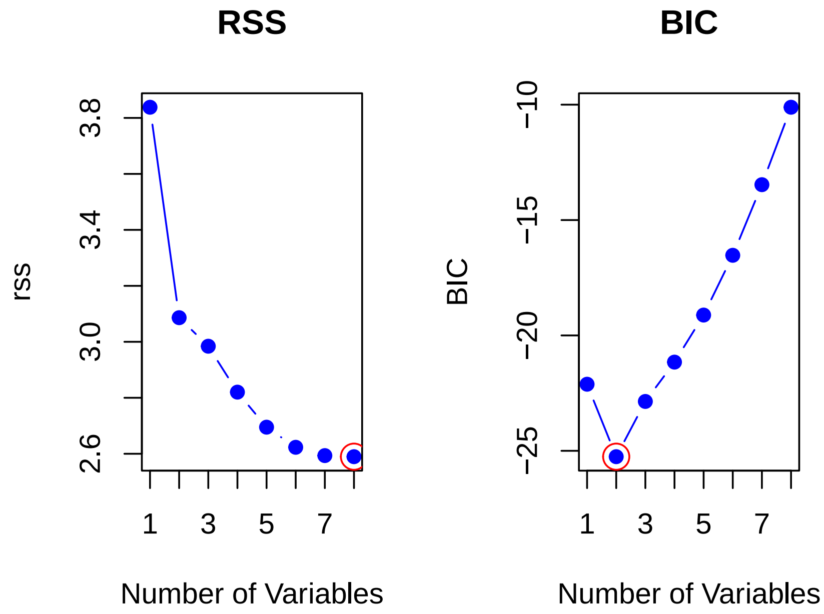
[Figure3]



[Figure4]



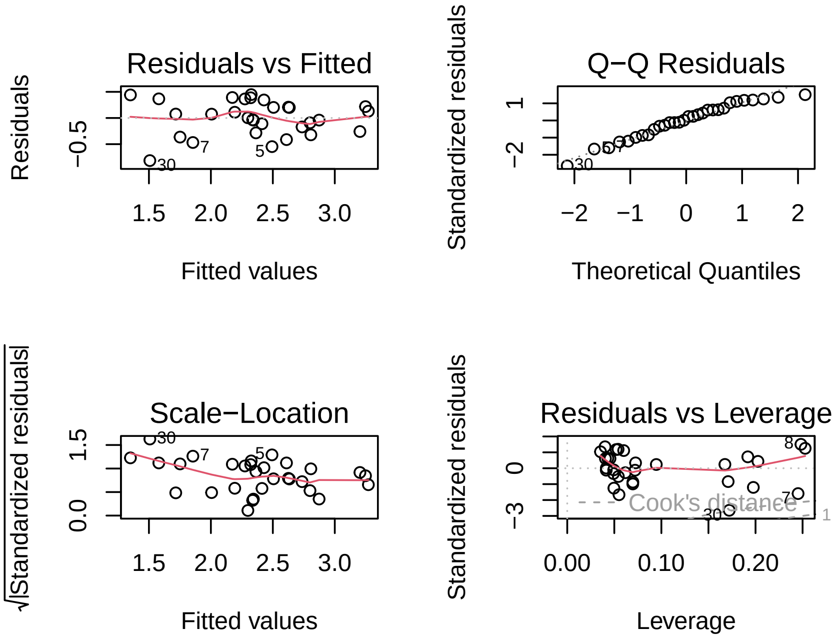
[Figure5]



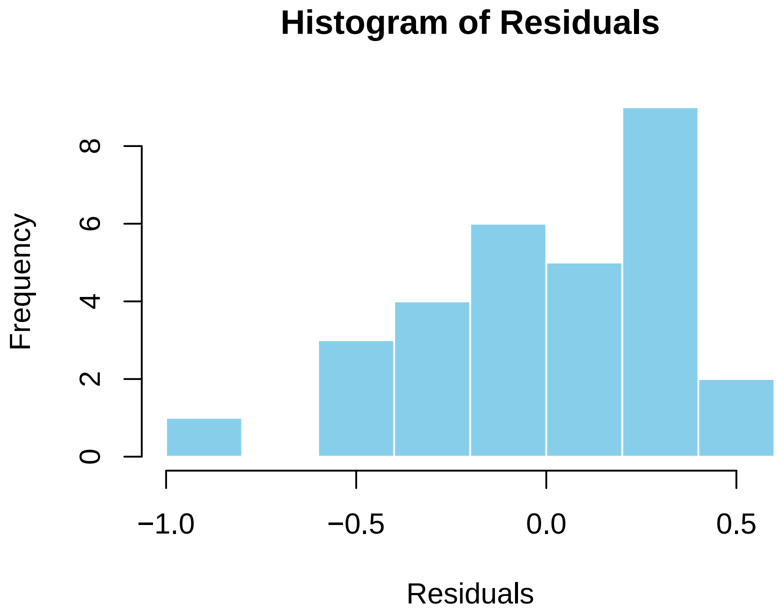
```
## [1] 2
```

```
## [1] "Elev"    "LogArea"
```

[Figure6]



[Figure7]



[Figure8]

```
## [1] "The 95% prediction interval is: "
```

```
##        fit      lwr      upr
## 1 18.17034 8.809747 37.47682
```

Appendix

#Appendix 1

```
## Start:  AIC=-30.77
## LogSpecies ~ 1
##
##                 Df Sum of Sq     RSS     AIC
## + LogArea        1    6.2251  3.8382 -57.686
## + LogMeanDepth   1    4.5865  5.4768 -47.021
## + Elev           1    3.7908  6.2725 -42.951
## + LogNLakes      1    2.9877  7.0755 -39.337
## + Photo          1    0.8108  9.2524 -31.289
## <none>                       10.0632 -30.769
## + LogLat         1    0.5671  9.4961 -30.509
## + LogCond        1    0.5045  9.5587 -30.312
## + Long           1    0.1144  9.9488 -29.112
##
## Step:  AIC=-57.69
## LogSpecies ~ LogArea
##
##                 Df Sum of Sq    RSS     AIC
## + Elev          1   0.75197 3.0862 -62.228
## + Photo         1   0.35292 3.4852 -58.580
## <none>                      3.8382 -57.686
## + LogNLakes     1   0.17227 3.6659 -57.064
## + LogLat        1   0.14037 3.6978 -56.804
## + LogCond       1   0.02510 3.8131 -55.883
## + LogMeanDepth  1   0.02059 3.8176 -55.848
## + Long          1   0.01802 3.8201 -55.827
##
## Step:  AIC=-62.23
## LogSpecies ~ LogArea + Elev
##
##                Df Sum of Sq    RSS     AIC
## <none>                      3.0862 -62.228
## + LogMeanDepth  1  0.101938 2.9842 -61.236
## + LogNLakes     1  0.081703 3.0045 -61.033
## + Photo         1  0.075557 3.0106 -60.971
## + LogCond       1  0.030049 3.0561 -60.521
## + LogLat        1  0.011749 3.0744 -60.342
## + Long          1  0.000063 3.0861 -60.228
```

```
##
## Call:
## lm(formula = LogSpecies ~ LogArea + Elev, data = selected_data)
##
## Coefficients:
## (Intercept)      LogArea         Elev
##   2.2118578    0.0684940   -0.0001702
```

#Appendix 2

```
## Start:  AIC=-29.37
## LogSpecies ~ 1
##
##                Df Sum of Sq     RSS     AIC
## + LogArea       1    6.2251  3.8382 -54.884
## + LogMeanDepth  1    4.5865  5.4768 -44.218
## + Elev          1    3.7908  6.2725 -40.148
## + LogNLakes     1    2.9877  7.0755 -36.534
## <none>                      10.0632 -29.368
## + Photo         1    0.8108  9.2524 -28.487
## + LogLat        1    0.5671  9.4961 -27.707
## + LogCond       1    0.5045  9.5587 -27.510
## + Long          1    0.1144  9.9488 -26.310
##
## Step:  AIC=-54.88
## LogSpecies ~ LogArea
##
##                Df Sum of Sq    RSS     AIC
## + Elev          1   0.75197 3.0862 -58.024
## <none>                      3.8382 -54.884
## + Photo         1   0.35292 3.4852 -54.376
## + LogNLakes     1   0.17227 3.6659 -52.860
## + LogLat        1   0.14037 3.6978 -52.600
## + LogCond       1   0.02510 3.8131 -51.679
## + LogMeanDepth  1   0.02059 3.8176 -51.644
## + Long          1   0.01802 3.8201 -51.624
##
## Step:  AIC=-58.02
## LogSpecies ~ LogArea + Elev
##
##                Df Sum of Sq    RSS     AIC
## <none>                      3.0862 -58.024
## + LogMeanDepth  1  0.101938 2.9842 -55.631
## + LogNLakes     1  0.081703 3.0045 -55.428
## + Photo         1  0.075557 3.0106 -55.367
## + LogCond       1  0.030049 3.0561 -54.917
## + LogLat        1  0.011749 3.0744 -54.738
## + Long          1  0.000063 3.0861 -54.624
```

```
##
## Call:
## lm(formula = LogSpecies ~ LogArea + Elev, data = selected_data)
##
## Coefficients:
## (Intercept)      LogArea         Elev
##   2.2118578    0.0684940   -0.0001702
```

#Appendix 3

```
## Start:  AIC=-55.49
## LogSpecies ~ LogMeanDepth + LogCond + Elev + LogLat + Long +
##     LogNLakes + Photo + LogArea
##
##                Df Sum of Sq    RSS     AIC
## - Long          1   0.00388 2.5934 -57.446
## - LogLat        1   0.03174 2.6213 -57.126
## - LogNLakes     1   0.03863 2.6282 -57.047
## - LogMeanDepth  1   0.09466 2.6842 -56.414
## - LogCond       1   0.11994 2.7095 -56.133
## <none>                      2.5896 -55.491
## - Elev          1   0.27629 2.8659 -54.450
## - Photo         1   0.30776 2.8973 -54.122
## - LogArea       1   0.43999 3.0295 -52.784
##
## Step:  AIC=-57.45
## LogSpecies ~ LogMeanDepth + LogCond + Elev + LogLat + LogNLakes +
##     Photo + LogArea
##
##                Df Sum of Sq    RSS     AIC
## - LogLat        1   0.02967 2.6231 -59.105
## - LogNLakes     1   0.03519 2.6286 -59.042
## - LogMeanDepth  1   0.09481 2.6883 -58.369
## - LogCond       1   0.11680 2.7102 -58.125
## <none>                      2.5934 -57.446
## - Elev          1   0.27409 2.8675 -56.432
## - Photo         1   0.32827 2.9217 -55.871
## - LogArea       1   0.48638 3.0798 -54.290
##
## Step:  AIC=-59.11
## LogSpecies ~ LogMeanDepth + LogCond + Elev + LogNLakes + Photo +
##     LogArea
##
##                Df Sum of Sq    RSS     AIC
## - LogMeanDepth  1   0.07184 2.6950 -60.294
## - LogNLakes     1   0.11054 2.7336 -59.867
## - LogCond       1   0.12469 2.7478 -59.712
## <none>                      2.6231 -59.105
## - Photo         1   0.29883 2.9219 -57.868
## - Elev          1   0.42025 3.0434 -56.647
## - LogArea       1   0.49591 3.1190 -55.910
##
## Step:  AIC=-60.29
## LogSpecies ~ LogCond + Elev + LogNLakes + Photo + LogArea
##
##             Df Sum of Sq    RSS     AIC
## - LogNLakes  1   0.12523 2.8202 -60.932
## - LogCond    1   0.16620 2.8612 -60.499
## <none>                   2.6950 -60.294
## - Photo      1   0.29812 2.9931 -59.147
## - Elev       1   0.38783 3.0828 -58.261
## - LogArea    1   2.13821 4.8332 -44.771
##
## Step:  AIC=-60.93
## LogSpecies ~ LogCond + Elev + Photo + LogArea
##
##           Df Sum of Sq    RSS     AIC
## - LogCond  1    0.1904 3.0106 -60.971
## <none>                 2.8202 -60.932
## - Photo    1    0.2359 3.0561 -60.521
## - Elev     1    0.5585 3.3787 -57.511
## - LogArea  1    3.3720 6.1922 -39.337
##
## Step:  AIC=-60.97
## LogSpecies ~ Elev + Photo + LogArea
##
##           Df Sum of Sq    RSS     AIC
## - Photo    1    0.0756 3.0862 -62.228
## <none>                 3.0106 -60.971
## - Elev     1    0.4746 3.4852 -58.580
## - LogArea  1    3.2458 6.2564 -41.028
##
## Step:  AIC=-62.23
## LogSpecies ~ Elev + LogArea
##
##           Df Sum of Sq    RSS     AIC
## <none>                 3.0862 -62.228
## - Elev     1    0.7520 3.8382 -57.686
## - LogArea  1    3.1863 6.2725 -42.951
```

```
## 
## Call:
## lm(formula = LogSpecies ~ Elev + LogArea, data = selected_data)
## 
## Coefficients:
## (Intercept)         Elev      LogArea  
##   2.2118578   -0.0001702    0.0684940
```

#Appendix 4

```
## Start:  AIC=-42.88
## LogSpecies ~ LogMeanDepth + LogCond + Elev + LogLat + Long + 
##     LogNLakes + Photo + LogArea
## 
##                Df Sum of Sq    RSS     AIC
## - Long          1   0.00388 2.5934 -46.237
## - LogLat        1   0.03174 2.6213 -45.916
## - LogNLakes     1   0.03863 2.6282 -45.837
## - LogMeanDepth  1   0.09466 2.6842 -45.205
## - LogCond       1   0.11994 2.7095 -44.923
## - Elev          1   0.27629 2.8659 -43.240
## - Photo         1   0.30776 2.8973 -42.913
## <none>                      2.5896 -42.881
## - LogArea       1   0.43999 3.0295 -41.574
## 
## Step:  AIC=-46.24
## LogSpecies ~ LogMeanDepth + LogCond + Elev + LogLat + LogNLakes + 
##     Photo + LogArea
## 
##                Df Sum of Sq    RSS     AIC
## - LogLat        1   0.02967 2.6231 -49.297
## - LogNLakes     1   0.03519 2.6286 -49.234
## - LogMeanDepth  1   0.09481 2.6883 -48.561
## - LogCond       1   0.11680 2.7102 -48.316
## - Elev          1   0.27409 2.8675 -46.624
## <none>                      2.5934 -46.237
## - Photo         1   0.32827 2.9217 -46.063
## - LogArea       1   0.48638 3.0798 -44.481
## 
## Step:  AIC=-49.3
## LogSpecies ~ LogMeanDepth + LogCond + Elev + LogNLakes + Photo + 
##     LogArea
## 
##                Df Sum of Sq    RSS     AIC
## - LogMeanDepth  1   0.07184 2.6950 -51.887
## - LogNLakes     1   0.11054 2.7336 -51.460
## - LogCond       1   0.12469 2.7478 -51.305
## - Photo         1   0.29883 2.9219 -49.461
## <none>                      2.6231 -49.297
## - Elev          1   0.42025 3.0434 -48.240
## - LogArea       1   0.49591 3.1190 -47.503
## 
## Step:  AIC=-51.89
## LogSpecies ~ LogCond + Elev + LogNLakes + Photo + LogArea
```

```
##
##             Df Sum of Sq    RSS     AIC
## - LogNLakes  1   0.12523 2.8202 -53.926
## - LogCond    1   0.16620 2.8612 -53.493
## - Photo      1   0.29812 2.9931 -52.141
## <none>                   2.6950 -51.887
## - Elev       1   0.38783 3.0828 -51.255
## - LogArea    1   2.13821 4.8332 -37.765
##
## Step:  AIC=-53.93
## LogSpecies ~ LogCond + Elev + Photo + LogArea
##
##           Df Sum of Sq    RSS     AIC
## - LogCond  1    0.1904 3.0106 -55.367
## - Photo    1    0.2359 3.0561 -54.917
## <none>                 2.8202 -53.926
## - Elev     1    0.5585 3.3787 -51.906
## - LogArea  1    3.3720 6.1922 -33.733
##
## Step:  AIC=-55.37
## LogSpecies ~ Elev + Photo + LogArea
##
##           Df Sum of Sq    RSS     AIC
## - Photo    1    0.0756 3.0862 -58.024
## <none>                 3.0106 -55.367
## - Elev     1    0.4746 3.4852 -54.376
## - LogArea  1    3.2458 6.2564 -36.824
##
## Step:  AIC=-58.02
## LogSpecies ~ Elev + LogArea
##
##           Df Sum of Sq    RSS     AIC
## <none>                 3.0862 -58.024
## - Elev     1    0.7520 3.8382 -54.884
## - LogArea  1    3.1863 6.2725 -40.148
```

```
##
## Call:
## lm(formula = LogSpecies ~ Elev + LogArea, data = selected_data)
##
## Coefficients:
## (Intercept)         Elev      LogArea
##   2.2118578   -0.0001702    0.0684940
```

#Appendix 5

```
## Start:  AIC=-30.77
## LogSpecies ~ 1
##
##                Df Sum of Sq     RSS     AIC
## + LogArea       1    6.2251  3.8382 -57.686
## + LogMeanDepth  1    4.5865  5.4768 -47.021
## + Elev          1    3.7908  6.2725 -42.951
## + LogNLakes     1    2.9877  7.0755 -39.337
## + Photo         1    0.8108  9.2524 -31.289
## <none>                      10.0632 -30.769
## + LogLat        1    0.5671  9.4961 -30.509
## + LogCond       1    0.5045  9.5587 -30.312
## + Long          1    0.1144  9.9488 -29.112
##
## Step:  AIC=-57.69
## LogSpecies ~ LogArea
##
##                Df Sum of Sq     RSS     AIC
## + Elev          1    0.7520  3.0862 -62.228
## + Photo         1    0.3529  3.4852 -58.580
## <none>                       3.8382 -57.686
## + LogNLakes     1    0.1723  3.6659 -57.064
## + LogLat        1    0.1404  3.6978 -56.804
## + LogCond       1    0.0251  3.8131 -55.883
## + LogMeanDepth  1    0.0206  3.8176 -55.848
## + Long          1    0.0180  3.8201 -55.827
## - LogArea       1    6.2251 10.0632 -30.769
##
## Step:  AIC=-62.23
## LogSpecies ~ LogArea + Elev
##
##                Df Sum of Sq    RSS     AIC
## <none>                      3.0862 -62.228
## + LogMeanDepth  1    0.1019 2.9842 -61.236
## + LogNLakes     1    0.0817 3.0045 -61.033
## + Photo         1    0.0756 3.0106 -60.971
## + LogCond       1    0.0300 3.0561 -60.521
## + LogLat        1    0.0117 3.0744 -60.342
## + Long          1    0.0001 3.0861 -60.228
## - Elev          1    0.7520 3.8382 -57.686
## - LogArea       1    3.1863 6.2725 -42.951
```

```
##
## Call:
## lm(formula = LogSpecies ~ LogArea + Elev, data = selected_data)
##
## Coefficients:
## (Intercept)      LogArea         Elev
##   2.2118578    0.0684940   -0.0001702
```

#Appendix 6

```
## Start:  AIC=-29.37
## LogSpecies ~ 1
##
##                Df Sum of Sq     RSS     AIC
## + LogArea       1    6.2251  3.8382 -54.884
## + LogMeanDepth  1    4.5865  5.4768 -44.218
## + Elev          1    3.7908  6.2725 -40.148
## + LogNLakes     1    2.9877  7.0755 -36.534
## <none>                      10.0632 -29.368
## + Photo         1    0.8108  9.2524 -28.487
## + LogLat        1    0.5671  9.4961 -27.707
## + LogCond       1    0.5045  9.5587 -27.510
## + Long          1    0.1144  9.9488 -26.310
##
## Step:  AIC=-54.88
## LogSpecies ~ LogArea
##
##                Df Sum of Sq     RSS     AIC
## + Elev          1    0.7520  3.0862 -58.024
## <none>                       3.8382 -54.884
## + Photo         1    0.3529  3.4852 -54.376
## + LogNLakes     1    0.1723  3.6659 -52.860
## + LogLat        1    0.1404  3.6978 -52.600
## + LogCond       1    0.0251  3.8131 -51.679
## + LogMeanDepth  1    0.0206  3.8176 -51.644
## + Long          1    0.0180  3.8201 -51.624
## - LogArea       1    6.2251 10.0632 -29.368
##
## Step:  AIC=-58.02
## LogSpecies ~ LogArea + Elev
##
##                Df Sum of Sq    RSS     AIC
## <none>                      3.0862 -58.024
## + LogMeanDepth  1    0.1019 2.9842 -55.631
## + LogNLakes     1    0.0817 3.0045 -55.428
## + Photo         1    0.0756 3.0106 -55.367
## + LogCond       1    0.0300 3.0561 -54.917
## - Elev          1    0.7520 3.8382 -54.884
## + LogLat        1    0.0117 3.0744 -54.738
## + Long          1    0.0001 3.0861 -54.624
## - LogArea       1    3.1863 6.2725 -40.148
```

```
##
## Call:
## lm(formula = LogSpecies ~ LogArea + Elev, data = selected_data)
##
## Coefficients:
## (Intercept)      LogArea         Elev
##   2.2118578    0.0684940   -0.0001702
```

#Appendix 7

```
## Start:  AIC=-55.49
## LogSpecies ~ LogMeanDepth + LogCond + Elev + LogLat + Long +
##     LogNLakes + Photo + LogArea
##
##                Df Sum of Sq    RSS     AIC
## - Long          1   0.00388 2.5934 -57.446
## - LogLat        1   0.03174 2.6213 -57.126
## - LogNLakes     1   0.03863 2.6282 -57.047
## - LogMeanDepth  1   0.09466 2.6842 -56.414
## - LogCond       1   0.11994 2.7095 -56.133
## <none>                      2.5896 -55.491
## - Elev          1   0.27629 2.8659 -54.450
## - Photo         1   0.30776 2.8973 -54.122
## - LogArea       1   0.43999 3.0295 -52.784
##
## Step:  AIC=-57.45
## LogSpecies ~ LogMeanDepth + LogCond + Elev + LogLat + LogNLakes +
##     Photo + LogArea
##
##                Df Sum of Sq    RSS     AIC
## - LogLat        1   0.02967 2.6231 -59.105
## - LogNLakes     1   0.03519 2.6286 -59.042
## - LogMeanDepth  1   0.09481 2.6883 -58.369
## - LogCond       1   0.11680 2.7102 -58.125
## <none>                      2.5934 -57.446
## - Elev          1   0.27409 2.8675 -56.432
## - Photo         1   0.32827 2.9217 -55.871
## + Long          1   0.00388 2.5896 -55.491
## - LogArea       1   0.48638 3.0798 -54.290
##
## Step:  AIC=-59.11
## LogSpecies ~ LogMeanDepth + LogCond + Elev + LogNLakes + Photo +
##     LogArea
##
##                Df Sum of Sq    RSS     AIC
## - LogMeanDepth  1   0.07184 2.6950 -60.294
## - LogNLakes     1   0.11054 2.7336 -59.867
## - LogCond       1   0.12469 2.7478 -59.712
## <none>                      2.6231 -59.105
## - Photo         1   0.29883 2.9219 -57.868
## + LogLat        1   0.02967 2.5934 -57.446
## + Long          1   0.00182 2.6213 -57.126
## - Elev          1   0.42025 3.0434 -56.647
## - LogArea       1   0.49591 3.1190 -55.910
##
## Step:  AIC=-60.29
## LogSpecies ~ LogCond + Elev + LogNLakes + Photo + LogArea
##
##                Df Sum of Sq    RSS     AIC
## - LogNLakes     1   0.12523 2.8202 -60.932
## - LogCond       1   0.16620 2.8612 -60.499
## <none>                      2.6950 -60.294
## - Photo         1   0.29812 2.9931 -59.147
## + LogMeanDepth  1   0.07184 2.6231 -59.105
## + LogLat        1   0.00671 2.6882 -58.369
## + Long          1   0.00284 2.6921 -58.326
## - Elev          1   0.38783 3.0828 -58.261
## - LogArea       1   2.13821 4.8332 -44.771
##
## Step:  AIC=-60.93
## LogSpecies ~ LogCond + Elev + Photo + LogArea
##
##                Df Sum of Sq    RSS     AIC
## - LogCond       1    0.1904 3.0106 -60.971
## <none>                      2.8202 -60.932
## - Photo         1    0.2359 3.0561 -60.521
## + LogNLakes     1    0.1252 2.6950 -60.294
## + LogMeanDepth  1    0.0865 2.7336 -59.867
## + LogLat        1    0.0623 2.7579 -59.602
## + Long          1    0.0035 2.8167 -58.969
## - Elev          1    0.5585 3.3787 -57.511
## - LogArea       1    3.3720 6.1922 -39.337
##
## Step:  AIC=-60.97
## LogSpecies ~ Elev + Photo + LogArea
##
##                Df Sum of Sq    RSS     AIC
## - Photo         1    0.0756 3.0862 -62.228
## <none>                      3.0106 -60.971
## + LogCond       1    0.1904 2.8202 -60.932
## + LogNLakes     1    0.1495 2.8612 -60.499
## + LogMeanDepth  1    0.1367 2.8740 -60.365
## + LogLat        1    0.0719 2.9387 -59.696
## + Long          1    0.0124 2.9982 -59.095
## - Elev          1    0.4746 3.4852 -58.580
## - LogArea       1    3.2458 6.2564 -41.028
##
## Step:  AIC=-62.23
## LogSpecies ~ Elev + LogArea
##
##                Df Sum of Sq    RSS     AIC
## <none>                      3.0862 -62.228
## + LogMeanDepth  1    0.1019 2.9842 -61.236
## + LogNLakes     1    0.0817 3.0045 -61.033
## + Photo         1    0.0756 3.0106 -60.971
## + LogCond       1    0.0300 3.0561 -60.521
## + LogLat        1    0.0117 3.0744 -60.342
## + Long          1    0.0001 3.0861 -60.228
## - Elev          1    0.7520 3.8382 -57.686
## - LogArea       1    3.1863 6.2725 -42.951
```

```
##
## Call:
## lm(formula = LogSpecies ~ Elev + LogArea, data = selected_data)
##
## Coefficients:
## (Intercept)         Elev      LogArea
##   2.2118578   -0.0001702    0.0684940
```

#Appendix 8

```
## Start:  AIC=-42.88
## LogSpecies ~ LogMeanDepth + LogCond + Elev + LogLat + Long +
##     LogNLakes + Photo + LogArea
##
##                Df Sum of Sq    RSS     AIC
## - Long          1   0.00388 2.5934 -46.237
## - LogLat        1   0.03174 2.6213 -45.916
## - LogNLakes     1   0.03863 2.6282 -45.837
## - LogMeanDepth  1   0.09466 2.6842 -45.205
## - LogCond       1   0.11994 2.7095 -44.923
## - Elev          1   0.27629 2.8659 -43.240
## - Photo         1   0.30776 2.8973 -42.913
## <none>                      2.5896 -42.881
## - LogArea       1   0.43999 3.0295 -41.574
##
## Step:  AIC=-46.24
## LogSpecies ~ LogMeanDepth + LogCond + Elev + LogLat + LogNLakes +
##     Photo + LogArea
##
##                Df Sum of Sq    RSS     AIC
## - LogLat        1   0.02967 2.6231 -49.297
## - LogNLakes     1   0.03519 2.6286 -49.234
## - LogMeanDepth  1   0.09481 2.6883 -48.561
## - LogCond       1   0.11680 2.7102 -48.316
## - Elev          1   0.27409 2.8675 -46.624
## <none>                      2.5934 -46.237
## - Photo         1   0.32827 2.9217 -46.063
## - LogArea       1   0.48638 3.0798 -44.481
## + Long          1   0.00388 2.5896 -42.881
##
## Step:  AIC=-49.3
## LogSpecies ~ LogMeanDepth + LogCond + Elev + LogNLakes + Photo +
##     LogArea
##
##                Df Sum of Sq    RSS     AIC
## - LogMeanDepth  1   0.07184 2.6950 -51.887
## - LogNLakes     1   0.11054 2.7336 -51.460
## - LogCond       1   0.12469 2.7478 -51.305
## - Photo         1   0.29883 2.9219 -49.461
## <none>                      2.6231 -49.297
## - Elev          1   0.42025 3.0434 -48.240
## - LogArea       1   0.49591 3.1190 -47.503
## + LogLat        1   0.02967 2.5934 -46.237
## + Long          1   0.00182 2.6213 -45.916
##
## Step:  AIC=-51.89
## LogSpecies ~ LogCond + Elev + LogNLakes + Photo + LogArea
##
##                Df Sum of Sq    RSS     AIC
## - LogNLakes     1   0.12523 2.8202 -53.926
## - LogCond       1   0.16620 2.8612 -53.493
## - Photo         1   0.29812 2.9931 -52.141
## <none>                      2.6950 -51.887
## - Elev          1   0.38783 3.0828 -51.255
## + LogMeanDepth  1   0.07184 2.6231 -49.297
## + LogLat        1   0.00671 2.6882 -48.561
## + Long          1   0.00284 2.6921 -48.518
## - LogArea       1   2.13821 4.8332 -37.765
##
```

```
## Step:  AIC=-53.93
## LogSpecies ~ LogCond + Elev + Photo + LogArea
##
##                Df Sum of Sq    RSS     AIC
## - LogCond       1    0.1904 3.0106 -55.367
## - Photo         1    0.2359 3.0561 -54.917
## <none>                      2.8202 -53.926
## - Elev          1    0.5585 3.3787 -51.906
## + LogNLakes     1    0.1252 2.6950 -51.887
## + LogMeanDepth  1    0.0865 2.7336 -51.460
## + LogLat        1    0.0623 2.7579 -51.195
## + Long          1    0.0035 2.8167 -50.562
## - LogArea       1    3.3720 6.1922 -33.733
##
## Step:  AIC=-55.37
## LogSpecies ~ Elev + Photo + LogArea
##
##                Df Sum of Sq    RSS     AIC
## - Photo         1    0.0756 3.0862 -58.024
## <none>                      3.0106 -55.367
## - Elev          1    0.4746 3.4852 -54.376
## + LogCond       1    0.1904 2.8202 -53.926
## + LogNLakes     1    0.1495 2.8612 -53.493
## + LogMeanDepth  1    0.1367 2.8740 -53.359
## + LogLat        1    0.0719 2.9387 -52.690
## + Long          1    0.0124 2.9982 -52.089
## - LogArea       1    3.2458 6.2564 -36.824
##
## Step:  AIC=-58.02
## LogSpecies ~ Elev + LogArea
##
##                Df Sum of Sq    RSS     AIC
## <none>                      3.0862 -58.024
## + LogMeanDepth  1    0.1019 2.9842 -55.631
## + LogNLakes     1    0.0817 3.0045 -55.428
## + Photo         1    0.0756 3.0106 -55.367
## + LogCond       1    0.0300 3.0561 -54.917
## - Elev          1    0.7520 3.8382 -54.884
## + LogLat        1    0.0117 3.0744 -54.738
## + Long          1    0.0001 3.0861 -54.624
## - LogArea       1    3.1863 6.2725 -40.148
```

```
##
## Call:
## lm(formula = LogSpecies ~ Elev + LogArea, data = selected_data)
##
## Coefficients:
## (Intercept)         Elev      LogArea
##   2.2118578   -0.0001702    0.0684940
```

#Appendix9

```
## Analysis of Variance Table
```

```
##
## Model 1: LogSpecies ~ LogArea + Elev
## Model 2: LogSpecies ~ LogArea * Elev
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     27 3.0862
## 2     26 3.0652  1  0.020997 0.1781 0.6765
```

#Appendix10

```
##
## Call:
## lm(formula = LogSpecies ~ LogArea + Elev, data = selected_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81430 -0.23671  0.03876  0.21517  0.44711
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.212e+00  1.081e-01  20.457  < 2e-16 ***
## LogArea      6.849e-02  1.297e-02   5.280 1.44e-05 ***
## Elev        -1.702e-04  6.637e-05  -2.565   0.0162 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3381 on 27 degrees of freedom
## Multiple R-squared:  0.6933, Adjusted R-squared:  0.6706
## F-statistic: 30.52 on 2 and 27 DF,  p-value: 1.176e-07
```