Final_Project_jung.801

Haechan Jung

2024-12-04

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

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

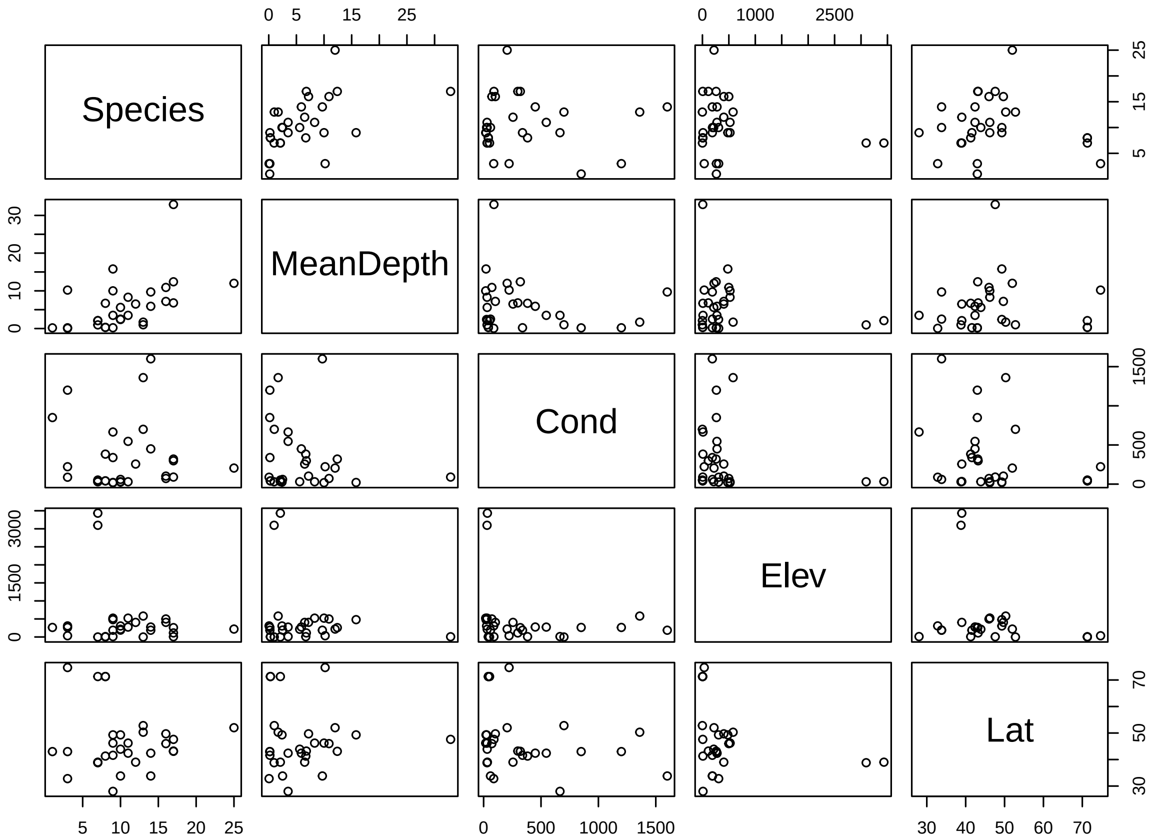
```
## # A tibble: 5 x 10
##   ...1         Species MeanDepth   Cond  Elev   Lat  Long NLakes Photo      Area
##   <chr>          <dbl>     <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>     <dbl>
## 1 Tower              3      0.2  1200     264  43    89.3     19 951.      0.008
## 2 Marion            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
```

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

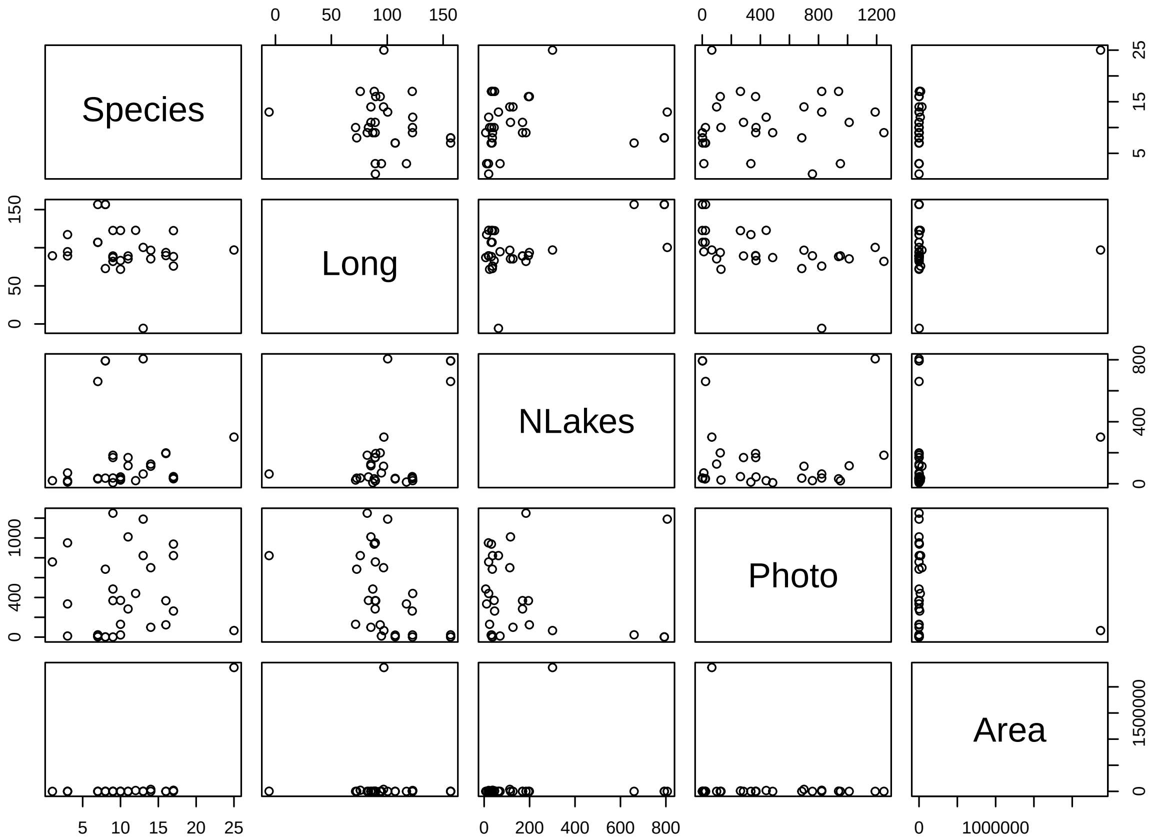
```
## [1] 0
```

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

```r
pairs(subset(df, select = c(Species, MeanDepth, Cond, Elev, Lat)))
```




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




```
range(df$Elev)
```

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

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

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

The gaps between the minimum and maximum values are big enough to apply transformation.Since Elev has -1 as the minimum value, it is necessary to add constant(1.1) before taking logarithm

```
print(names(model_fwd_AIC$model))
```

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

```
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
extractAIC(model_bth_AIC)
```

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

```r
extractAIC(model_bth_BIC)
```

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

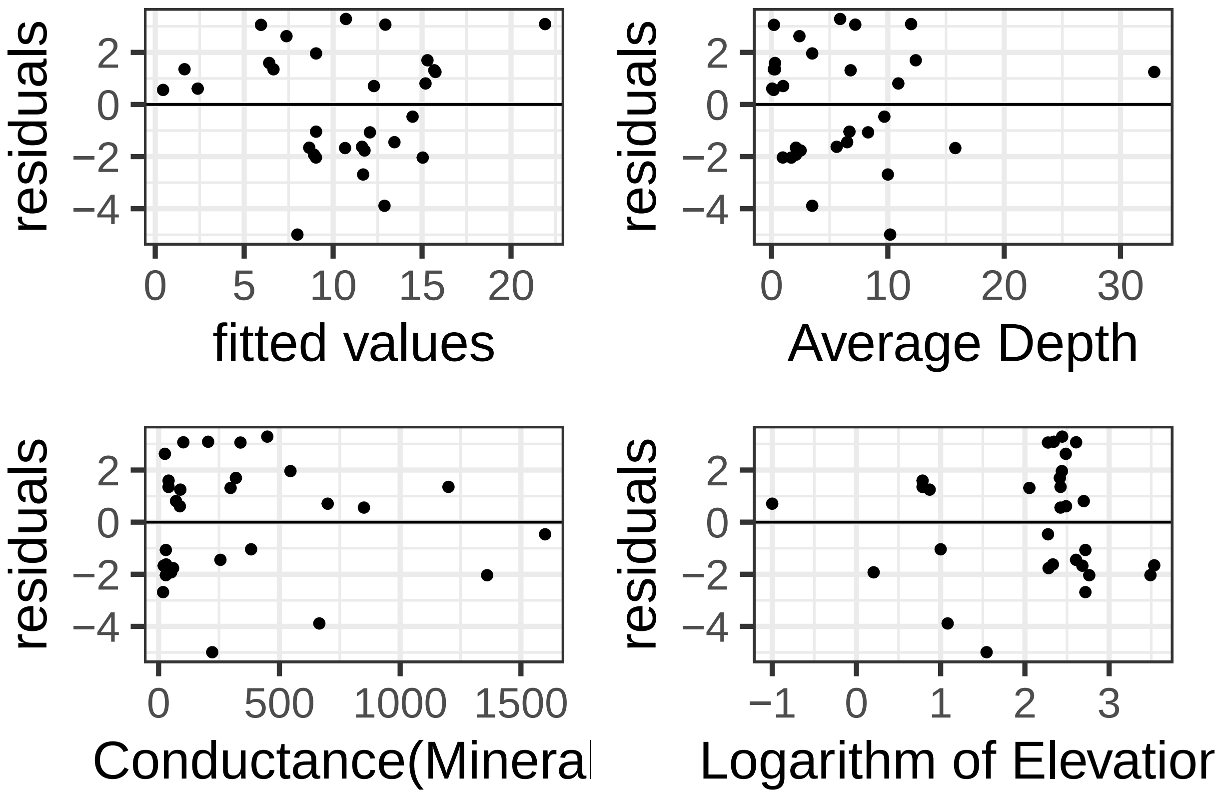
The lowest AIC is 61. We can determine the final model:

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

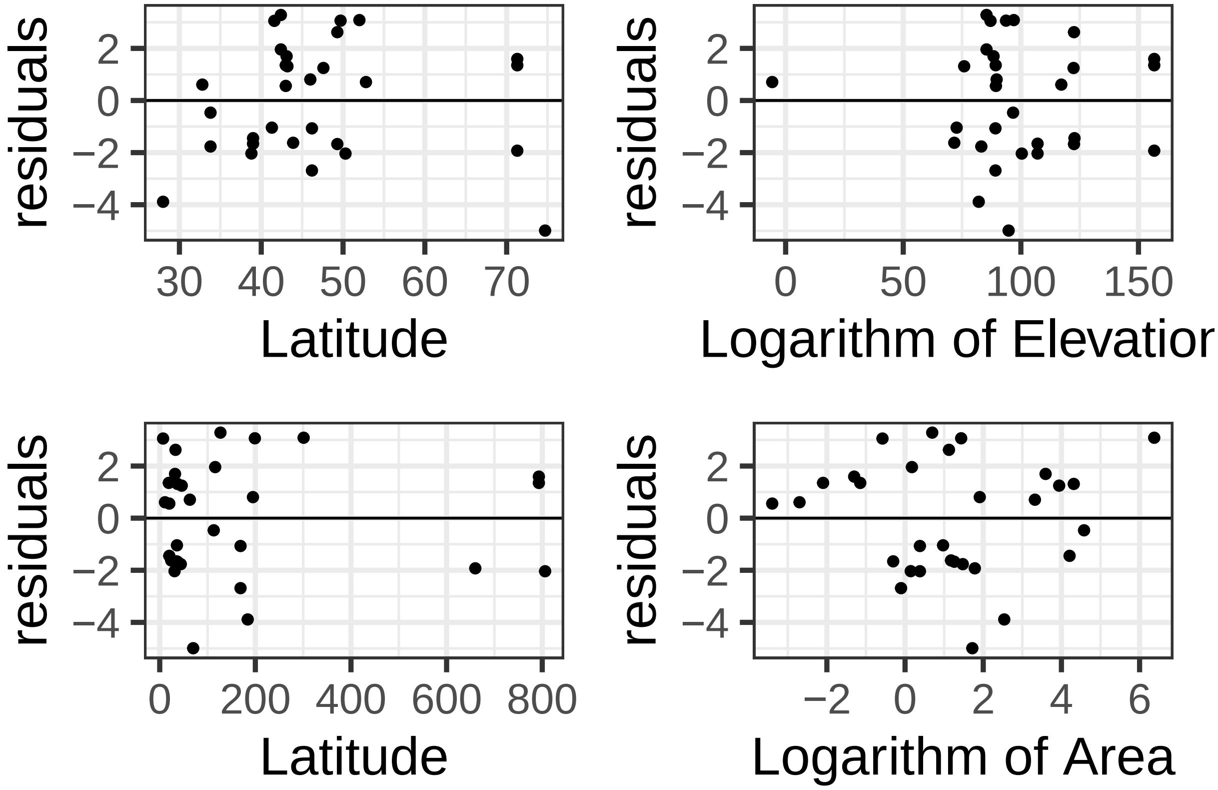
Diagnosis - Residual plots

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

```r
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)
```

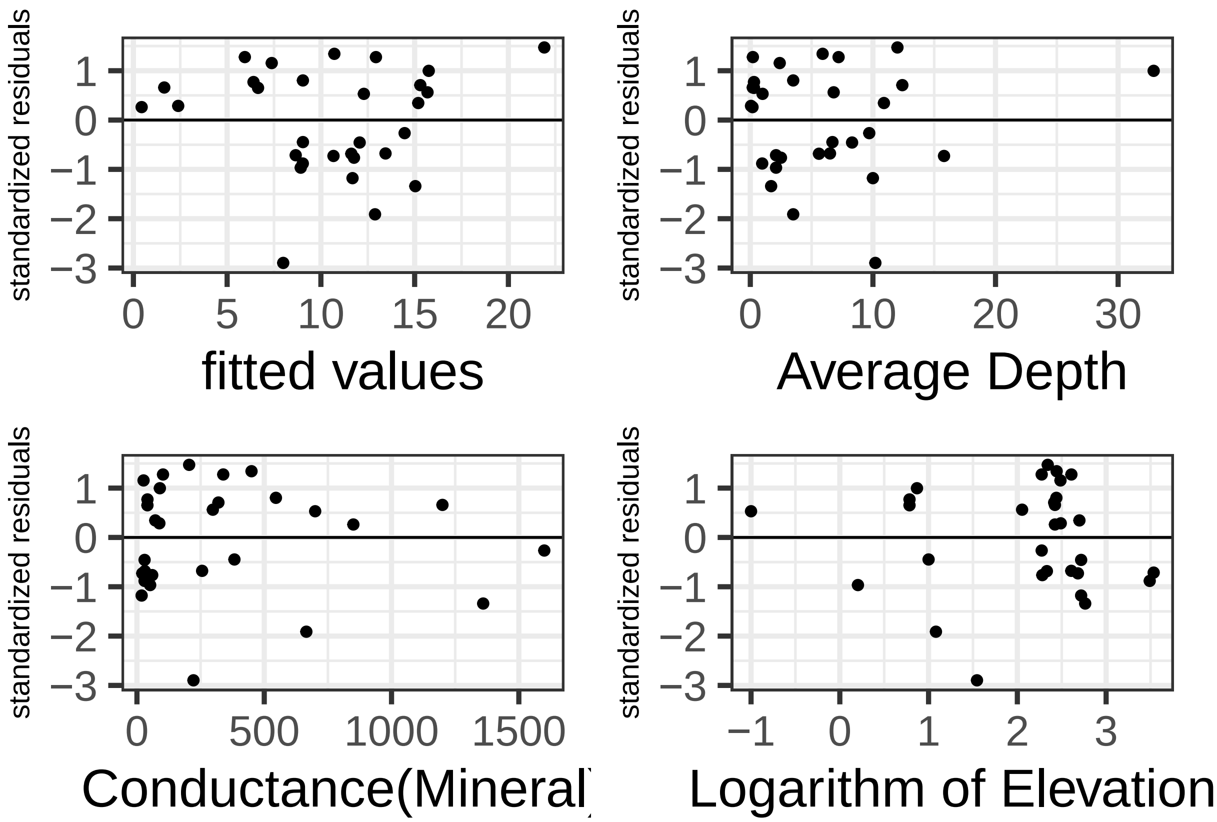



```
lat_plt = base + aes(x = Lat) + xlab("Latitude")
long_plt = base + aes(x = Long) + xlab("Logarithm of Elevation")
nLakes_plt = base + aes(x = NLakes) + xlab("Latitude")
logArea_plt = base + aes(x = logArea) + xlab("Logarithm of Area")
(lat_plt + long_plt) / (nLakes_plt + logArea_plt)
```

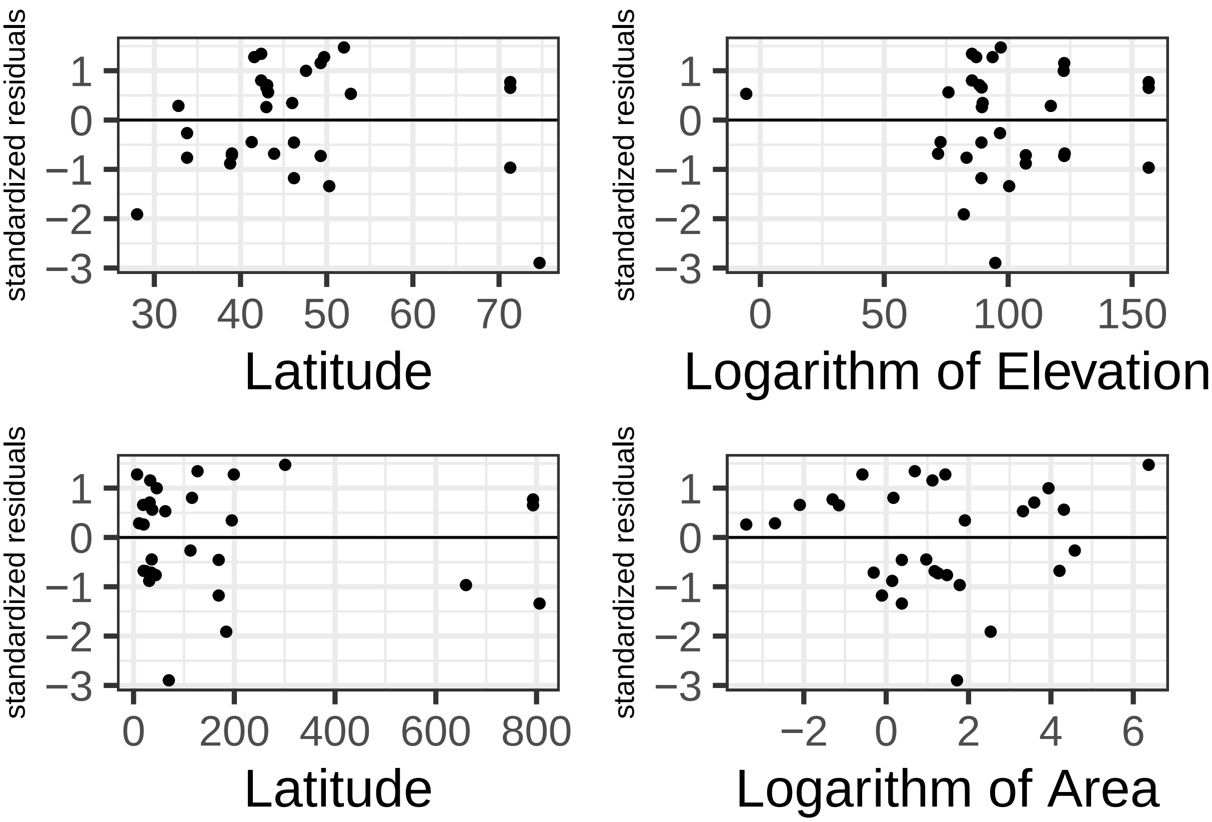


An obvious pattern that indicate the lack of linearlity does not show in null plot and the other residual plots.

```
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)
```

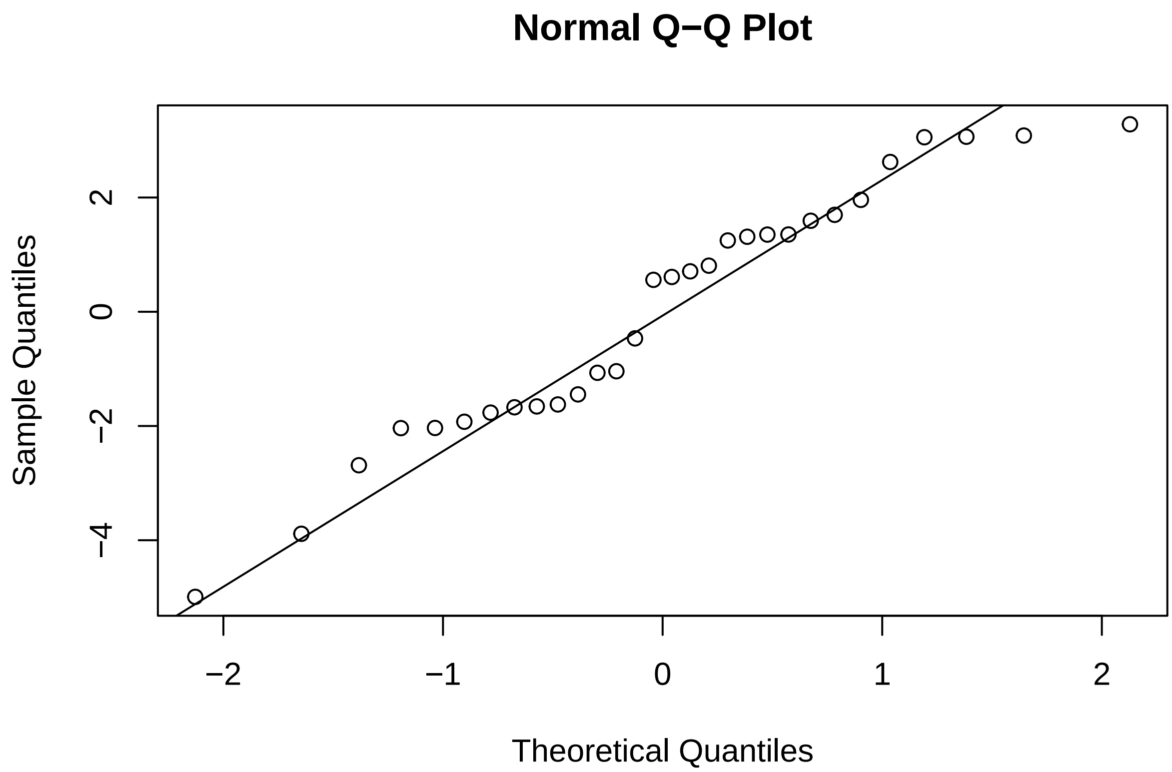



```r
lat_plt = base + aes(x = Lat) + xlab("Latitude")
long_plt = base + aes(x = Long) + xlab("Logarithm of Elevation")
nLakes_plt = base + aes(x = NLakes) + xlab("Latitude")
logArea_plt = base + aes(x = logArea) + xlab("Logarithm of Area")
(lat_plt + long_plt) / (nLakes_plt + logArea_plt)
```



Again, an obvious pattern was not indicated in the standardized residual plots.

```
qqnorm(lake.lm$residuals)
qqline(lake.lm$residuals)
```



Except for an outlier, the qqlot shows normality of the model.

5. Interpretation

```
summary(lake.lm)$sigma^2
```

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

Additional meter in the average lake depth is associated with 0.2 increase in the average number of crustacean species.

Additional micro Siemans in the conductance is associated with 0.002 decrease in the average number of crustacean species.

10 meters increase in the elevation is associated with 1.50 increase in the average number of crustacean species.

A degree increase in the north latitude is associated with 0.09 decrease in the average number of crustacean species.

Additional degree in the west longitude is associated with 0.07 decrease in the average number of crustacean species.

Additional lake within 20km is associated with 0.013 increase in the average number of crustacean species.

10 hectares increase in the surface area is associated with 1.56 increase in the average number of crustacean species.

The data analysis cannot find a meaningful meaningful linear relationship between rate of photosynthesis and number of crustacean species.

Prediction

```
predict(lake.lm, newdata = data.frame(MeanDepth = 153, Cond = 167, logElev = log10(372 + 1.1), Lat = 46
```

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

A 95% prediction interval; the estimated number of crustceans species is 52.23, and the interval is (23.26, 81.20).