Data Analysis on Crustaceans in Lakes

Haechan Jung

2024-12-04

Introduciton

Biology researchers analyze investigation data to demonstrate their theories as researching living things. They try to collect data as much as possible since more data will enhance the quality of the research result. Still, there are several realistic constraints such as unaffordable expenses and insufficient technology. This data analysis report proposes a prediction model to overcome the limitations; we can predict the number of crustacean species in a new lake with several predictor variables. The model is designed based on a data set that has 30 random lakes and variables:

Variable	Description
Speices	Number of Crustacean Species
MeanDepth	Average Lake depth in Meters
Cond	Specific Conductance (A Measure of Mineral Content), in Micro Siemans
Elev	Lake Elevation, in Meters
Lat	Latitude, Degrees North
Long	Longitude, Degrees West
NLakes	Number of Lakes Within 20km
Photo	Rate of Photosynthesis (Measured by Using C14)
Area	Surface Area of Lake, in Hectares

The data set has no missing value, and all of the covariates are quantitative values. Species was used as a a response, and the others except for Photo were used as predictors in the model. Logarithmic transformation is applied to Elev and Area variables since they have a relatively wider range than other variables.

Methods

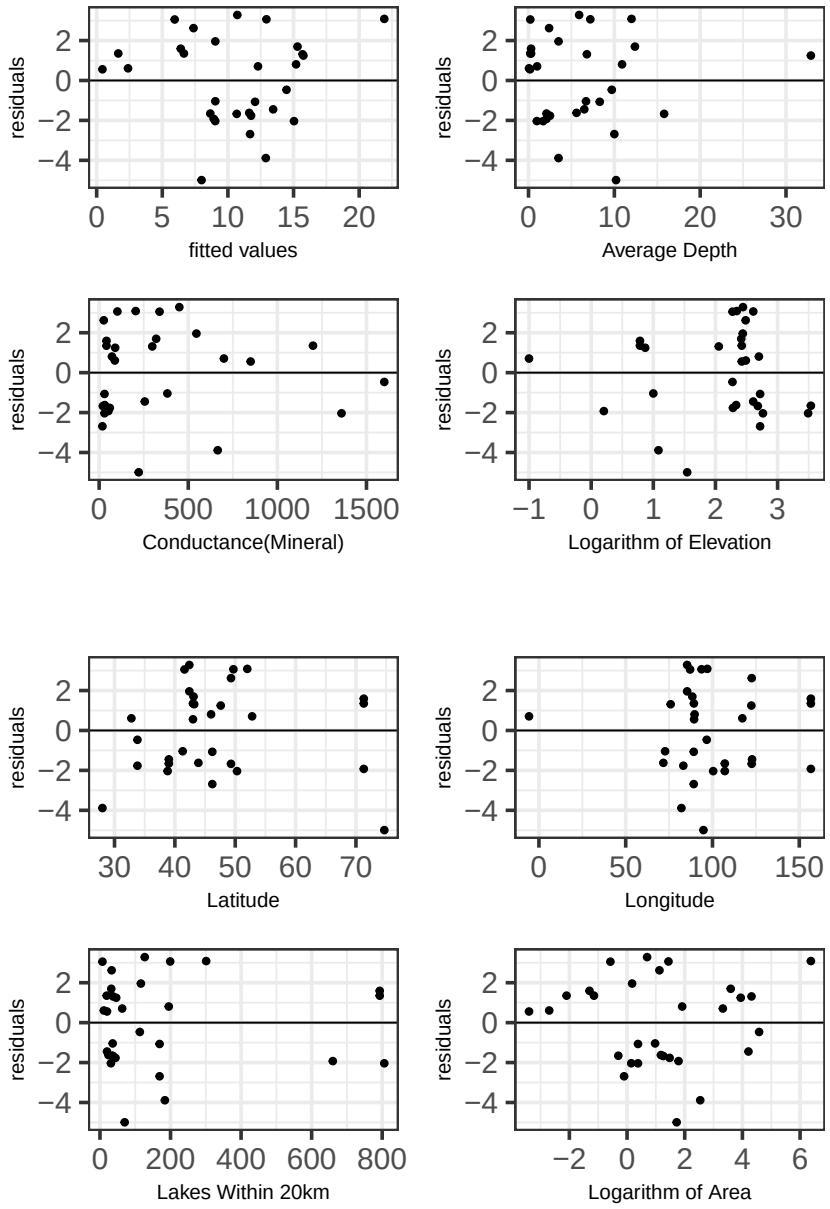
The prediction model was developed through a comparison of the best models from three different model selection techniques: forward selection, backward selection, and stepwise selection. All of these three are to find the model that has the relatively smallest value of AIC. The two most frequent models were shortlisted from the results, and the final model was settled after a comparison of AIC values.

The final linear regression model is:

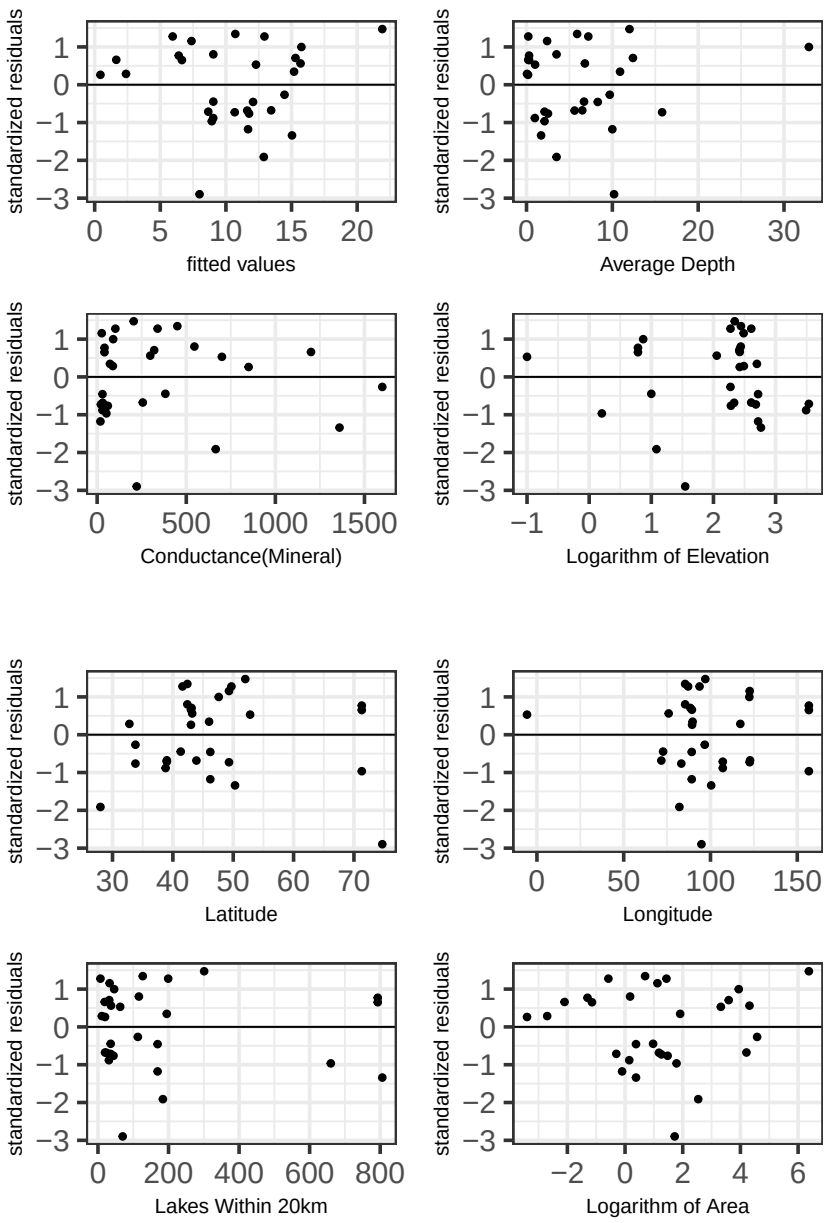
$$species_i = \beta_0 + \beta_1\ meanDepth_i + \beta_2\ cond_i + \beta_3\ log\ elev_i + \beta_4\ lat_i$$
$$+\ \beta_5\ long_i + \beta_6\ nLakes_i + \beta_7\ log\ area_i + e_i, \quad e_i \overset{iid}{\sim} (0, \sigma^2)$$

Diagnosis

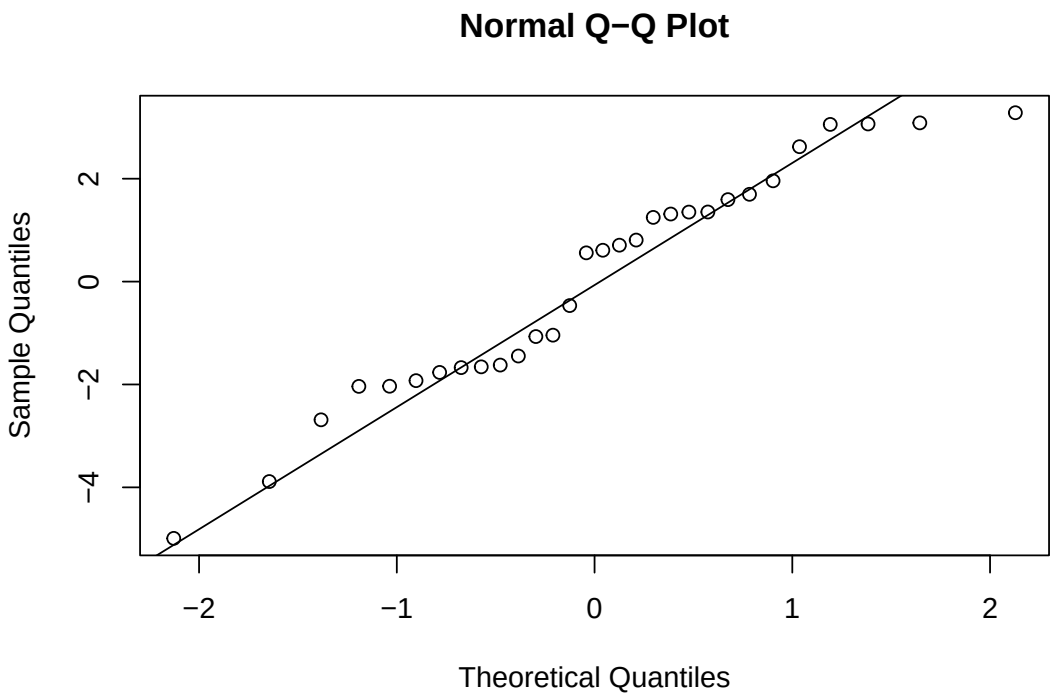
The reliability of the final model was diagnosed with the residual plot, standardized residual plot, and qqplot.



Dispersion of points in residual plots can check if a violation of the model assumption exists or not. An apparent pattern of dispersion that does not show const variability indicates the lack of linearity of a model. Such an obvious pattern or trend does not show in the above residual plots.



Constant variance is another assumption to be checked for the final model. If a model has constant variance, dispersion in standardized residual plots will show constant variability of fitted values and predictors. An obvious pattern or trend that indicates inconsistency does not show in the above plots.



The normality assumption can be assessed by qqplot. If the dataset is normally distributed, the plotted points will be located around the 45-degree straight line. Except for an outlier, the cluster of points shows the normality of the model.

Analysis

From the final linear regression model, we can determine the mean and variance function of the number of species:

$$E(species_i \mid meanDepth,\ cond,\ log\ elev,\ lat,\ long,\ nLakes,\ log\ area) =$$
$$13.75 + 0.203\ meanDepth_i - 0.002\ cond_i + 1.497\ log\ elev_i - 0.093\ lat_i$$
$$- 0.066\ long_i + 0.013\ nLakes_i + 1.562\ log\ area_i + e_i, \qquad e_i \overset{iid}{\sim} (0, \sigma^2)$$
$$Var(species_i \mid meanDepth,\ cond,\ log\ elev,\ lat,\ long,\ nLakes,\ log\ area) = 6.318$$

Also, we can conclude:

- An additional meter in the average lake depth is associated with a 0.203 increase in the average number of crustacean species.
- Additional micro Siemens in the conductance is associated with a 0.002 decrease in the average number of crustacean species.
- A 10-meter increase in the elevation is associated with a 1.497 increase in the average number of crustacean species.
- A degree increase in the north latitude is associated with a 0.093 decrease in the average number of crustacean species.

- An additional degree in the west longitude is associated with a 0.066 decrease in the average number of crustacean species.
- An additional lake within 20km is associated with a 0.013 increase in the average number of crustacean species.
- A 10-hectare increase in the surface area is associated with a 1.562 increase in the average number of crustacean species.
- The data analysis cannot find a meaningful linear relationship between the rate of photosynthesis and the number of crustacean species.

Prediction

The given new lake has characteristics: MeanDepth = 153, Cond = 167, Elev = 372, Lat = 46, Long = -3, NLakes = 44, Area = 58000. The predicted number of species in the new lake is approximately 52.23, and the actual number will fall between 23.26 and 81.20 with 95% confidence.

Appendix

```
library(alr4)
library(tidyverse)
library(readr)
library(GGally)
library(dplyr)
library(leaps)
library(MASS)
library(broom)
library(patchwork)
```

```
#Check if all covariates are quantitative
df = read_csv("C:/STAT_3301_Data_Storage/projectdata24.csv")
head(df, 5)
```

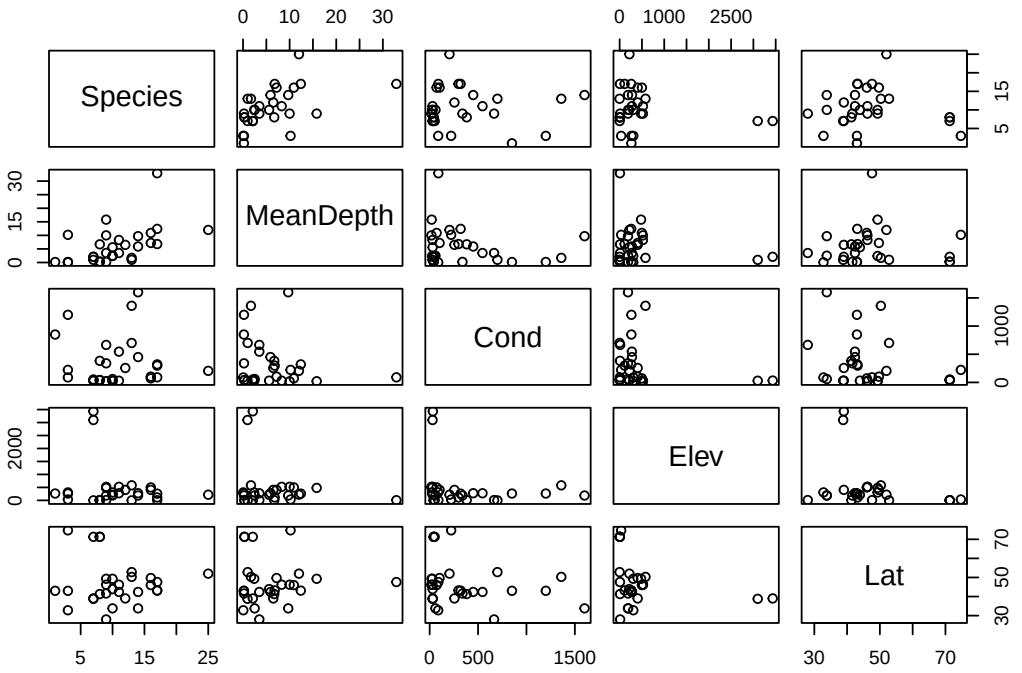
```
## # A tibble: 5 x 10
##   ...1         Species MeanDepth   Cond  Elev   Lat  Long NLakes Photo      Area
##   <chr>          <dbl>     <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>     <dbl>
## 1 Tower              3       0.2  1200     264  43    89.3     19 951.     0.008
## 2 Marion            10       2.4    26     305  49.3 123.      33  22     13.3
## 3 Miramar_1          3       0.06   88     307  32.8 117.      11 335.     0.002
## 4 Mendota           17      12.4   320     259  43.1  88.4     32 938   3940
## 5 NARL_IBP_A+B       8       0.3    41.2     5  71.3 157.     793   1.6    0.0714
```

```
# Check if a missing value exists
sum(is.na(df))
```

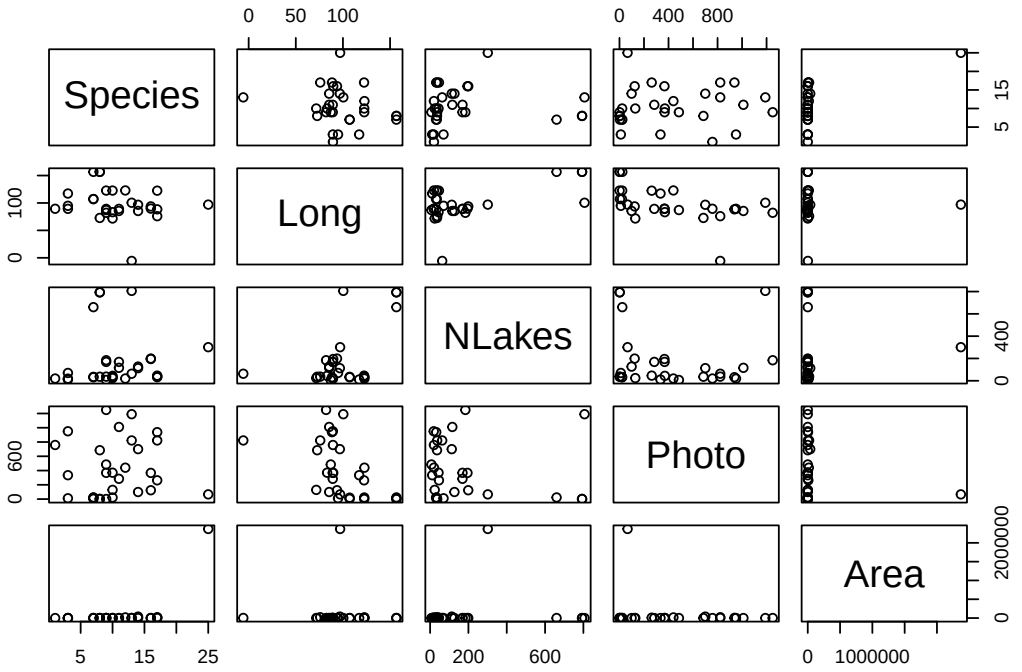
```
## [1] 0
```

```
# Exclude name column which is unnecessary for further data analysis.
df = subset(df, select = c(Species, MeanDepth, Cond, Elev, Lat, Long, NLakes, Photo, Area))
```

```
# Check relationships between the response and predictors
pairs(subset(df, select = c(Species, MeanDepth, Cond, Elev, Lat)))
```




```
pairs(subset(df, select = c(Species, Long, NLakes, Photo, Area)))
```



```r
# The gaps between the minimum and maximum values are big enough to apply transformation
range(df$Elev)
```

```
## [1]   -1 3433
```

```r
range(df$Area)
```

```
## [1] 4.00e-04 2.37e+06
```

```r
# Since Elev has -1 as the minimum value,
# it is necessary to add constant(1.1) before taking logarithm
df = df %>% mutate(logElev = log10(Elev + 1.1), logArea = log10(Area))
lake = subset(df, select = c(Species, MeanDepth, Cond, logElev, Lat, Long, NLakes, Photo, logArea))
```

```r
# Comparison of best models from the selection process
print(names(model_fwd_AIC$model))
```

```
## [1] "Species"   "logArea"   "logElev"   "NLakes"    "Long"      "MeanDepth"
## [7] "Cond"      "Lat"
```

```r
print(names(model_fwd_BIC$model))
```

```
## [1] "Species" "logArea"
```

```r
print(names(model_bwd_AIC$model))
```

```
## [1] "Species"   "MeanDepth" "Cond"      "logElev"   "Lat"       "Long"
## [7] "NLakes"    "logArea"
```

```r
print(names(model_bwd_BIC$model))
```

```
## [1] "Species"   "MeanDepth" "logElev"   "Long"      "NLakes"    "logArea"
```

```r
print(names(model_bth_AIC$model))
```

```
## [1] "Species"   "MeanDepth" "Cond"      "logElev"   "Lat"       "Long"
## [7] "NLakes"    "logArea"
```

```r
print(names(model_bth_BIC$model))
```

```
## [1] "Species"   "MeanDepth" "logElev"   "Long"      "NLakes"    "logArea"
```

```r
# Comparing AIC values of the two most frequent models
extractAIC(model_bth_AIC)
```

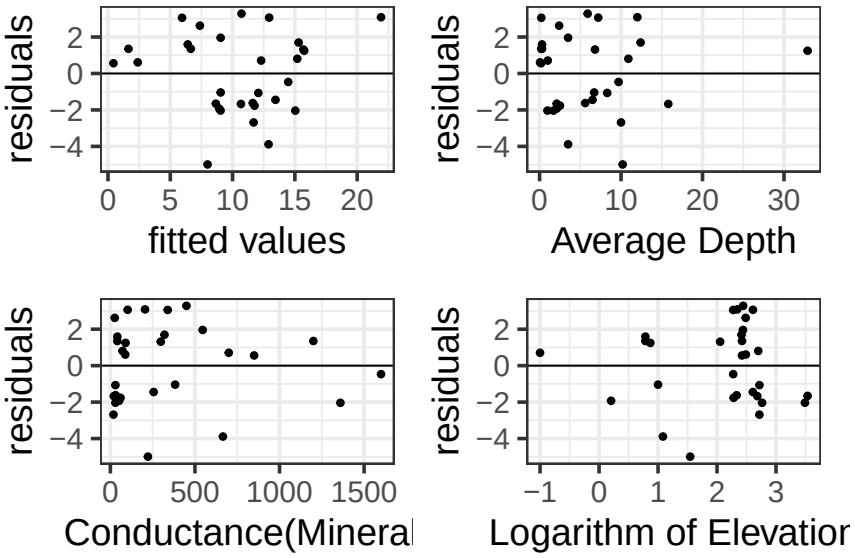
```
## [1]  8.00000 61.99705
```

```r
extractAIC(model_bth_BIC)
```

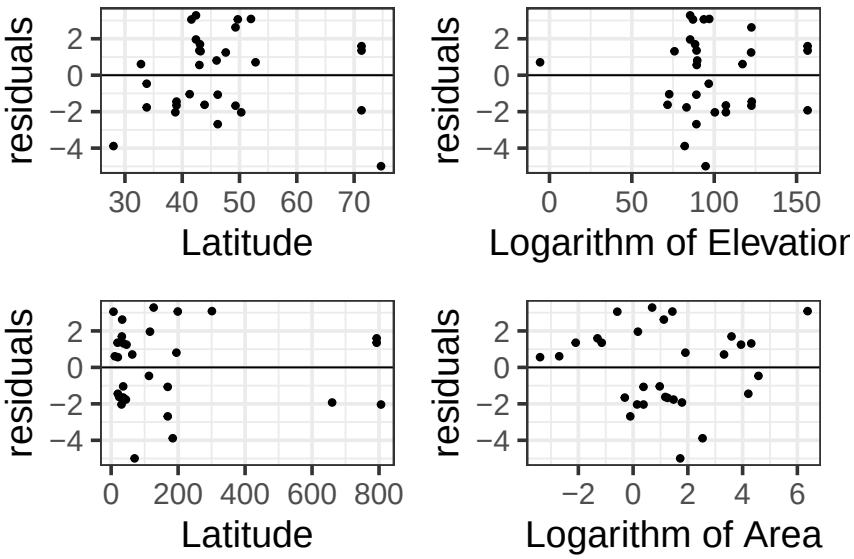
```
## [1]  6.00000 63.18806
```

```r
# Final linear regression model
lake.lm = lm(Species ~ MeanDepth + Cond + logElev + Lat + Long + NLakes + logArea, data = lake)
```

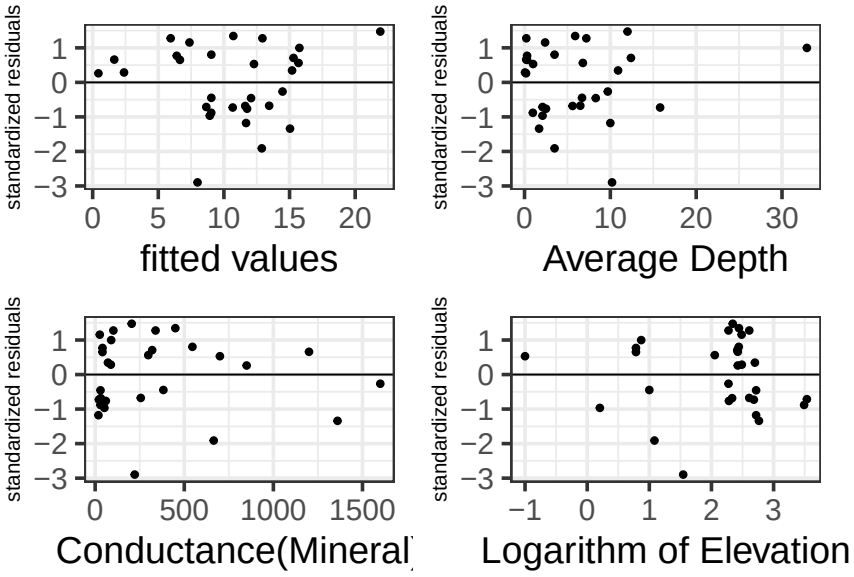
```r
# Standardized residual plots
base = augment(lake.lm) %>% ggplot(aes(y = .resid)) + geom_hline(yintercept = 0) +
  theme_bw(20) + ylab("residuals") + geom_point()
fit_plt = base + aes(x = .fitted) + xlab("fitted values")
meanDepth_plt = base + aes(x = MeanDepth) + xlab("Average Depth")
cond_plt = base + aes(x = Cond) + xlab("Conductance(Mineral)")
logElev_plt = base + aes(x = logElev) + xlab("Logarithm of Elevation")
(fit_plt + meanDepth_plt) / (cond_plt + logElev_plt)
```



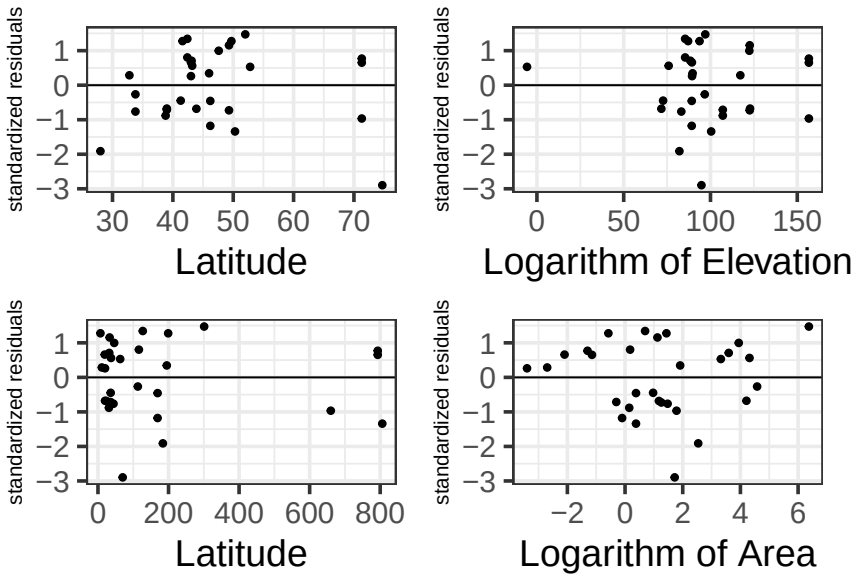

```r
lat_plt = base + aes(x = Lat) + xlab("Latitude")
long_plt = base + aes(x = Long) + xlab("Logarithm of Elevation")
nLakes_plt = base + aes(x = NLakes) + xlab("Latitude")
logArea_plt = base + aes(x = logArea) + xlab("Logarithm of Area")
(lat_plt + long_plt) / (nLakes_plt + logArea_plt)
```



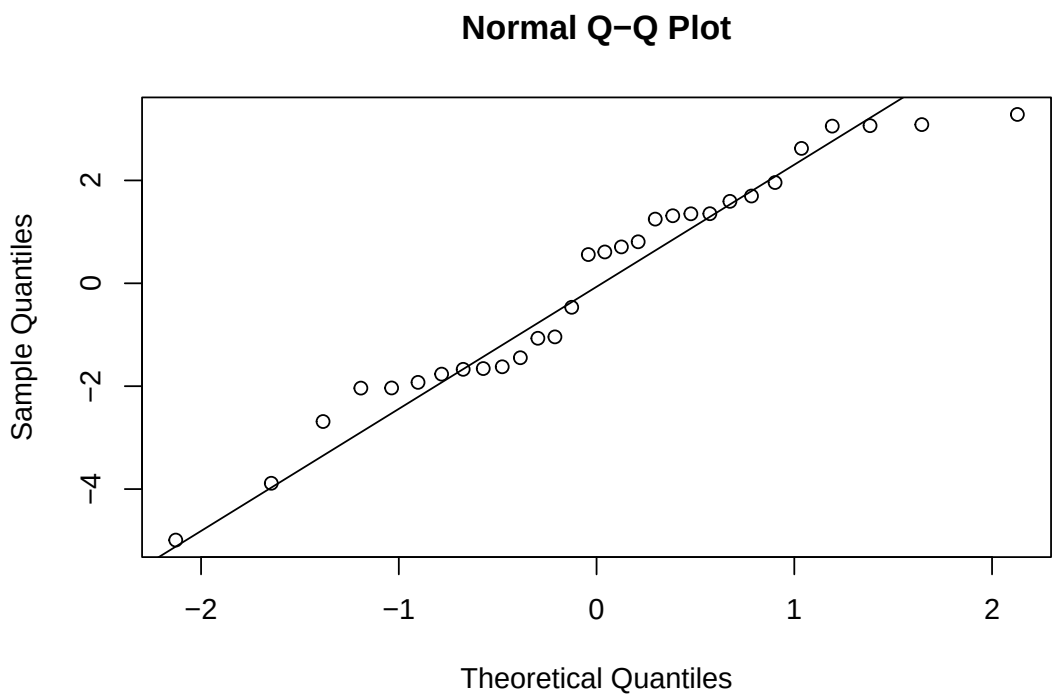

```
base = base + aes(y = .std.resid) + ylab("standardized residuals") + theme(axis.title.y = element_text(s
fit_plt = base + aes(x = .fitted) + xlab("fitted values")
meanDepth_plt = base + aes(x = MeanDepth) + xlab("Average Depth")
cond_plt = base + aes(x = Cond) + xlab("Conductance(Mineral)")
logElev_plt = base + aes(x = logElev) + xlab("Logarithm of Elevation")
(fit_plt + meanDepth_plt) / (cond_plt + logElev_plt)
```




```
lat_plt = base + aes(x = Lat) + xlab("Latitude")
long_plt = base + aes(x = Long) + xlab("Logarithm of Elevation")
nLakes_plt = base + aes(x = NLakes) + xlab("Latitude")
logArea_plt = base + aes(x = logArea) + xlab("Logarithm of Area")
(lat_plt + long_plt) / (nLakes_plt + logArea_plt)
```




```
# Check if it is normally distributed
qqnorm(lake.lm$residuals)
qqline(lake.lm$residuals)
```




```
# mean function
lake.lm$coefficients
```

```
##  (Intercept)     MeanDepth          Cond       logElev           Lat          Long
```

```
## 13.750690339  0.203162152 -0.002354263  1.497319252 -0.093269316 -0.066225840
##       NLakes      logArea
##  0.013338024  1.562474726
```

```r
# variance function
summary(lake.lm)$sigma^2
```

```
## [1] 6.317919
```

```r
# prediction of number of species in the new lake
predict(lake.lm, newdata = data.frame(MeanDepth = 153,
                                      Cond = 167,
                                      logElev = log10(372 + 1.1),
                                      Lat = 46,
                                      Long = -3,
                                      NLakes = 44,
                                      logArea = log10(58000)),
                                      interval = "prediction",
                                      level = 0.95)
```

```
##        fit      lwr      upr
## 1 52.23008 23.25979 81.20037
```