Crustaceans Count Analysis

Samuel Gadkar (gadkar.2)

2024-12-04

Introduction

Over time, biodiversity has been a prominent issue in society. Specifically, there have been numerous concerns about whether additional mass extinctions are occurring right now, even ones we may need to be made aware of. Lake crustaceans are one of the species that researchers are concerned with. To get ahead of the curve on this issue, we have been allotted data on 30 different lakes across the Earth. With this data, we aim to find a model that can predict the numerous crustaceans in a lake so researchers can analyze their concerns with mass extinctions further.

Data

The data includes nine different data fields about each of the 30 lakes. The following are included: number of crustacean species (Species), mean lake depth in meters (MeanDepth), specific conductance (a measure of mineral content) in microsiemans (Cond), lake elevation in meters (Elev), latitude degrees north (Lat), longitude degrees west (Long), number of lakes within 20 kilometers (NLakes), rate of photosynthesis measured by using C^{14} (Photo), and surface area of lake in hectares (Area). (Appendix - Section 1)

Starting with the number of crustaceans, the data for this variable ranged from as low as 3 to a high of 30, with an average number of crustacean species of approximately 10.5. As for mean lake depth, this data was more variable, with a minimum of a little more than 0 meters and a maximum of 148 meters. However, the elevation tended to be on the lower end, as the median value was only 8.5 meters. Moving onto the measure of conductance, this variable was even more distributed. It ranged from 8 microsiemans to a maximum of 1,600 microsiemans. The median value for this variable was also on the lower end of the range, with only 67.5 microsiemans of mineral content.

Lake elevation also had much variability, with a minimum of .5 meters and a maximum of more than 3400 meters. The median elevation was closer to the minimum, about 280 meters high. The variable of latitude degrees north is more normally spread out with a minimum of 28 degrees north and a maximum of a little more than 70 degrees north. The mean of this variable is about 47 degrees north. Longitude is similar to the distribution, ranging from about 72 to 157 degrees west and having an average of about 106 degrees.

Three more variables were provided for our analysis. There was an enormous spread in the data regarding the number of lakes within 20km of the lake analyzed. This value ranged from 5 to nearly 9000, meaning a median of only 45 lakes. The rate of photosynthesis had a similar spread in the distribution with a min of -18.60 C^{14} to 1500 C^{14}, having a median of approximately 230 C^{14}. The final variable was the surface area of the lake. This ranged from .008 hectares to more than 8 million hectares, yet the median was only 12 hectares.

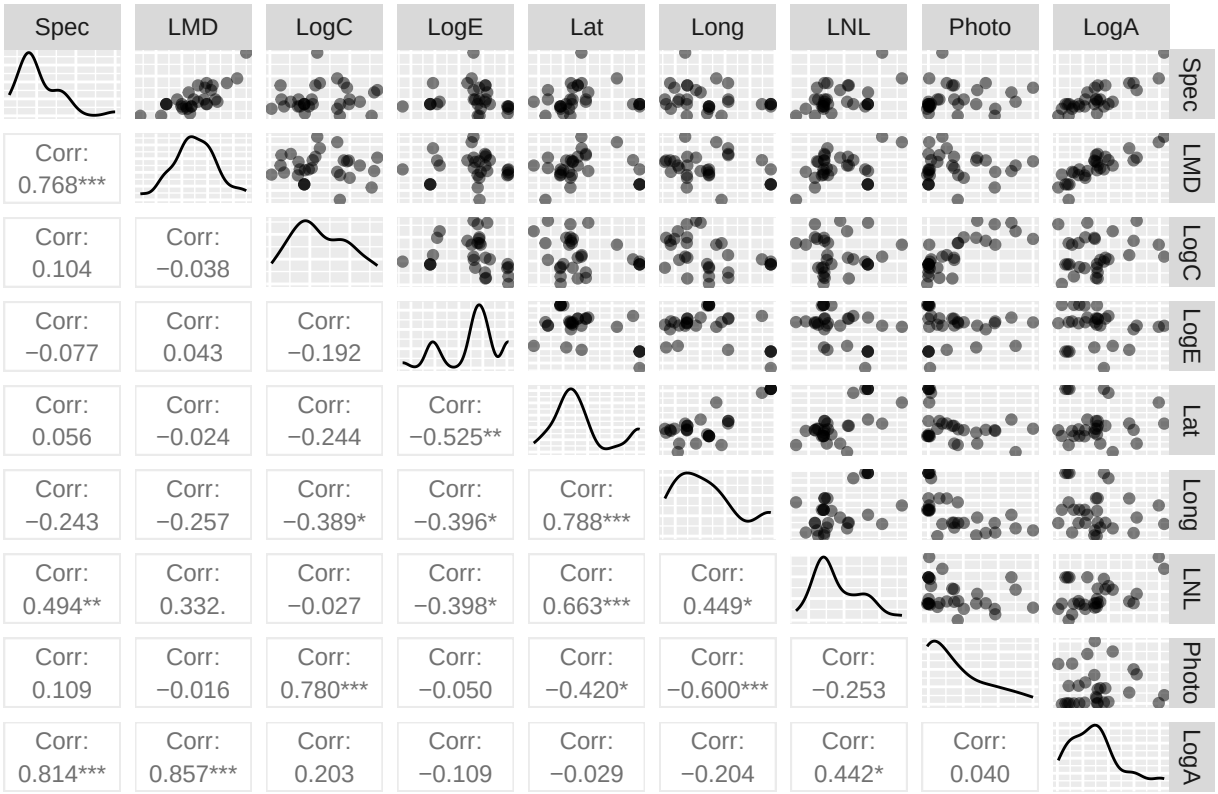
Exploratory Data Analysis

This analysis aims to uncover the model, using these variables, which is best able to predict the number of crustaceans in a lake. To begin exploring the data, I first created a scatterplot matrix. A scatterplot matrix plots each possible combination of pairs of variables. For example, it would plot 'Species' with 'MeanDepth', then 'Species' with 'Cond', then 'Species' with 'Elev', and so on. It would then repeat this step for each variable in the dataset. (Appendix - Section 2)

When looking at the relevant scatterplots along the first row and correlations across the first column, some noticeable plots stray from a roughly linear shape. Following this, I created eight scatterplots where I placed 'Species' on the y-axis and one of the additional eight variables on the x-axis. From there, it was clear that some plots were far from a linear model. Thus, I took the log of the variable on the x-axis for those I felt it was necessary. Doing so would shift the graph to make it look more linear but does not take away from the understanding of the data. I decided on the following variables to be transformed: 'MeanDepth', 'Cond', 'Elev', 'NLakes', and 'Area'. (Appendix - Section 3)

Following this step, I decided to create a new scatterplot matrix that displays all the plots with the necessary changes in variables such that their plots with 'Species' look closer to a linear shape. The updated scatterplot matrix is below:

Updated Scatterplot Matrix Including Log Transformations



Main Effects Model Selection

When finding the model that would best represent and predict the number of crustaceans in a lake, I began with one of the simpler models, which could be a single predictor model. Three variables were very beneficial when going through the eight possible single predictor models: 'LogMeanDepth', 'LogNLakes', and 'LogArea'. Out of the three, 'LogArea' seemed to be the best variable for this model as the p-value for this predictor

was practically zero and the lowest among all eight predictors. This means that out of all possible single predictor models, we were confident that adding this variable would be beneficial and meaningful to the model. It is logical to say that the surface area of a lake influences the number of crustaceans. Specifically, more surface area would likely lead to more crustaceans. (Appendix - Section 4)

The next step was to find the best two predictor models to estimate the number of crustaceans. When doing this, the best model included 'LogArea' and 'LogMeanDepth' as its predictors. It also makes logical sense to add 'LogMeanDepth' to the model. The larger the mean depth of a lake, the bigger the lake, and thus, more area for the crustaceans to live. In addition, when looking back at the single predictor models from earlier, both 'LogMeanDepth' and 'LogArea' had p-values of practically zero. Plus, when the two predictor models were formed, these variables created the best estimate of 'Species'. (Appendix - Section 5)

The three predictors model I decided to look into was the final critical main effects model. This model included 'LogMeanDepth', 'Long', and 'LogNLakes'. Additionally, using BIC and Stepwise Regression, a model including these three predictors is the "best" model out of any possible model for predicting the number of crustaceans. Despite this, there are some apparent flaws with this assumption. Firstly, it makes no sense to include longitude as that has practically no impact on the habitat and thus should not be included in the model. In addition, when we test to see if this model change is beneficial, there isn't enough evidence to say it is. I also explored different three predictor models which included 'LogMeanDepth', 'LogArea', and some other predictor. When analyzing these additional linear models, the current two predictor models needed to be improved or not a logical predictor for the number of species. Thus, when selecting a primary effects model, the two predictor models, including 'LogArea' and 'LogMeanDepth', provide the most accurate and straightforward interpretation. I have included this model below:

$$\begin{aligned} Species &= \beta_0 + \beta_{LMD} * LogMeanDepth + \beta_{LA} * LogArea \\ &= 7.9780 + 0.8417 * LogMeanDepth + 0.6270 * LogArea \end{aligned}$$

Interaction Model Selection

After finalizing a primary effects model, I decided to look into potential interactions between the variables. An interaction term is when the value of one variable impacts another variable. There is only one possible interaction term that can come from the model above: the interaction between 'LogMeanDepth' and 'LogArea'. I performed a test to see if this additional term was beneficial to the model, and based on this, I can say that I am pretty confident that there is an interaction between the two variables. Specifically, the p-value for adding this to the model is slightly more than .02. Thus, adding this to the model would provide a more accurate estimation of the number of crustaceans. The new and improved model is as follows: (Appendix - Section 6)

$$\begin{aligned} Species &= \beta_0 + \beta_{LMD} * LogMeanDepth + \beta_{LA} * LogArea + \beta_{LMD:LA} * (LogMeanDepth * LogArea) \\ &= 7.53204 + 1.00787 * LogMeanDepth + 0.20519 * LogArea + 0.13149 * (LogMeanDepth * LogArea) \end{aligned}$$

Understanding The Final Model

Assumptions About Final Model:

When it comes to linear models, there are certain assumptions that we must make to allow us to use this type of model. First, given the predictors, we must assume that the error terms are all independent. (Model Fit and Diagnostics)

Since the plotted points are mostly scattered randomly around $y = 0$, we can say that the errors are independent. The following assumption we must say is that the mean function is:

$$E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_1 * x_2) = \beta_0 + \beta_{LMD} * x_1 + \beta_{LA} * x_2 + \beta_{LMD:LA} * (x_1 * x_2)$$

The third assumption we must make about our model is that the variance of Y is the same at all values of $X_1 \cdots X_3$. To ensure this, we must look at the residual plot above. There must be primarily random scattering of the points, but we should also look at and ensure there is no funneling in the plot. We can look at the above residual plot and see there isn't funneling. Finally, we will also assume that the distribution of 'Y' is normal at each combination of 'X's'. (Model Fit and Diagnostics)

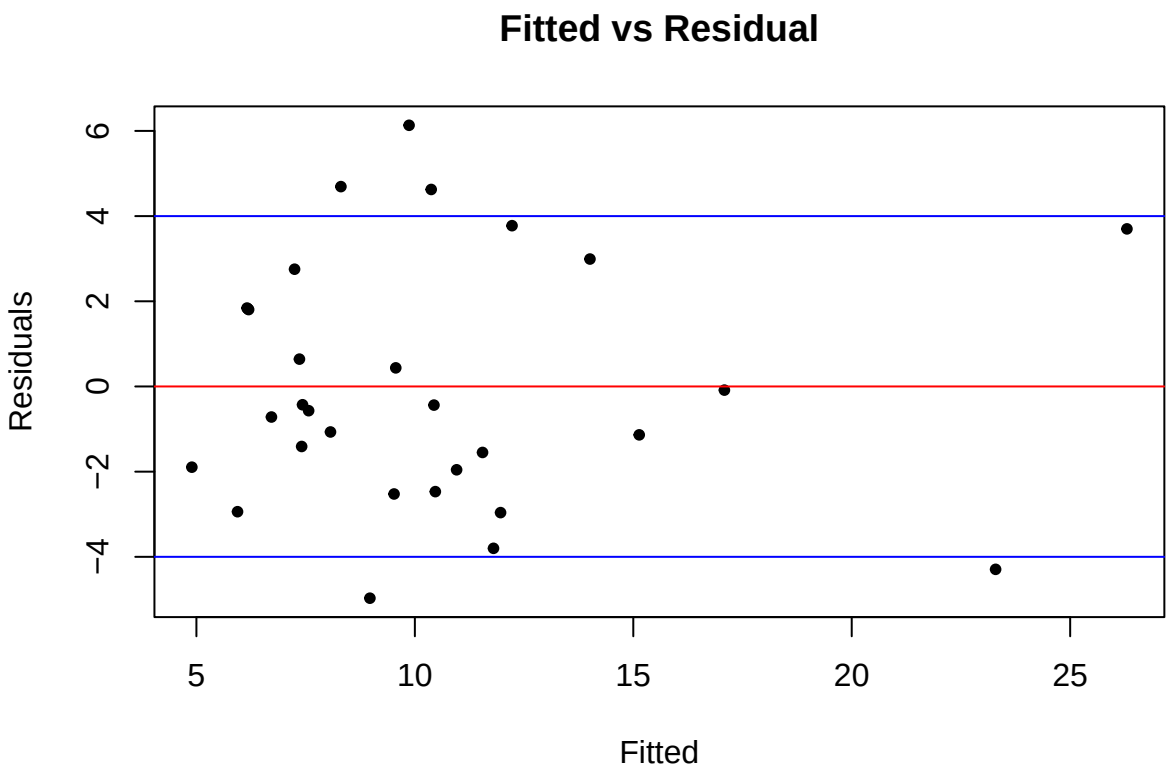
When looking at this new plot, we see that these points follow the linear line quite well. Our model satisfies all the necessary assumptions. Another way we could state this is as follows:

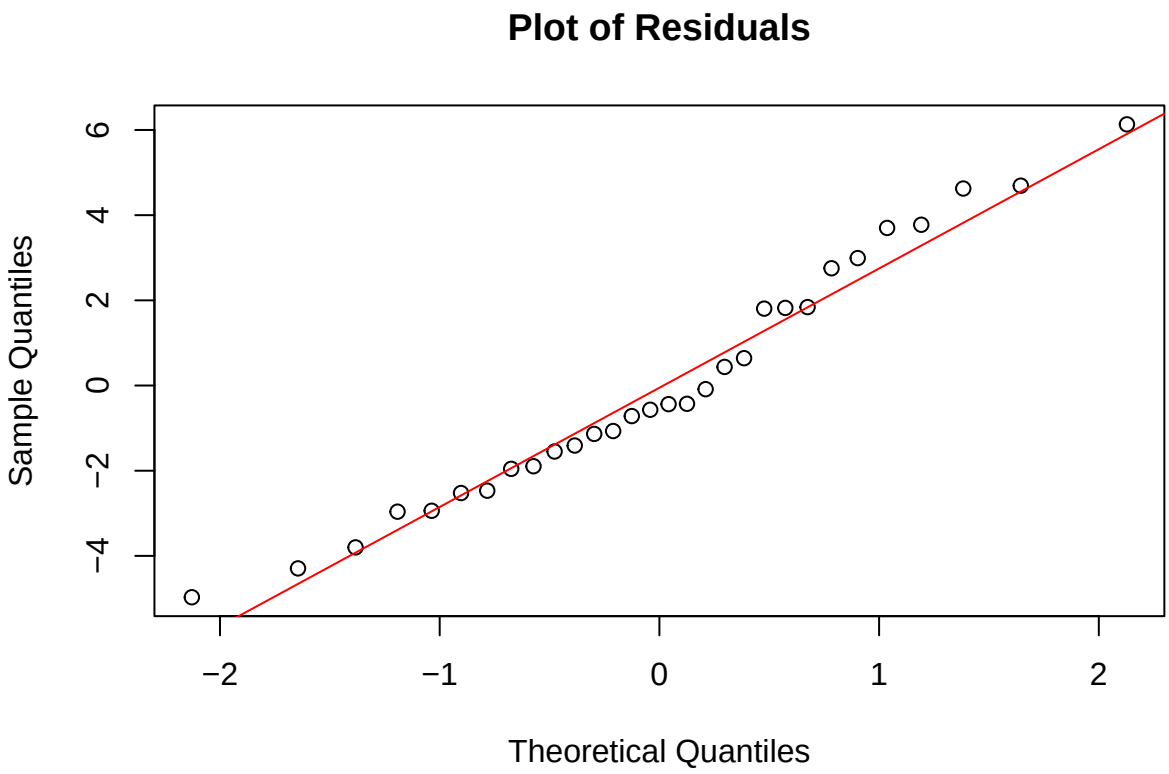
$$Y_i = \beta_0 + \beta_{LMD} X_{1i} + \beta_{LA} X_{2i} + \beta_{LMD:LA}(X_{1i} * X_{2i}) + e_i, \ where \ e_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \ and \ i = 1, \cdots, 30$$

It should be noted that X_1 is the natural log of the mean depth and X_2 is the natural log of the surface area.

Model Fit and Diagnostics

Look at the charts below for a different visual representation of the final linear model. It includes a visualization of the fitted values versus residuals, displaying how the residual plot seems to reveal mostly random around 0. Additionally, the second graph demonstrates very normal relationship, showing how our model seems to be an accurate predictor for the true number of species: (Appendix - Section 7)





Definition of Variables:

Species - The number of crustacean species which live in the lake

LogMeanDepth - The natural log of the average depth of the lake in meters

LogArea - The natural log of the surface area of the lake in hectares

Interaction Term (LogMeanDepth * LogArea) - The effect of the natural log of average depth on the number of crustaceans, factoring in the different levels of the natural log of surface area

Interpretation of Parameters:

$\beta_0 = 7.53204$ -> The estimated number of species given that LogMeanDepth and LogArea are 0, is about 7.53204 crustaceans.

$\beta_{LMD} = 1.00787$ -> A 1% increase in the average depth of the lake leads to approximately .01003 increase in the number of crustaceans.

$\beta_{LA} = .20519$ -> A 1% increase in the surface area of the lake leads to approximately .00204 increase in the number of crustaceans.

$\beta_{LMD:LA} = .13149$ -> When there is a 1% increase in both the average depth and surface area, then there is approximately an additional .000013 increase in the number of crustaceans.

Predicted Number of Crustaceans For Sample Lake

Now that I have finalized what I believe is the best model for predicting the number of crustaceans, it is time to test it on a sample lake. The lake provided has the following values for each variable:

MeanDepth = 153

Cond = 167

Lat = 46

Long = -3

Elev = 372

NLakes = 44

Photo = 263

Area = 58,000

Using the necessary values from the sample, my prediction values, confidence interval, and prediction interval for the number of crustacean species is as follows:

The predicted number of crustaceans for the sample lakes is approximately 22.10768.

I am 95% confident that the true value will fall between 18.19838 and 26.01698.

The predicted number of crustaceans for a new lake is approximately 22.10768.

I am 95% confident that the true value for a new lake will fall between 14.75091 and 29.46445.

Conclusion

Thus, after thorough research and analysis, I have concluded that the model, which is not only accurate but also relatively simple, is the following:

$$Species = 7.53204 + 1.00787 * LogMeanDepth + 0.20519 * LogArea + 0.13149 * (LogMeanDepth * LogArea)$$

Appendix

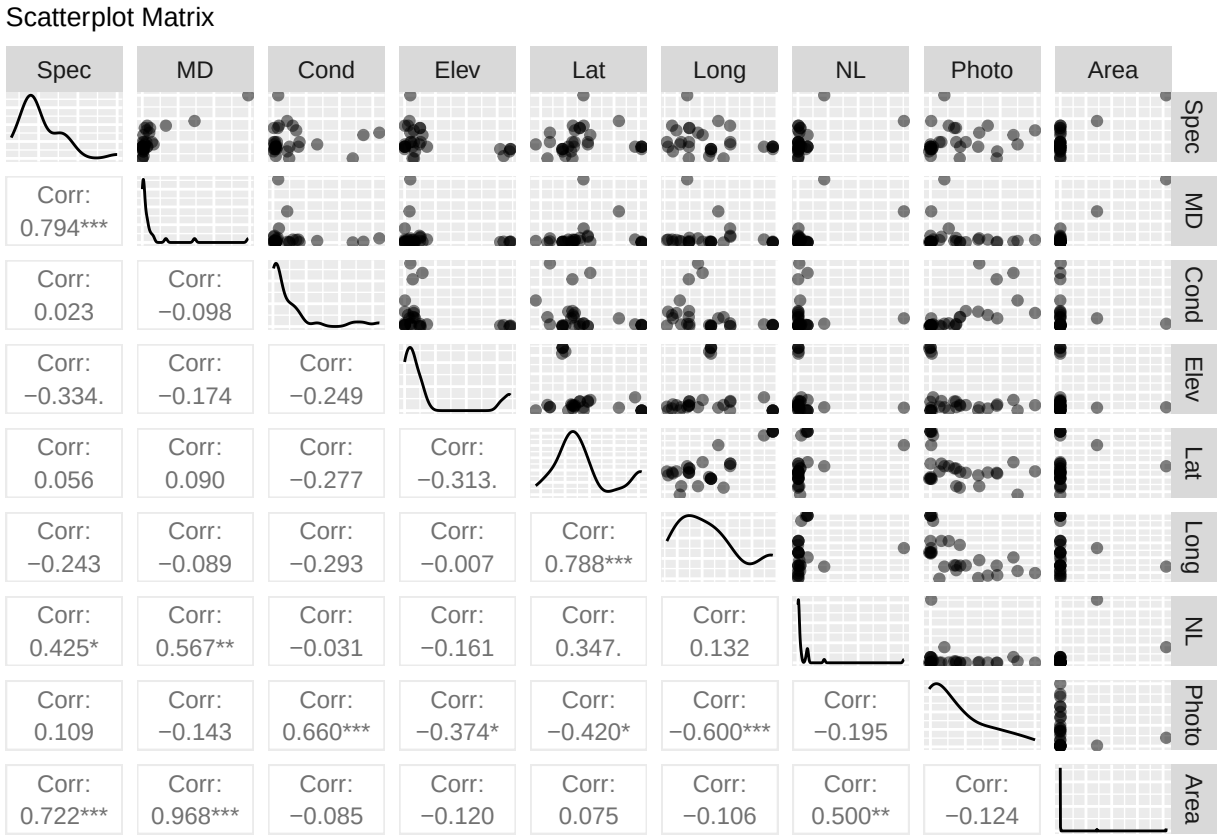
Section 1

```
summary(crust)
```

```
##      ...1              Species        MeanDepth            Cond
##  Length:30          Min.   : 3.00   Min.   :  0.040   Min.   :   8.00
##  Class :character   1st Qu.: 7.00   1st Qu.:  1.018   1st Qu.:  35.02
##  Mode  :character   Median : 8.50   Median :  2.300   Median :  67.50
##                     Mean   :10.43   Mean   : 12.031   Mean   : 264.17
##                     3rd Qu.:13.75   3rd Qu.:  6.950   3rd Qu.: 285.50
##                     Max.   :30.00   Max.   :148.000   Max.   :1600.00
##       Elev             Lat             Long            NLakes
##  Min.   :   0.5   Min.   :28.00   Min.   : 71.70   Min.   :   5.0
##  1st Qu.: 123.8   1st Qu.:39.27   1st Qu.: 88.33   1st Qu.:  35.0
##  Median : 279.5   Median :43.75   Median :103.00   Median :  45.0
##  Mean   : 757.7   Mean   :47.35   Mean   :106.08   Mean   : 546.7
##  3rd Qu.: 516.5   3rd Qu.:49.30   3rd Qu.:121.10   3rd Qu.: 265.8
##  Max.   :3433.0   Max.   :71.30   Max.   :156.70   Max.   :8805.0
##      Photo              Area
##  Min.   : -18.60   Min.   :      0
##  1st Qu.:   5.93   1st Qu.:      0
##  Median : 231.50   Median :     12
##  Mean   : 377.60   Mean   : 372274
##  3rd Qu.: 640.48   3rd Qu.:     76
##  Max.   :1500.00   Max.   :8240000
```

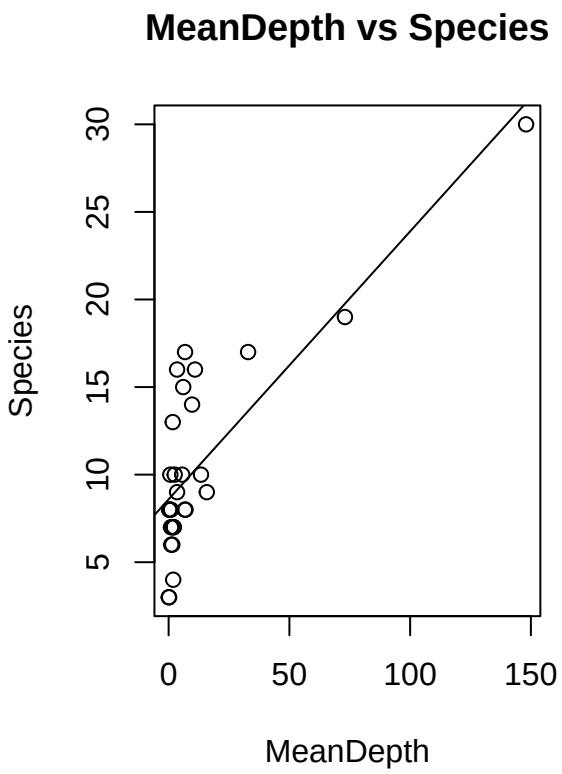
Section 2

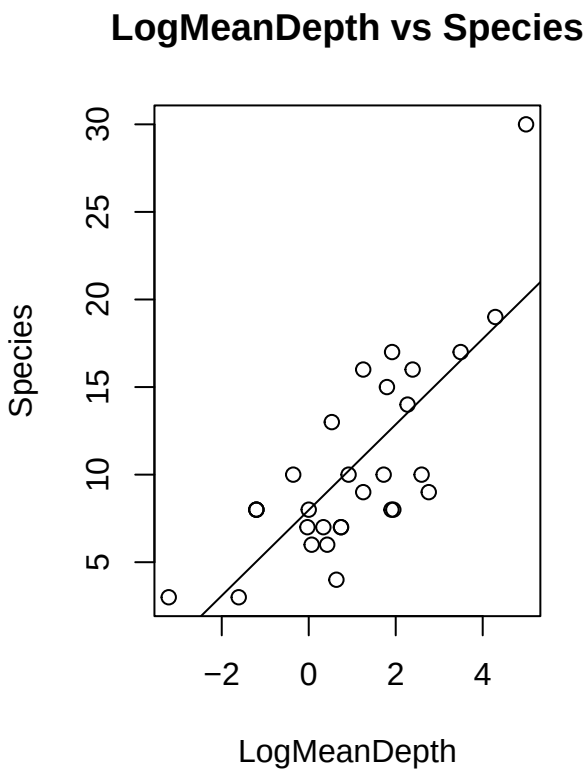
```
ggpairs(
  data = crust,
  columns = c("Species", "MeanDepth", "Cond", "Elev", "Lat", "Long", "NLakes", "Photo", "Area"),
  title = "Scatterplot Matrix",
  labeller = label_wrap_gen(width = 10),
  upper = list(continuous = wrap("points", alpha = 0.5)),
  lower = list(continuous = wrap("cor", size = 3)),
  columnLabels = c("Spec", "MD", "Cond", "Elev", "Lat", "Long", "NL", "Photo", "Area")
) +
theme(
  axis.text = element_blank(),
  axis.ticks = element_blank(),
  plot.title = element_text(size = 10)
)
```



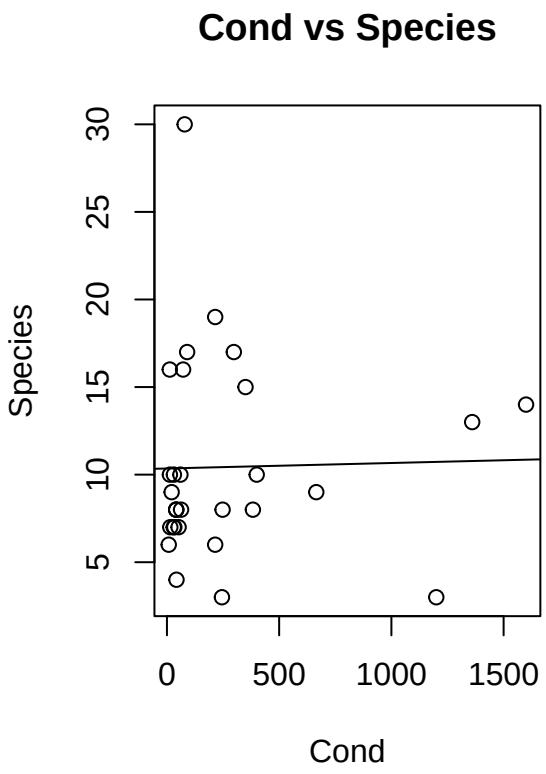
Section 3

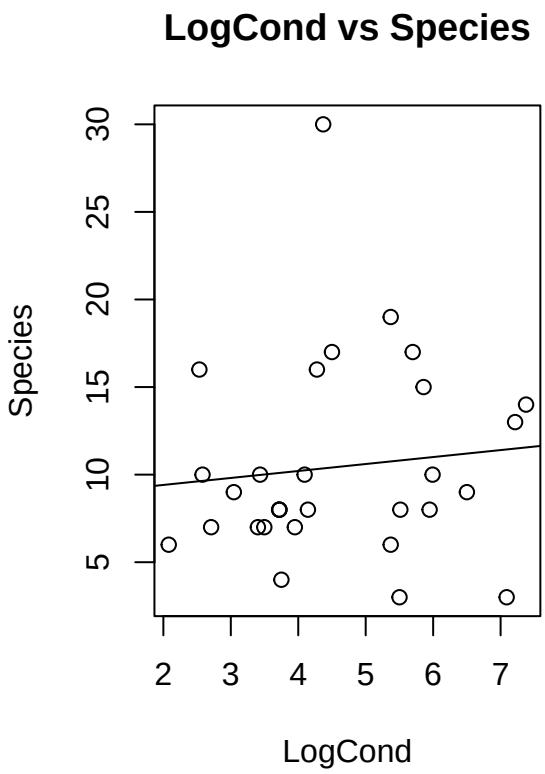
```
par(mfrow = c(1, 2))
plot(crust$MeanDepth,
     crust$Species,
     main = "MeanDepth vs Species",
     xlab = "MeanDepth",
     ylab = "Species")
not_log_lm <- lm(Species ~ MeanDepth, data = crust)
abline(not_log_lm)
plot(log(crust$MeanDepth),
     crust$Species,
     main = "LogMeanDepth vs Species",
     xlab = "LogMeanDepth",
     ylab = "Species")
log_lm <- lm(Species ~ log(MeanDepth), data = crust)
abline(log_lm)
```



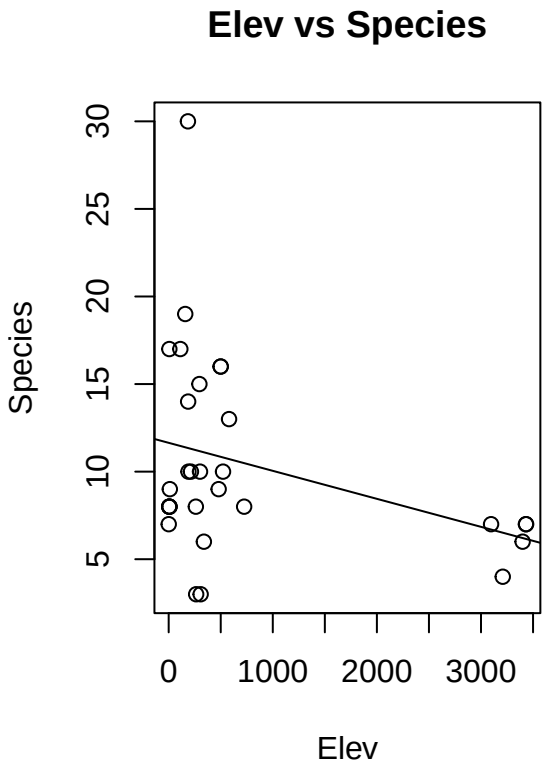


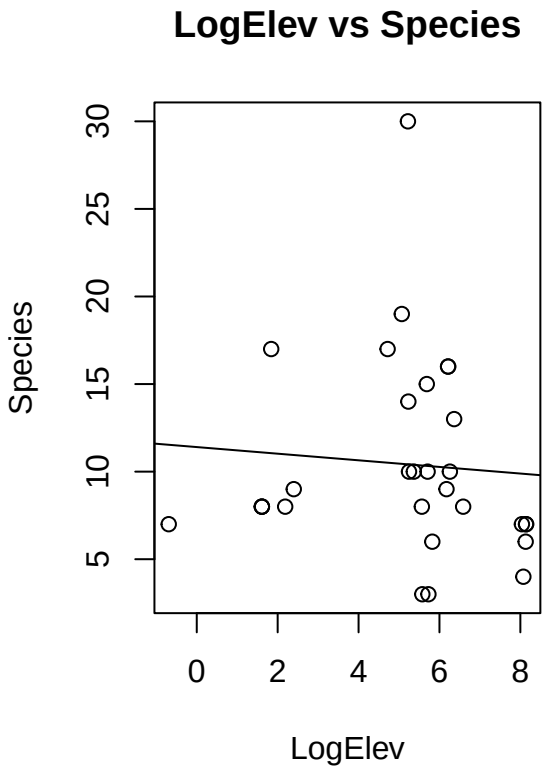

```
par(mfrow = c(1, 1))

par(mfrow = c(1, 2))
plot(crust$Cond,
     crust$Species,
     main = "Cond vs Species",
     xlab = "Cond",
     ylab = "Species")
not_log_lm <- lm(Species ~ Cond, data = crust)
abline(not_log_lm)
plot(log(crust$Cond),
     crust$Species,
     main = "LogCond vs Species",
     xlab = "LogCond",
     ylab = "Species")
log_lm <- lm(Species ~ log(Cond), data = crust)
abline(log_lm)
```



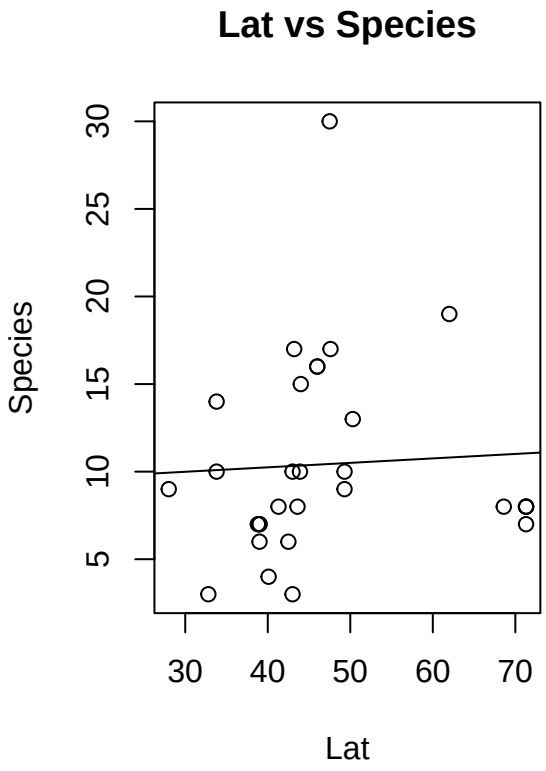


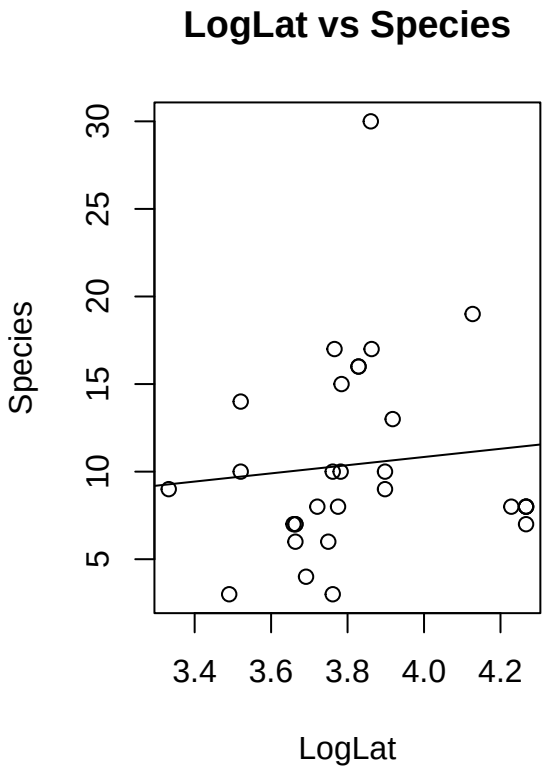

```
par(mfrow = c(1, 1))

par(mfrow = c(1, 2))
plot(crust$Elev,
     crust$Species,
     main = "Elev vs Species",
     xlab = "Elev",
     ylab = "Species")
not_log_lm <- lm(Species ~ Elev, data = crust)
abline(not_log_lm)
plot(log(crust$Elev),
     crust$Species,
     main = "LogElev vs Species",
     xlab = "LogElev",
     ylab = "Species")
log_lm <- lm(Species ~ log(Elev), data = crust)
abline(log_lm)
```



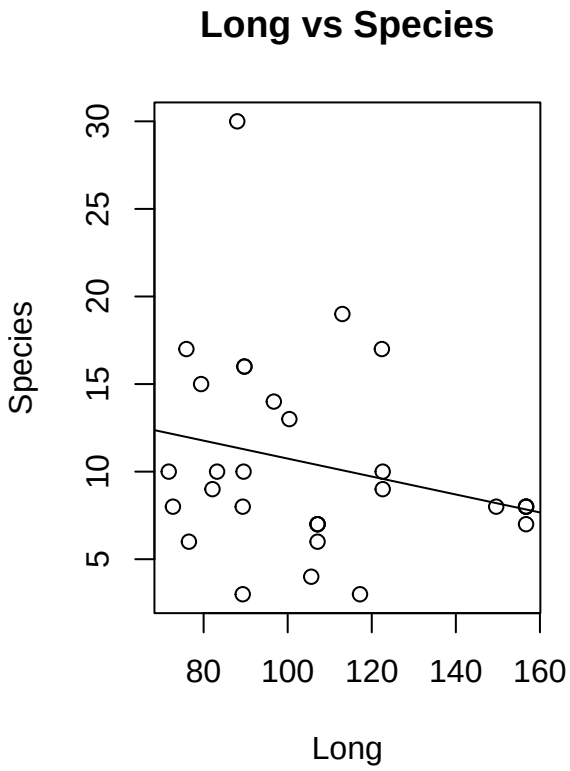


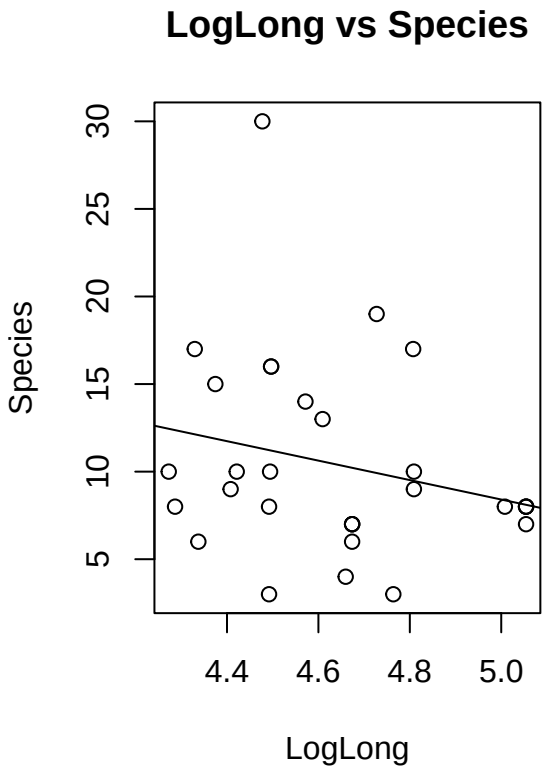

```
par(mfrow = c(1, 1))

par(mfrow = c(1, 2))
plot(crust$Lat,
     crust$Species,
     main = "Lat vs Species",
     xlab = "Lat",
     ylab = "Species")
not_log_lm <- lm(Species ~ Lat, data = crust)
abline(not_log_lm)
plot(log(crust$Lat),
     crust$Species,
     main = "LogLat vs Species",
     xlab = "LogLat",
     ylab = "Species")
log_lm <- lm(Species ~ log(Lat), data = crust)
abline(log_lm)
```



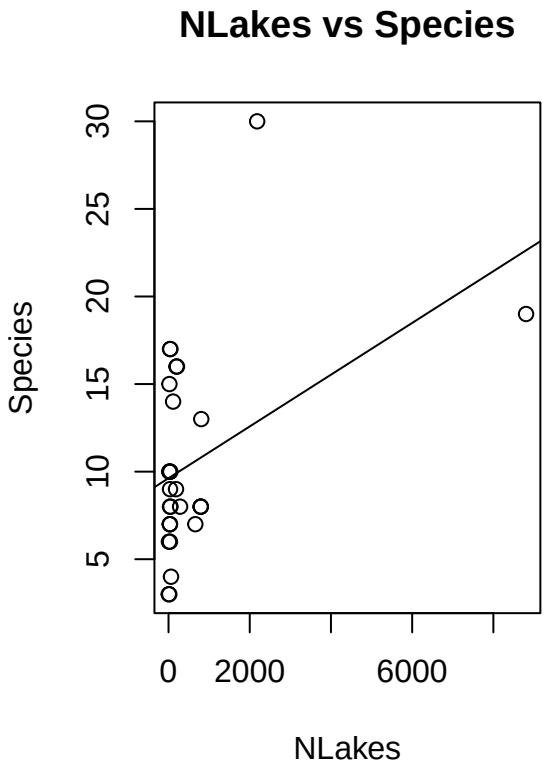


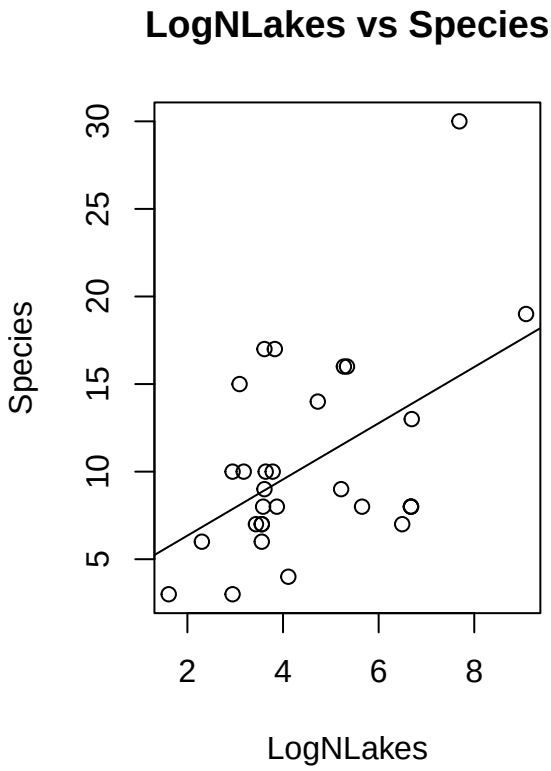

```
par(mfrow = c(1, 1))

par(mfrow = c(1, 2))
plot(crust$Long,
     crust$Species,
     main = "Long vs Species",
     xlab = "Long",
     ylab = "Species")
not_log_lm <- lm(Species ~ Long, data = crust)
abline(not_log_lm)
plot(log(crust$Long),
     crust$Species,
     main = "LogLong vs Species",
     xlab = "LogLong",
     ylab = "Species")
log_lm <- lm(Species ~ log(Long), data = crust)
abline(log_lm)
```



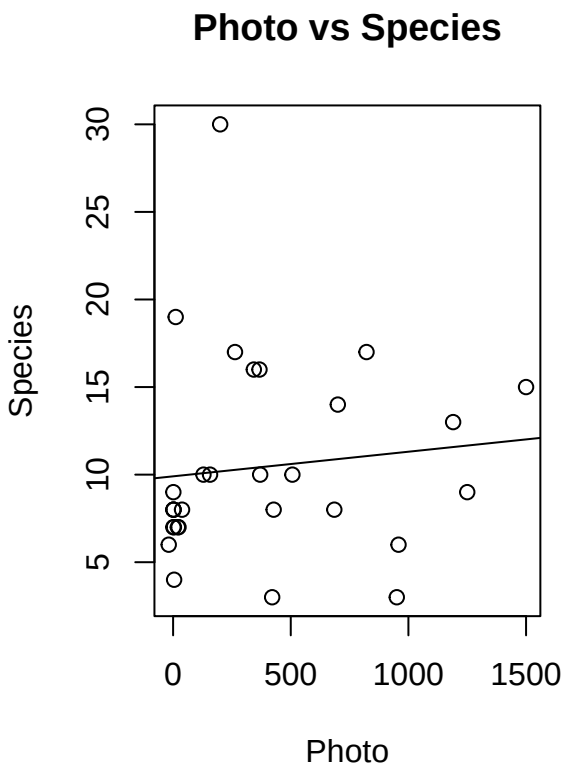


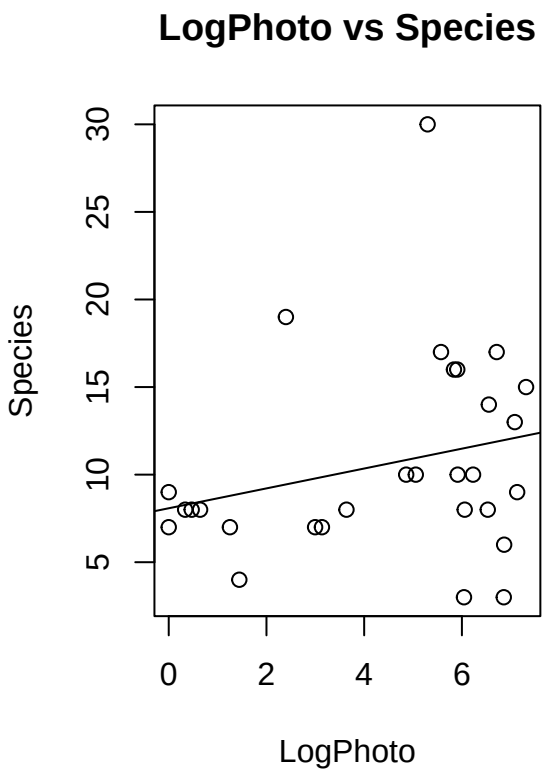

```
par(mfrow = c(1, 1))

par(mfrow = c(1, 2))
plot(crust$NLakes,
     crust$Species,
     main = "NLakes vs Species",
     xlab = "NLakes",
     ylab = "Species")
not_log_lm <- lm(Species ~ NLakes, data = crust)
abline(not_log_lm)
plot(log(crust$NLakes),
     crust$Species,
     main = "LogNLakes vs Species",
     xlab = "LogNLakes",
     ylab = "Species")
log_lm <- lm(Species ~ log(NLakes), data = crust)
abline(log_lm)
```



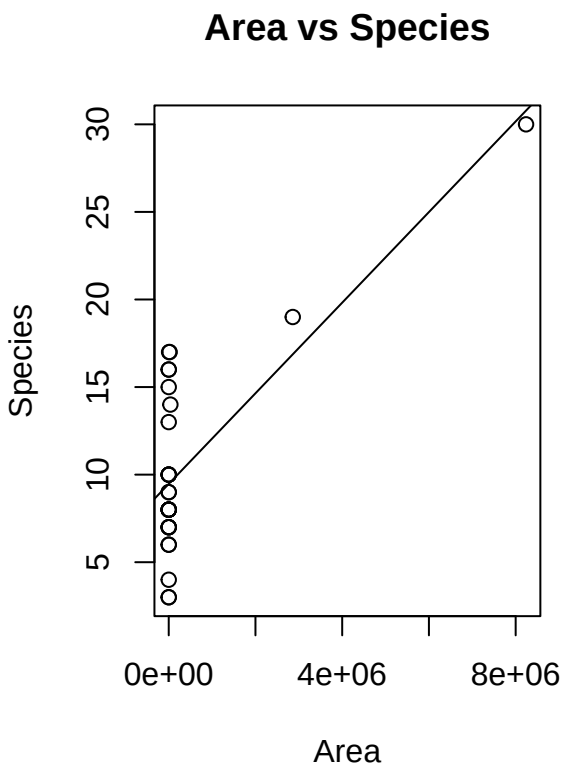


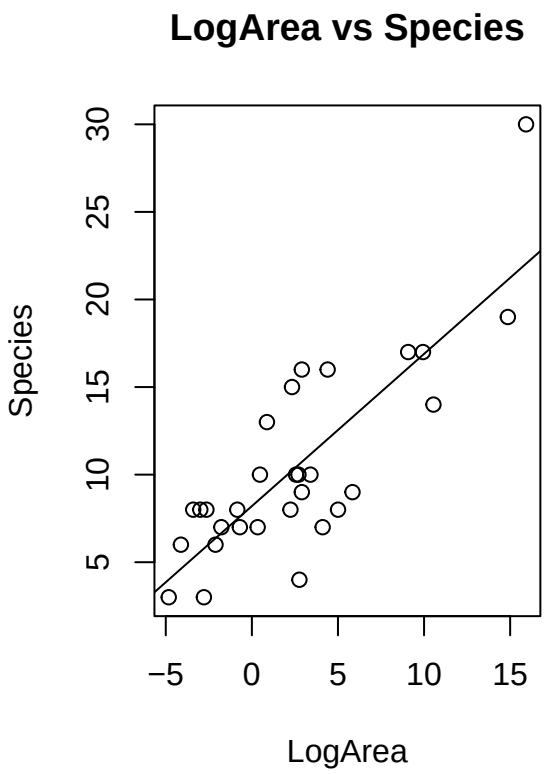

```r
par(mfrow = c(1, 1))

par(mfrow = c(1, 2))
plot(crust$Photo,
     crust$Species,
     main = "Photo vs Species",
     xlab = "Photo",
     ylab = "Species")
not_log_lm <- lm(Species ~ Photo, data = crust)
abline(not_log_lm)
plot(log(crust$Photo),
     crust$Species,
     main = "LogPhoto vs Species",
     xlab = "LogPhoto",
     ylab = "Species")
log_lm <- lm(Species ~ log(Photo), data = crust)
abline(log_lm)
```






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

par(mfrow = c(1, 2))
plot(crust$Area,
     crust$Species,
     main = "Area vs Species",
     xlab = "Area",
     ylab = "Species")
not_log_lm <- lm(Species ~ Area, data = crust)
abline(not_log_lm)
plot(log(crust$Area),
     crust$Species,
     main = "LogArea vs Species",
     xlab = "LogArea",
     ylab = "Species")
log_lm <- lm(Species ~ log(Area), data = crust)
abline(log_lm)
```






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

Section 4

```
lm_1 <- lm(Species ~ LogMeanDepth, data = crust_2)
summary(lm_1)
```

```
##
## Call:
## lm(formula = Species ~ LogMeanDepth, data = crust_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7217 -2.6390 -0.0962  2.8938  9.8108
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.9766     0.7746  10.298 5.01e-11 ***
## LogMeanDepth   2.4439     0.3848   6.352 7.14e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.676 on 28 degrees of freedom
```

```
## Multiple R-squared:  0.5903, Adjusted R-squared:  0.5757
## F-statistic: 40.34 on 1 and 28 DF,  p-value: 7.14e-07
```

```
lm_2 <- lm(Species ~ LogCond, data = crust_2)
summary(lm_2)
```

```
##
## Call:
## lm(formula = Species ~ LogCond, data = crust_2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.439 -2.980 -2.096  2.213 19.645
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.6158     3.4584   2.491   0.0189 *
## LogCond       0.3981     0.7223   0.551   0.5859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.712 on 28 degrees of freedom
## Multiple R-squared:  0.01073,    Adjusted R-squared:  -0.0246
## F-statistic: 0.3038 on 1 and 28 DF,  p-value: 0.5859
```

```
lm_3 <- lm(Species ~ LogElev, data = crust_2)
summary(lm_3)
```

```
##
## Call:
## lm(formula = Species ~ LogElev, data = crust_2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.349 -3.098 -2.053  3.390 19.584
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.4022     2.5950   4.394 0.000145 ***
## LogElev      -0.1890     0.4632  -0.408 0.686425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.726 on 28 degrees of freedom
## Multiple R-squared:  0.005908,   Adjusted R-squared:  -0.0296
## F-statistic: 0.1664 on 1 and 28 DF,  p-value: 0.6864
```

```
lm_4 <- lm(Species ~ Lat, data = crust_2)
summary(lm_4)
```

```
##
## Call:
```

```
## lm(formula = Species ~ Lat, data = crust_2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.323 -3.220 -1.881  3.557 19.563
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.22832    4.21539   2.189   0.0371 *
## Lat          0.02545    0.08623   0.295   0.7701
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.734 on 28 degrees of freedom
## Multiple R-squared:  0.003101,   Adjusted R-squared:  -0.0325
## F-statistic: 0.08709 on 1 and 28 DF,  p-value: 0.7701
```

```
lm_5 <- lm(Species ~ Long, data = crust_2)
summary(lm_5)
```

```
##
## Call:
## lm(formula = Species ~ Long, data = crust_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2929 -3.3809 -0.7131  2.8836 18.6405
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.86849    4.21661   3.763  0.00079 ***
## Long        -0.05124    0.03858  -1.328  0.19484
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.57 on 28 degrees of freedom
## Multiple R-squared:  0.05927,    Adjusted R-squared:  0.02567
## F-statistic: 1.764 on 1 and 28 DF,  p-value: 0.1948
```

```
lm_6 <- lm(Species ~ LogNLakes, data = crust_2)
summary(lm_6)
```

```
##
## Call:
## lm(formula = Species ~ LogNLakes, data = crust_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5550 -2.8125 -0.8819  2.0434 14.5258
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.1388     2.5923   1.211  0.23608
```

```
## LogNLakes     1.6044     0.5337   3.006  0.00554 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.993 on 28 degrees of freedom
## Multiple R-squared:  0.244,  Adjusted R-squared:  0.217
## F-statistic: 9.036 on 1 and 28 DF,  p-value: 0.005535
```

```
lm_7 <- lm(Species ~ Photo, data = crust_2)
summary(lm_7)
```

```
##
## Call:
## lm(formula = Species ~ Photo, data = crust_2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.239 -2.906 -1.905  2.598 19.816
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.902457   1.385889   7.145 8.93e-08 ***
## Photo       0.001406   0.002419   0.581    0.566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.708 on 28 degrees of freedom
## Multiple R-squared:  0.01192,    Adjusted R-squared:  -0.02337
## F-statistic: 0.3377 on 1 and 28 DF,  p-value: 0.5658
```

```
lm_8 <- lm(Species ~ LogArea, data = crust_2)
summary(lm_8)
```

```
##
## Call:
## lm(formula = Species ~ LogArea, data = crust_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6014 -2.0235 -0.3995  1.9157  7.9571
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.2026     0.6786  12.088 1.25e-12 ***
## LogArea       0.8691     0.1170   7.425 4.36e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.333 on 28 degrees of freedom
## Multiple R-squared:  0.6632, Adjusted R-squared:  0.6512
## F-statistic: 55.13 on 1 and 28 DF,  p-value: 4.364e-08
```

Section 5

```
data = crust
data <- data[, !(names(data) %in% c("Unnamed..0"))]
data$LogMeanDepth <- log(data$MeanDepth + 1)
data$LogCond <- log(data$Cond + 1)
data$LogElev <- log(data$Elev + 1)
data$LogNLakes <- log(data$NLakes + 1)
data$LogArea <- log(data$Area + 1)
y <- data$Species
x <- c("LogMeanDepth", "LogCond", "LogElev", "Lat", "Long", "LogNLakes", "Photo", "LogArea")
formula <- as.formula(paste("Species ~", paste(x, collapse = " + ")))
best <- regsubsets(formula, data = data, nvmax = 8)
print(summary(best))
```

```
## Subset selection object
## Call: regsubsets.formula(formula, data = data, nvmax = 8)
## 8 Variables  (and intercept)
##              Forced in Forced out
## LogMeanDepth     FALSE      FALSE
## LogCond          FALSE      FALSE
## LogElev          FALSE      FALSE
## Lat              FALSE      FALSE
## Long             FALSE      FALSE
## LogNLakes        FALSE      FALSE
## Photo            FALSE      FALSE
## LogArea          FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          LogMeanDepth LogCond LogElev Lat Long LogNLakes Photo LogArea
## 1  ( 1 ) " "          " "     " "     " " " "  " "       " "   "*"
## 2  ( 1 ) "*"          " "     " "     " " " "  " "       " "   "*"
## 3  ( 1 ) "*"          " "     " "     " " "*"  "*"       " "   " "
## 4  ( 1 ) "*"          " "     " "     " " "*"  "*"       " "   "*"
## 5  ( 1 ) " "          "*"     " "     " " "*"  "*"       "*"   "*"
## 6  ( 1 ) "*"          "*"     " "     " " "*"  "*"       "*"   "*"
## 7  ( 1 ) "*"          "*"     " "     "*" "*"  "*"       "*"   "*"
## 8  ( 1 ) "*"          "*"     "*"     "*" "*"  "*"       "*"   "*"
```

```
one_pred_lm <- lm(Species ~ LogArea, data = crust_2)
summary(one_pred_lm)
```

```
##
## Call:
## lm(formula = Species ~ LogArea, data = crust_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6014 -2.0235 -0.3995  1.9157  7.9571
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   8.2026     0.6786  12.088 1.25e-12 ***
## LogArea       0.8691     0.1170   7.425 4.36e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.333 on 28 degrees of freedom
## Multiple R-squared:  0.6632, Adjusted R-squared:  0.6512
## F-statistic: 55.13 on 1 and 28 DF,  p-value: 4.364e-08
```

```
main_lm <- lm(Species ~ LogMeanDepth + LogArea, data = crust_2)
summary(main_lm)
```

```
##
## Call:
## lm(formula = Species ~ LogMeanDepth + LogArea, data = crust_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2443 -1.8753 -0.5591  2.5246  7.8315
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.9780     0.6952  11.476 6.85e-12 ***
## LogMeanDepth   0.8417     0.6710   1.254  0.22044
## LogArea        0.6270     0.2251   2.785  0.00966 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.299 on 27 degrees of freedom
## Multiple R-squared:  0.6817, Adjusted R-squared:  0.6582
## F-statistic: 28.92 on 2 and 27 DF,  p-value: 1.939e-07
```

```
anova(one_pred_lm, main_lm)
```

```
## Analysis of Variance Table
##
## Model 1: Species ~ LogArea
## Model 2: Species ~ LogMeanDepth + LogArea
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 311.00
## 2     27 293.87  1    17.127 1.5736 0.2204
```

Section 6

```
main_lm <- lm(Species ~ LogMeanDepth + LogArea, data = crust_2)
summary(main_lm)
```

```
##
## Call:
## lm(formula = Species ~ LogMeanDepth + LogArea, data = crust_2)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2443 -1.8753 -0.5591  2.5246  7.8315
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.9780     0.6952  11.476 6.85e-12 ***
## LogMeanDepth   0.8417     0.6710   1.254  0.22044
## LogArea        0.6270     0.2251   2.785  0.00966 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.299 on 27 degrees of freedom
## Multiple R-squared:  0.6817, Adjusted R-squared:  0.6582
## F-statistic: 28.92 on 2 and 27 DF,  p-value: 1.939e-07
```

```
inter_lm <- lm(Species ~ LogMeanDepth * LogArea, data = crust_2)
summary(inter_lm)
```

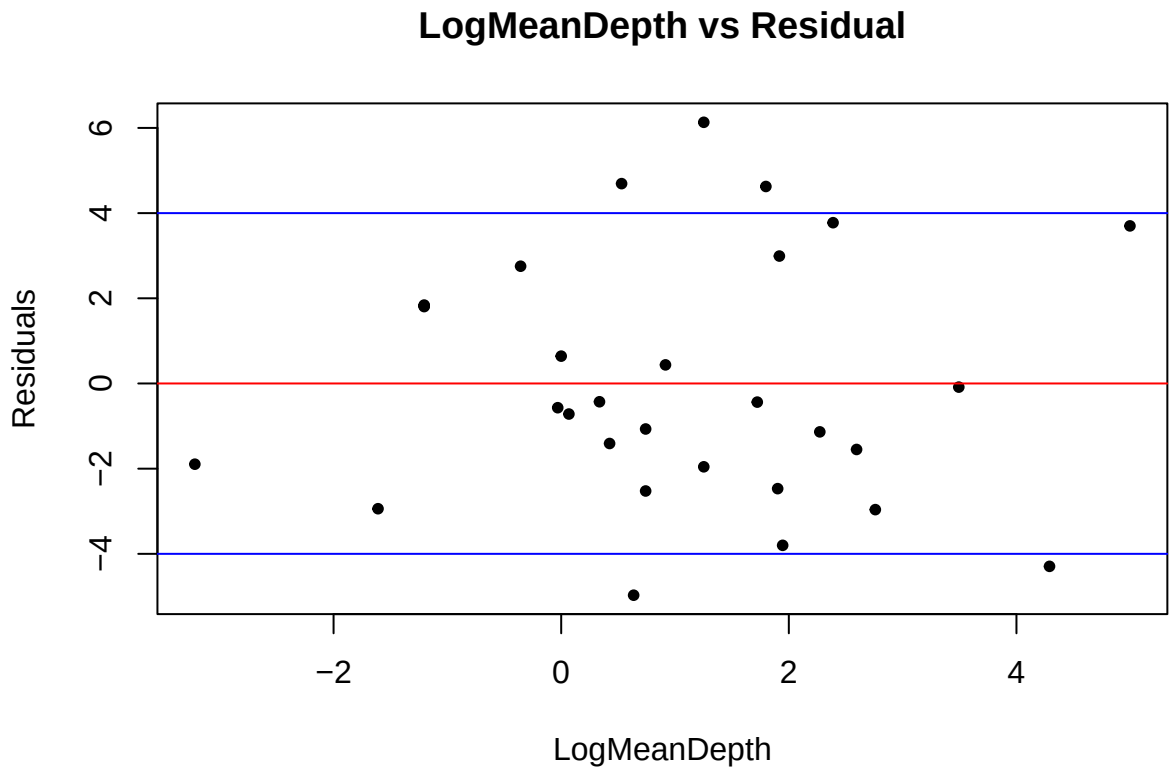
```
##
## Call:
## lm(formula = Species ~ LogMeanDepth * LogArea, data = crust_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9710 -1.9410 -0.5033  1.8362  6.1320
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7.53204    0.66447  11.335 1.47e-11 ***
## LogMeanDepth          1.00787    0.62037   1.625   0.1163
## LogArea               0.20519    0.26947   0.761   0.4532
## LogMeanDepth:LogArea  0.13149    0.05382   2.443   0.0217 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.032 on 26 degrees of freedom
## Multiple R-squared:  0.7412, Adjusted R-squared:  0.7113
## F-statistic: 24.82 on 3 and 26 DF,  p-value: 8.536e-08
```

```
anova(main_lm, inter_lm)
```

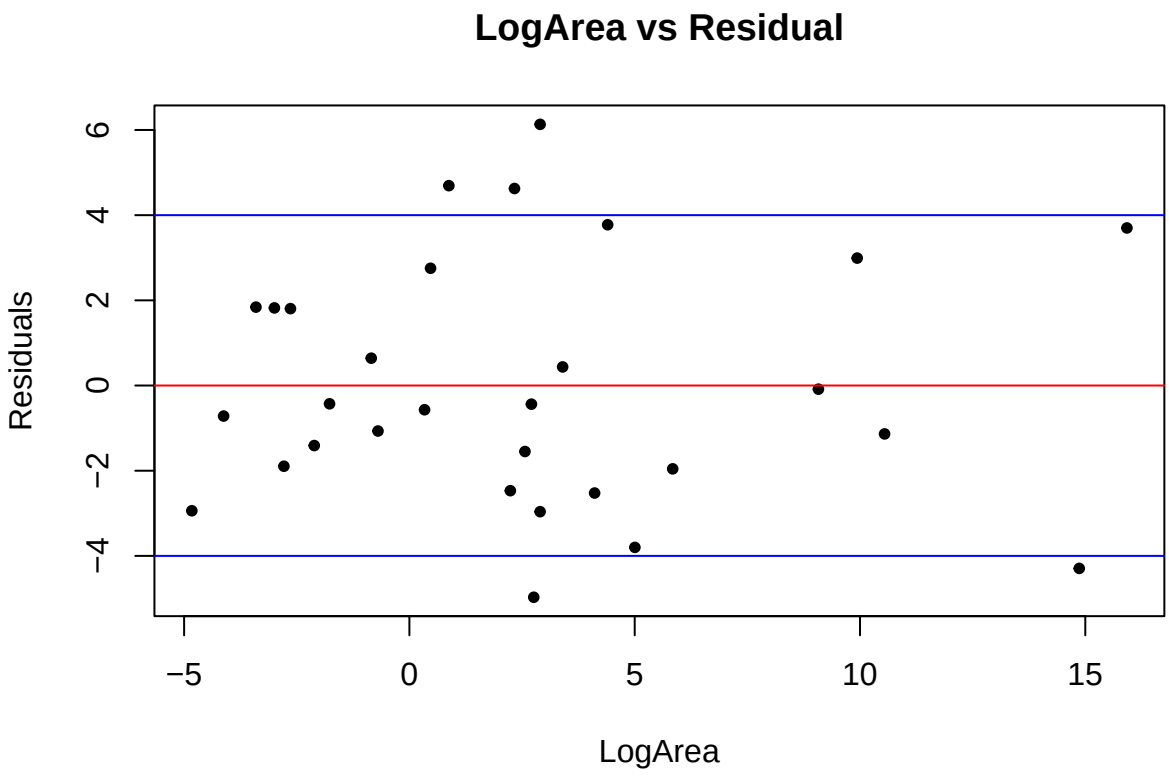
```
## Analysis of Variance Table
##
## Model 1: Species ~ LogMeanDepth + LogArea
## Model 2: Species ~ LogMeanDepth * LogArea
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     27 293.87
## 2     26 239.00  1    54.867 5.9688 0.02166 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Section 7

```
plot(crust_2$LogMeanDepth, resid(crust_2_lm),
     xlab = "LogMeanDepth",
     ylab = "Residuals",
     main = "LogMeanDepth vs Residual",
     pch = 20)
abline(h = 0, col = "red")
abline(h = 4, col = "blue")
abline(h = -4, col = "blue")
```

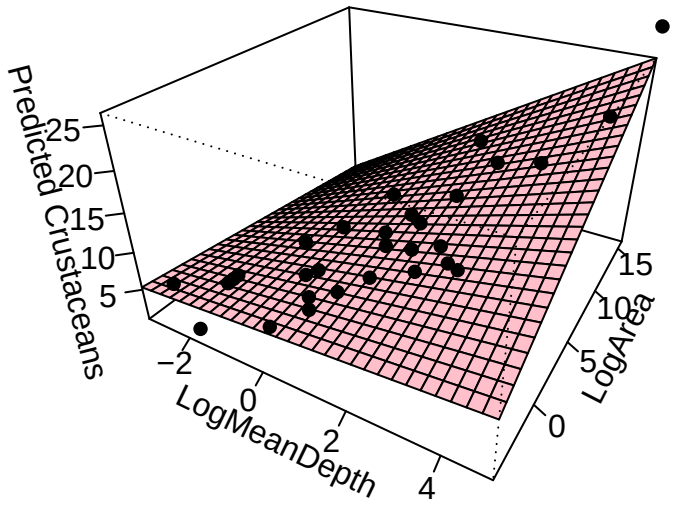



```
plot(crust_2$LogArea, resid(crust_2_lm),
     xlab = "LogArea",
     ylab = "Residuals",
     main = "LogArea vs Residual",
     pch = 20)
abline(h = 0, col = "red")
abline(h = 4, col = "blue")
abline(h = -4, col = "blue")
```




```
beta_0 <- 7.53204
beta_LMD <- 1.00787
beta_LA <- 0.20519
beta_Inter <- 0.13149
x <- seq(min(crust_2$LogMeanDepth), max(crust_2$LogMeanDepth), length.out = 30)
y <- seq(min(crust_2$LogArea), max(crust_2$LogArea), length.out = 30)
z <- outer(x, y, function(x, y) {
  beta_0 + beta_LMD * x + beta_LA * y + beta_Inter * (x * y)
})
model_3d <- persp(
  x, y, z,
  xlab = "LogMeanDepth", ylab = "LogArea", zlab = "Predicted Crustaceans",
  main = "Final Model",
  col = "pink", theta = 30, phi = 30, expand = 0.6, ticktype = "detailed"
)
points(trans3d(
  crust_2$LogMeanDepth, crust_2$LogArea, crust_2$Species,
  model_3d),
  col = "black", pch = 16)
```

Final Model


Predicted Crustaceans
25
20
15
10
5
−2
0
2
4
LogMeanDepth
15
10
5
0
LogArea