605-Wk11-Discussion

Jose Mawyin

11/9/2019

Linear Regression of Factors Important to Lake Population of Zooplankton Species

1. Introduction
2. Data
3. Linear Regresssion Using All Factors
4. Backward Elimination of Factors
5. Comments

1.Introduction

This is a case study of how to use linear regression to see how we different independent variables affect a system's response. In this case, the system's response is the number of crustacean zooplankton species and the independent variables are 10 different parameters gathered at 69 world lakes.

2.Data

These data give the number of known crustacean zooplankton species for 69 world lakes. Also included are a number of characteristics of each lake. There are missing values.

Format

This data frame uses lake name as row label and contains the following columns:

Species: Number of zooplankton species

MaxDepth: Maximum lake depth, m

MeanDepth: Mean lake depth, m

Cond: Specific conductance, micro Siemans

Elev: Elevation, m

Lat: N latitude, degrees

Long: W longitude, degrees

Dist: distance to nearest lake, km

NLakes: number of lakes within 20 km

Photo: Rate of photosynthesis, mostly by the 14C method

Area: Lake area, in hectares

Source

Dodson, S. (1992), Predicting curstacean zooplankton species richness, Limnology and Oceanography, 37, 848–856.

```
data(lakes)
head(lakes)
```

```
##             Species MaxDepth MeanDepth Cond Elev  Lat  Long Dist NLakes
## Superior         30      406       148   79  185 47.5  88.0 0.12   2183
## Michigan         32      281        84  226  180 44.0  87.0 0.12    633
## Great_Slave      19      613        73  215  159 62.0 113.0 0.12   8805
## Erie             31       64        17  242  178 42.0  81.0 0.12    105
## Winnipeg         25       38        12  205  219 52.0  97.0 0.12    301
## Tahoe            12      404       313   92 1914 39.1 120.1 0.25     87
##             Photo    Area
## Superior      200 8240000
## Michigan      550 5800000
## Great_Slave    11 2860000
## Erie          696 2580000
## Winnipeg       66 2370000
## Tahoe         219   48800
```

3. Linear Regresssion Using All Factors

First, let's try linar regression using all 10 independent variables: Area, Cond, Dist, Elev, Lat, Long, MaxDepth, MeanDepth, NLakes and Photo

```
attach(lakes)
Species.lm <- lm(Species ~  Area+ Cond+ Dist+ Elev+ Lat+ Long+ MaxDepth+ MeanDepth+ NLakes+ Photo+ Spec
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared
## on the right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 11 in
## model.matrix: no columns are assigned
```

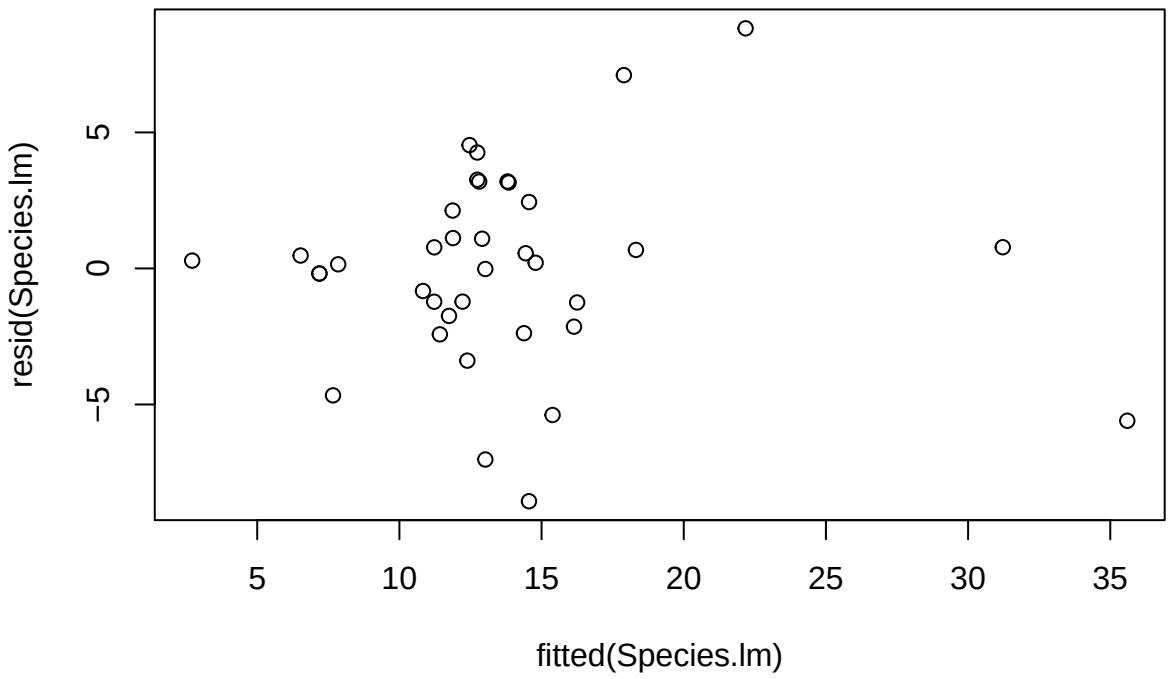
We see how all 10 variables account for 73% percent of the data's variation. The residuals, difference between the actual measured value stored in the data frame and the value that the fitted regression line predicts for that corresponding data point, plot appear evenly spaced around 0. Our quantile-versus-quantile (Q-Q) plot follow a straight line.

```
summary(Species.lm)
```

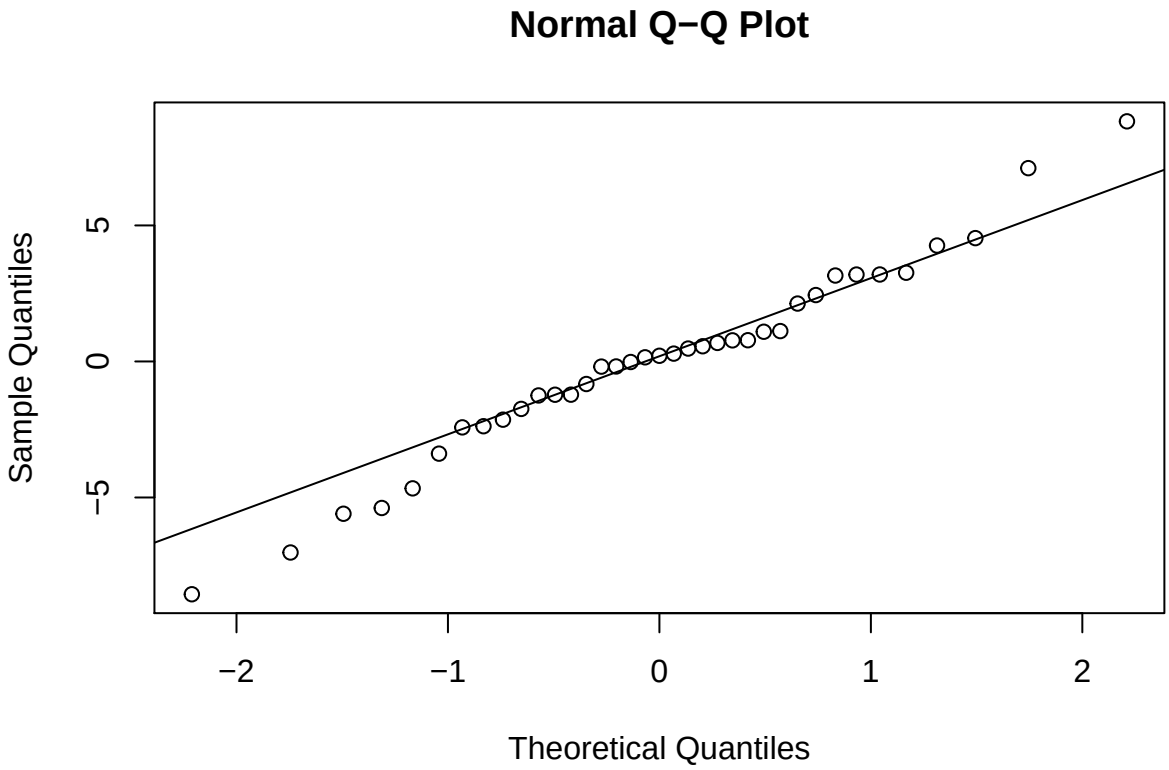
```
##
## Call:
## lm(formula = Species ~ Area + Cond + Dist + Elev + Lat + Long +
##     MaxDepth + MeanDepth + NLakes + Photo + Species)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5572 -1.7447  0.2083  2.1274  8.8253
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.097e+01  6.098e+00   3.439  0.00198 **
## Area         2.572e-06  6.944e-07   3.704  0.00101 **
## Cond        -1.784e-04  2.528e-03  -0.071  0.94428
## Dist        -7.282e-01  5.243e-01  -1.389  0.17666
```

```
## Elev        -1.904e-03  1.004e-03  -1.896  0.06915 .
## Lat         -1.696e-01  1.046e-01  -1.621  0.11711
## Long        -8.350e-04  3.443e-02  -0.024  0.98084
## MaxDepth     1.372e-02  3.230e-02   0.425  0.67465
## MeanDepth   -1.699e-02  4.583e-02  -0.371  0.71390
## NLakes      -7.002e-04  1.810e-03  -0.387  0.70206
## Photo        2.461e-03  2.637e-03   0.933  0.35921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.224 on 26 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.7354, Adjusted R-squared:  0.6337
## F-statistic: 7.227 on 10 and 26 DF,  p-value: 2.407e-05
```

```
plot(fitted(Species.lm),resid(Species.lm))
```




```
qqnorm(resid(Species.lm))
qqline(resid(Species.lm))
```



Let's see how the removal of some of the independent variables affect the linear regressio model.

4. Removal of Factors

Removing Longitude (degrees)

Longitude spcefies the east–west position of a point on the Earth's surface.

```
Species.lm <- update(Species.lm, .~. - Long, data=lakes)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared
## on the right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 10 in
## model.matrix: no columns are assigned
```

```
summary(Species.lm)
```

```
##
## Call:
## lm(formula = Species ~ Area + Cond + Dist + Elev + Lat + MaxDepth +
##     MeanDepth + NLakes + Photo + Species, data = lakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5571 -1.7621  0.1997  2.1330  8.8234
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.090e+01  5.358e+00   3.901 0.000574 ***
## Area         2.575e-06  6.677e-07   3.857 0.000646 ***
## Cond        -1.803e-04  2.479e-03  -0.073 0.942557
## Dist        -7.253e-01  5.012e-01  -1.447 0.159361
## Elev        -1.909e-03  9.638e-04  -1.981 0.057884 .
## Lat         -1.699e-01  1.015e-01  -1.674 0.105753
## MaxDepth     1.365e-02  3.158e-02   0.432 0.669078
## MeanDepth   -1.699e-02  4.497e-02  -0.378 0.708524
## NLakes      -6.972e-04  1.772e-03  -0.393 0.697125
## Photo        2.479e-03  2.489e-03   0.996 0.328062
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.146 on 27 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.7354, Adjusted R-squared:  0.6472
## F-statistic: 8.339 on 9 and 27 DF,  p-value: 7.901e-06
```

Removing Specific conductance (micro Siemans)

Specific conductance is a measure of a solution's ability to conduct electricity.

```
Species.lm <- update(Species.lm, .~. - Cond, data=lakes)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared
## on the right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 9 in
## model.matrix: no columns are assigned
```

```
summary(Species.lm)
```

```
##
## Call:
## lm(formula = Species ~ Area + Dist + Elev + Lat + MaxDepth +
##     MeanDepth + NLakes + Photo + Species, data = lakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0858 -1.7569 -0.1199  1.9633  9.6562
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.911e+01  4.529e+00   4.221 0.000146 ***
## Area         2.472e-06  6.238e-07   3.963 0.000315 ***
## Dist        -8.844e-01  4.581e-01  -1.931 0.060992 .
## Elev        -2.069e-03  8.051e-04  -2.570 0.014212 *
## Lat         -1.296e-01  7.893e-02  -1.642 0.108783
## MaxDepth     2.379e-02  2.747e-02   0.866 0.391933
## MeanDepth   -2.707e-02  3.951e-02  -0.685 0.497402
## NLakes      -1.380e-03  1.562e-03  -0.884 0.382272
## Photo        1.224e-03  2.005e-03   0.610 0.545358
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.008 on 38 degrees of freedom
##   (22 observations deleted due to missingness)
## Multiple R-squared:  0.7079, Adjusted R-squared:  0.6464
## F-statistic: 11.51 on 8 and 38 DF,  p-value: 4.064e-08
```

Removing Rate of Photosynthesis

```
Species.lm <- update(Species.lm, .~. - Photo, data=lakes)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared
## on the right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 8 in
## model.matrix: no columns are assigned
```

```
summary(Species.lm)
```

```
##
## Call:
## lm(formula = Species ~ Area + Dist + Elev + Lat + MaxDepth +
##     MeanDepth + NLakes + Species, data = lakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9278 -1.8682 -0.5028  2.6050 11.3907
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.497e+01  3.082e+00   4.858 8.65e-06 ***
## Area         2.253e-06  6.212e-07   3.627 0.000588 ***
## Dist        -1.336e+00  3.500e-01  -3.817 0.000318 ***
## Elev        -1.812e-03  6.172e-04  -2.937 0.004674 **
## Lat         -5.848e-02  6.340e-02  -0.922 0.359903
## MaxDepth     4.293e-02  2.628e-02   1.633 0.107566
## MeanDepth   -4.248e-02  4.032e-02  -1.054 0.296144
## NLakes      -2.470e-03  1.498e-03  -1.649 0.104201
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.357 on 61 degrees of freedom
## Multiple R-squared:   0.61,  Adjusted R-squared:  0.5653
## F-statistic: 13.63 on 7 and 61 DF,  p-value: 1.825e-10
```

Removing Latitude (degrees)

The geographic coordinate that specifies the north–south position of a point on the Earth's surface.

```
Species.lm <- update(Species.lm, .~. - Lat, data=lakes)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared
## on the right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 7 in
## model.matrix: no columns are assigned
```

```
summary(Species.lm)
```

```
##
## Call:
## lm(formula = Species ~ Area + Dist + Elev + MaxDepth + MeanDepth +
##     NLakes + Species, data = lakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7609 -1.9114 -0.0831  2.9181 11.5200
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.223e+01  8.007e-01  15.267  < 2e-16 ***
## Area         2.261e-06  6.204e-07   3.645 0.000549 ***
## Dist        -1.295e+00  3.467e-01  -3.734 0.000412 ***
## Elev        -1.659e-03  5.936e-04  -2.795 0.006904 **
## MaxDepth     4.715e-02  2.585e-02   1.824 0.072953 .
## MeanDepth   -4.978e-02  3.948e-02  -1.261 0.212117
## NLakes      -2.855e-03  1.437e-03  -1.988 0.051268 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.351 on 62 degrees of freedom
## Multiple R-squared:  0.6046, Adjusted R-squared:  0.5663
## F-statistic:  15.8 on 6 and 62 DF,  p-value: 6.48e-11
```

At the end of this process of eliminaition we have seen how the three most important independent variables that predict the number of lake species are: Lake area, distance to nearest lake and Elevation with all these variables having a P value less than our critical value of 0.05.

5. Comments

What happens when remove the next two independent variables with P values that do not meet our critical values?

Removing Mean Lake Depth (m)

```
Species.lmA <- update(Species.lm, .~. - MeanDepth, data=lakes)
```

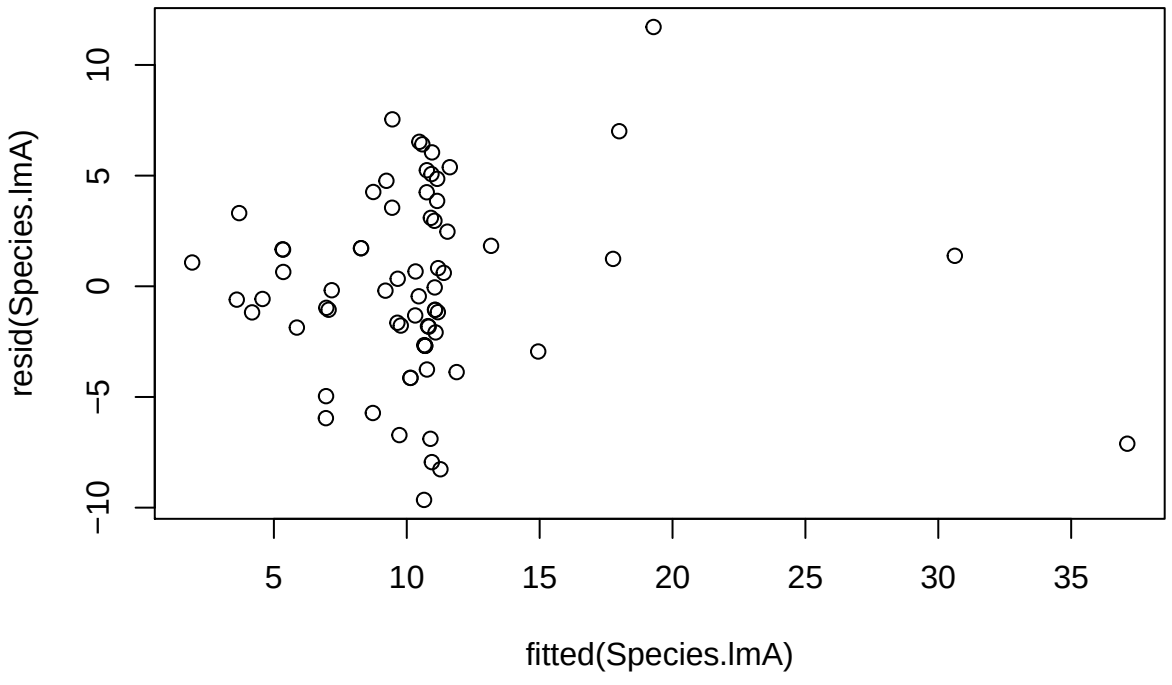
```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared
## on the right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 6 in
## model.matrix: no columns are assigned
```

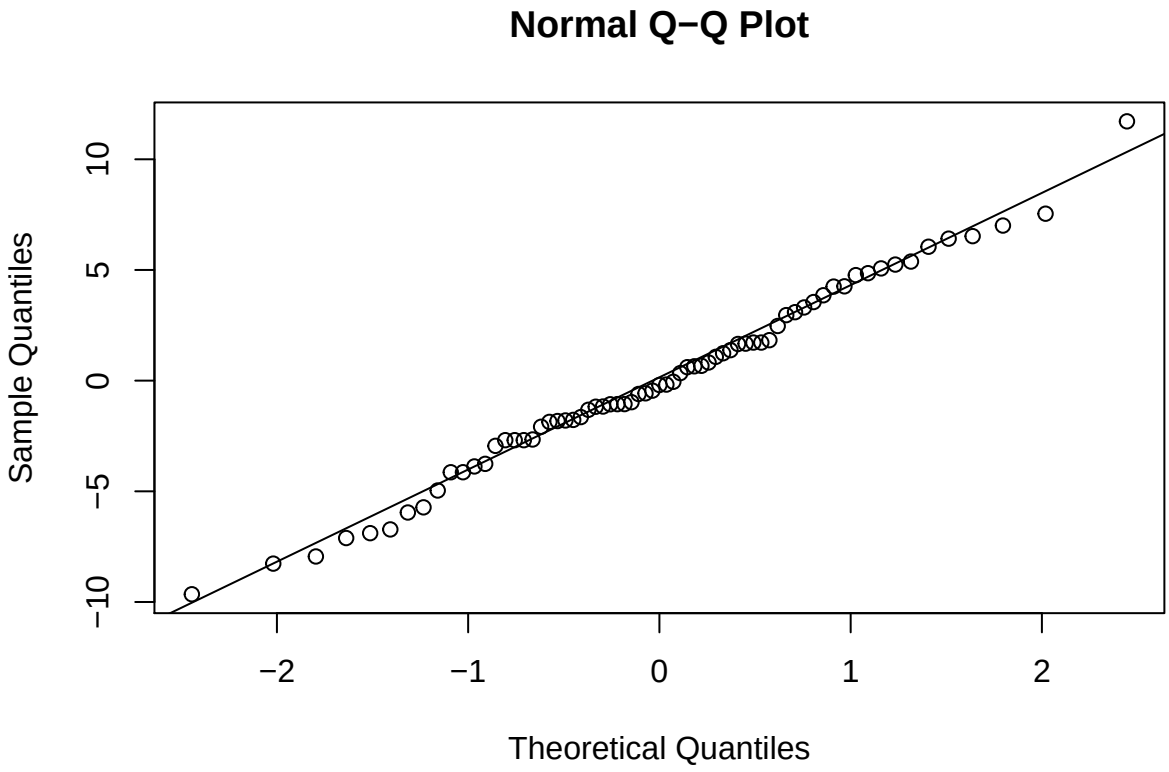
```r
summary(Species.lmA)
```

```
##
## Call:
## lm(formula = Species ~ Area + Dist + Elev + MaxDepth + NLakes +
##     Species, data = lakes)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.651 -2.659 -0.197  2.960 11.714
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.199e+01  7.828e-01  15.319  < 2e-16 ***
## Area         2.653e-06  5.396e-07   4.916 6.62e-06 ***
## Dist        -1.165e+00  3.327e-01  -3.502 0.000854 ***
## Elev        -1.690e-03  5.958e-04  -2.836 0.006130 **
## MaxDepth     1.599e-02  7.609e-03   2.102 0.039572 *
## NLakes      -1.273e-03  7.026e-04  -1.812 0.074686 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.372 on 63 degrees of freedom
## Multiple R-squared:  0.5945, Adjusted R-squared:  0.5623
## F-statistic: 18.47 on 5 and 63 DF,  p-value: 3e-11
```

```r
plot(fitted(Species.lmA),resid(Species.lmA))
```




```r
qqnorm(resid(Species.lmA))
qqline(resid(Species.lmA))
```



Our P values improved.

Let's see what happens when we remove the next critical value.

Removing Number of Lakes within 20 km

```
Species.lmB <- update(Species.lm, .~. - NLakes, data=lakes)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared
## on the right-hand side and was dropped
```

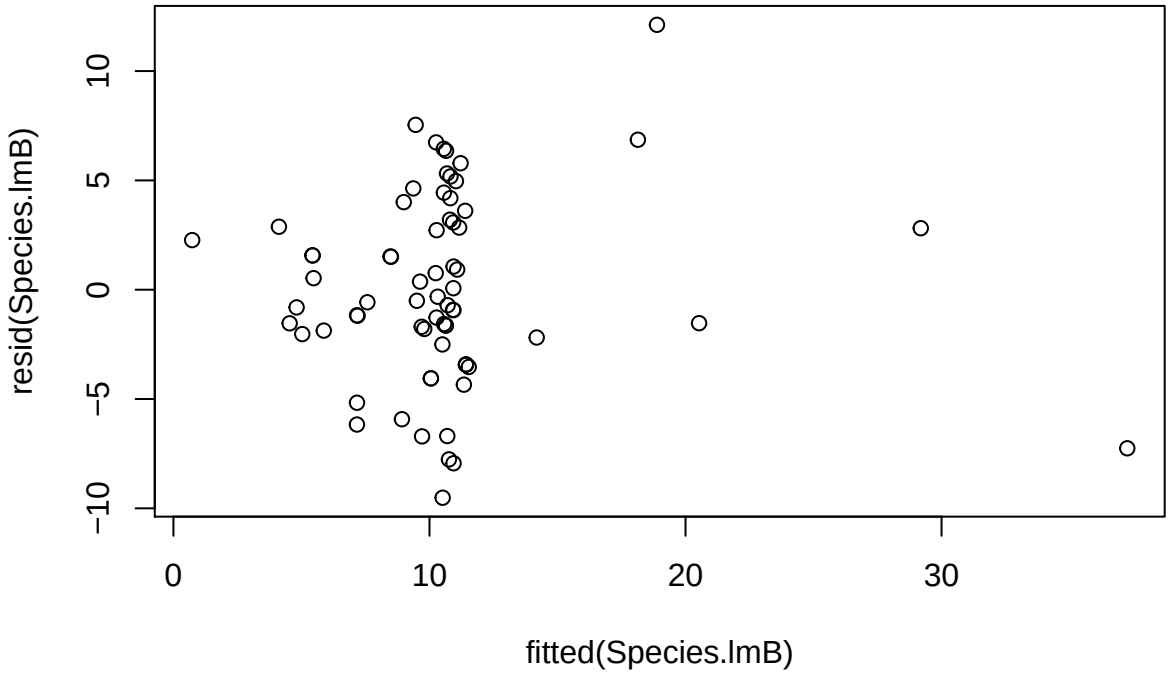
```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 6 in
## model.matrix: no columns are assigned
```

```
summary(Species.lmB)
```

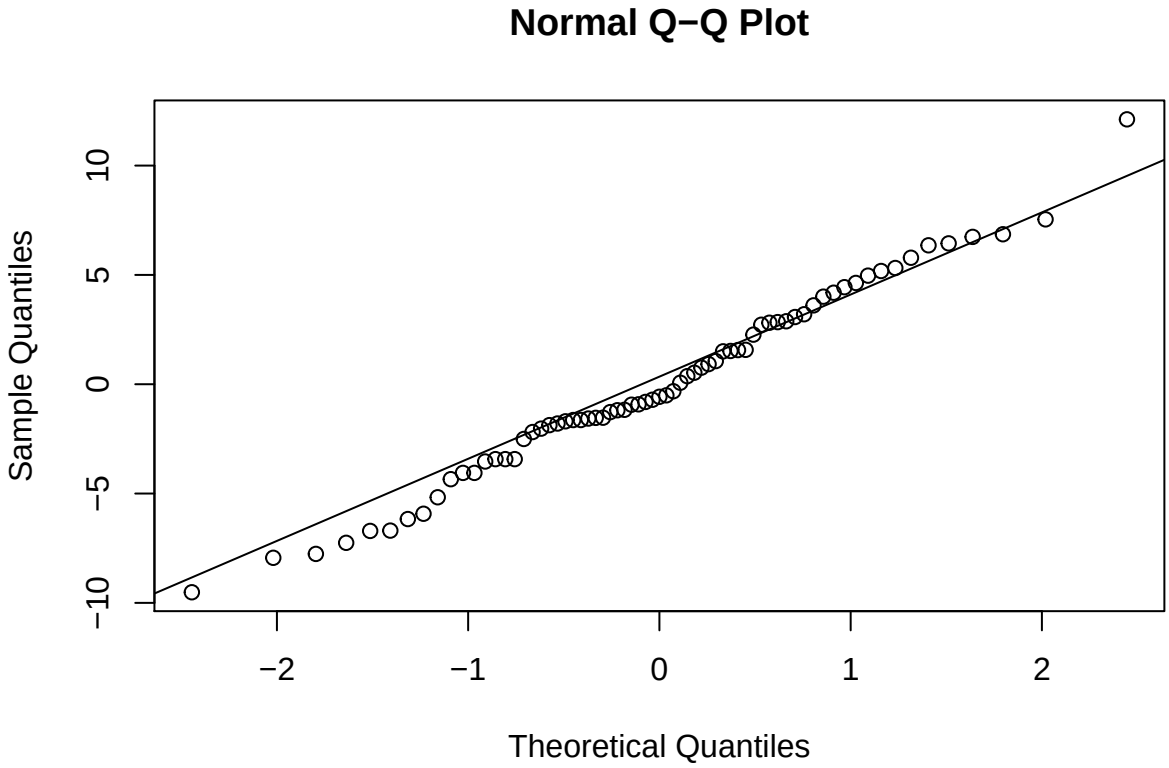
```
##
## Call:
## lm(formula = Species ~ Area + Dist + Elev + MaxDepth + MeanDepth +
##     Species, data = lakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5146 -2.1875 -0.5741  2.8801 12.1150
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.168e+01  7.688e-01  15.186  < 2e-16 ***
## Area         2.838e-06  5.610e-07   5.059 3.89e-06 ***
```

```
## Dist        -9.840e-01  3.166e-01  -3.108  0.00283 **
## Elev        -1.613e-03  6.068e-04  -2.658  0.00996 **
## MaxDepth    -4.213e-04  9.990e-03  -0.042  0.96649
## MeanDepth    1.878e-02  1.967e-02   0.955  0.34334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.452 on 63 degrees of freedom
## Multiple R-squared:  0.5794, Adjusted R-squared:  0.546
## F-statistic: 17.36 on 5 and 63 DF,  p-value: 9.12e-11
```

```
plot(fitted(Species.lmB),resid(Species.lmB))
```




```
qqnorm(resid(Species.lmB))
qqline(resid(Species.lmB))
```



Our critival P values increased slighlty. There is a limit to this eliminatin process at which the model does not improve even when including factors with higher than critical value P-factors.