PSTAT 126 Project 4

Wilber Delgado & Mason Delan

12-13-2023

```
library("readxl")
WHR = read_excel("D:/WHR2023.xls")
```

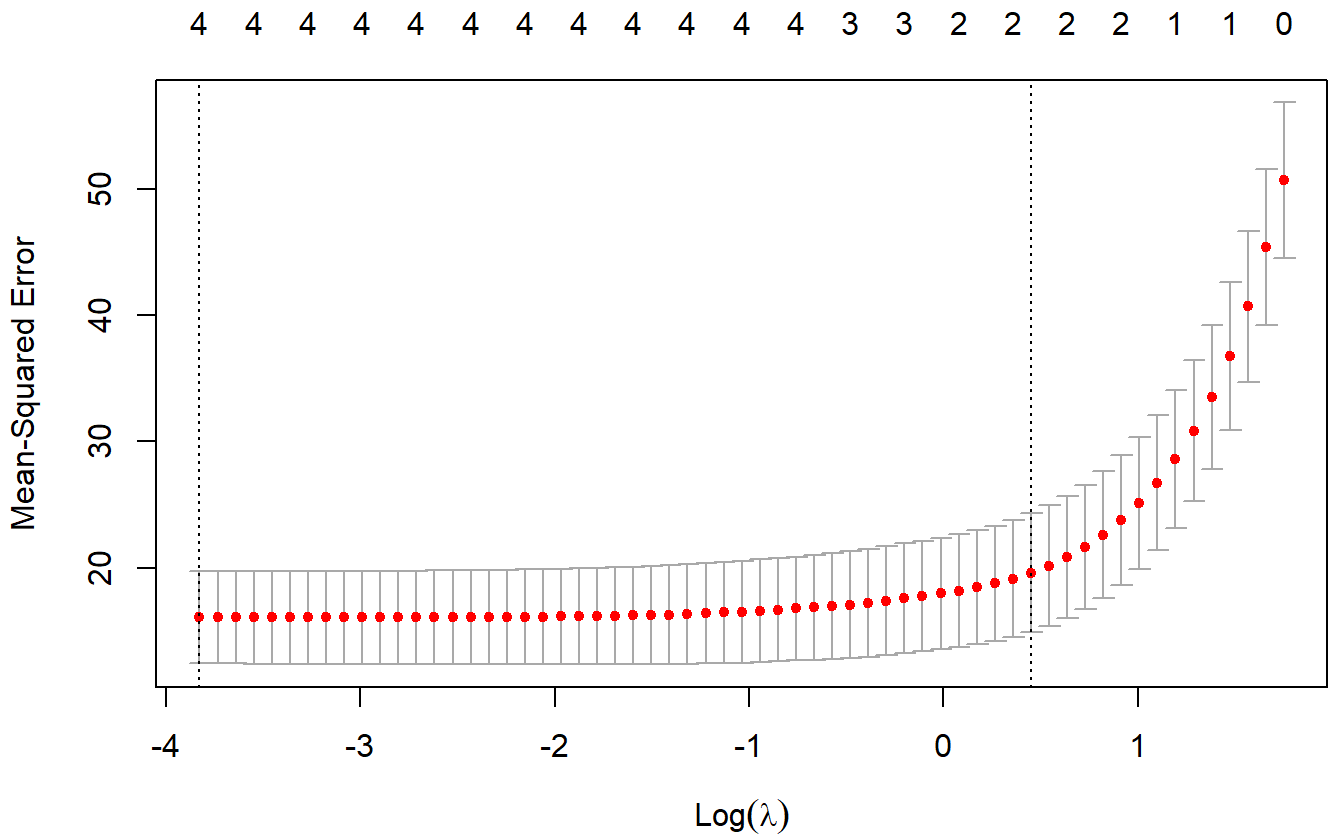
For our project we are working with the World Happiness Report, which is data gathered from around the world, in which people rated various aspects of their lives and experiences such as Freedom to make life choices, Perceptions of corruption, quality of life, economic experiences, etc. We retrieved this data set from https://worldhappiness.report/data/ (https://worldhappiness.report/data/). Because the data set is so big and also in the form of a Time Series dataset, we decided to drop the countries from the original data and the gather a random sample of 500 which can represent the average response of each country during any given year.

Put some data to the side

```
##
## Call:
## lm(formula = `Healthy life expectancy at birth` ~ year + `Life Ladder` +
##     `Log GDP per capita` + `Negative affect`, data = training_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.037  -1.466   0.292   1.954  14.029
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -270.94242   88.05331  -3.077 0.002236 **
## year                    0.14262    0.04393   3.247 0.001266 **
## `Life Ladder`           1.66570    0.29096   5.725 2.05e-08 ***
## `Log GDP per capita`    3.80328    0.28589  13.303  < 2e-16 ***
## `Negative affect`       8.10927    2.41066   3.364 0.000843 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.948 on 396 degrees of freedom
## Multiple R-squared:  0.6974, Adjusted R-squared:  0.6943
## F-statistic: 228.2 on 4 and 396 DF,  p-value: < 2.2e-16
```

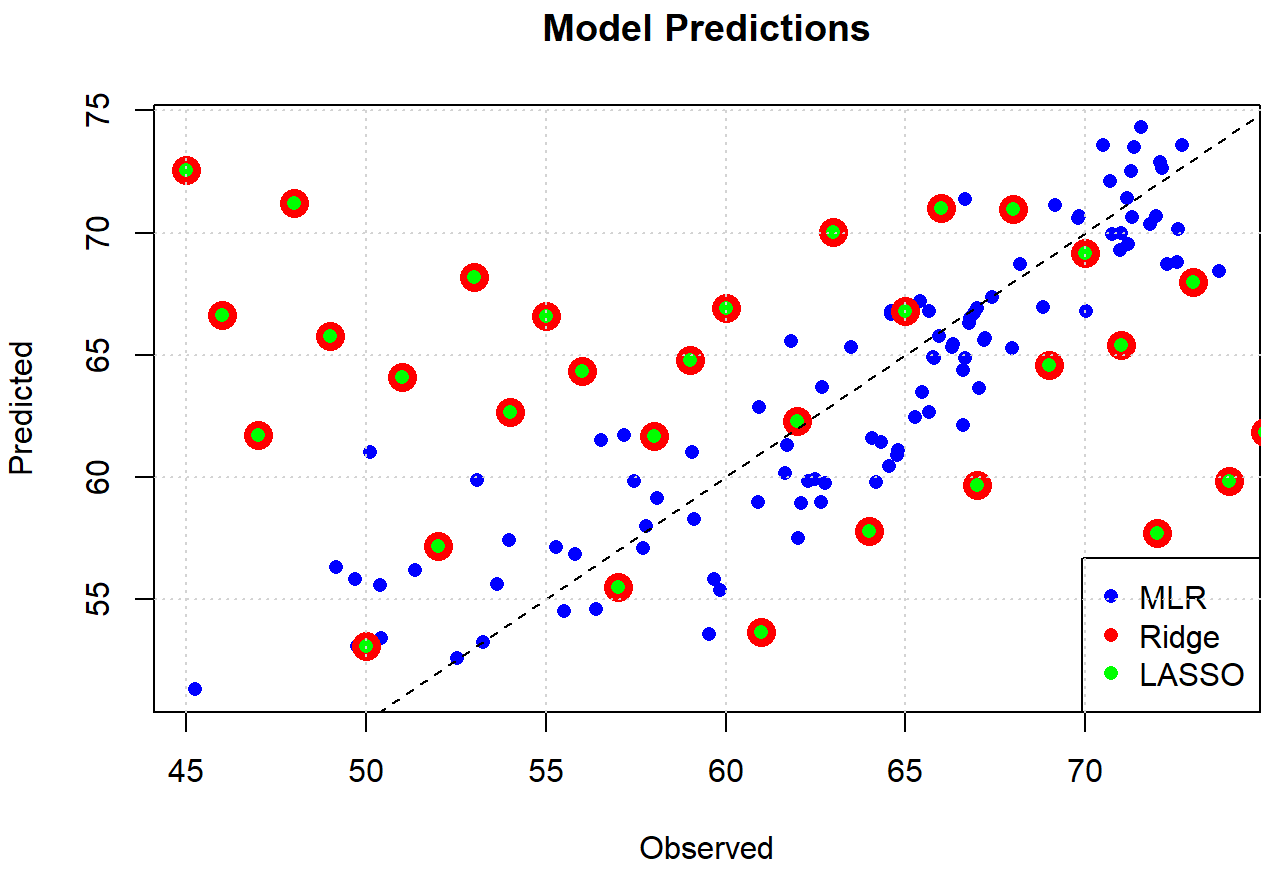
```
## Loaded glmnet 4.1-8
```

```
## [1] 0.02174512
```

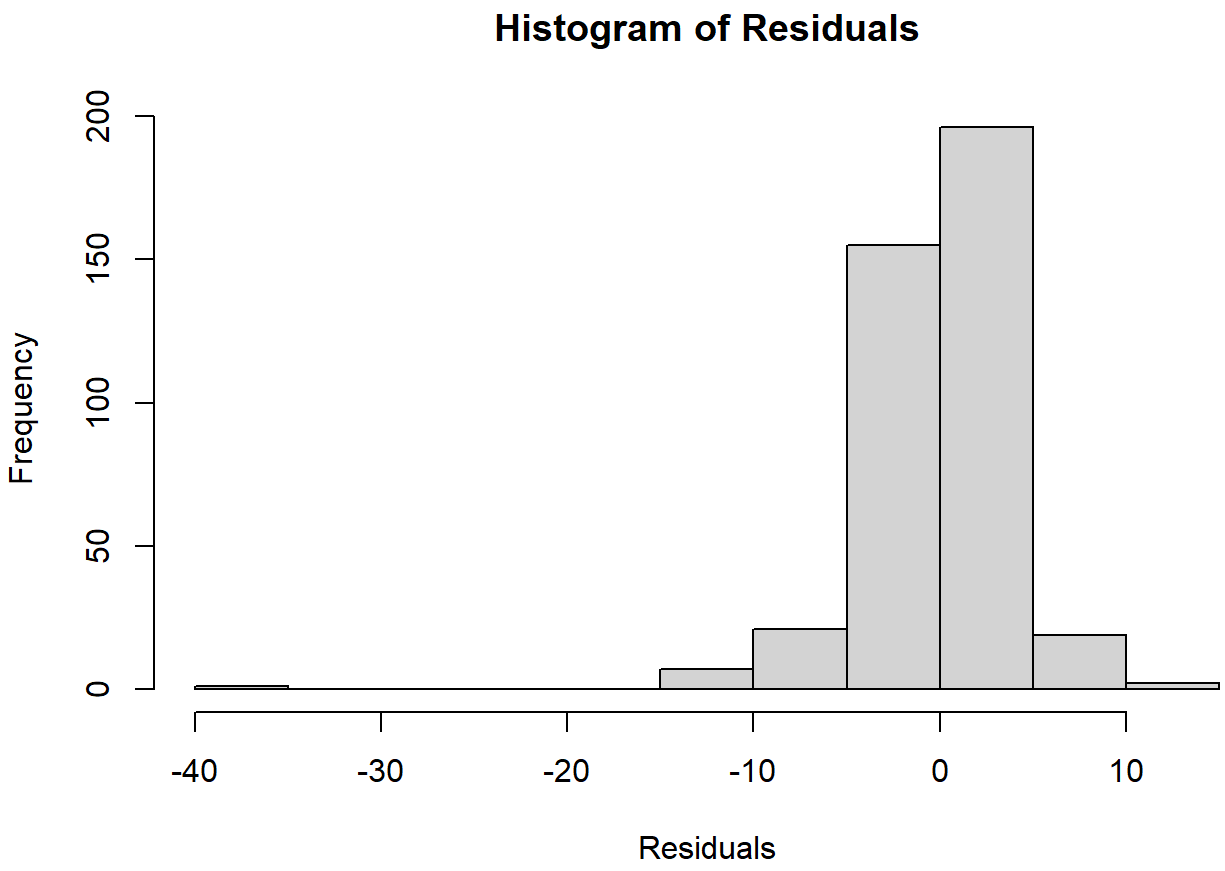


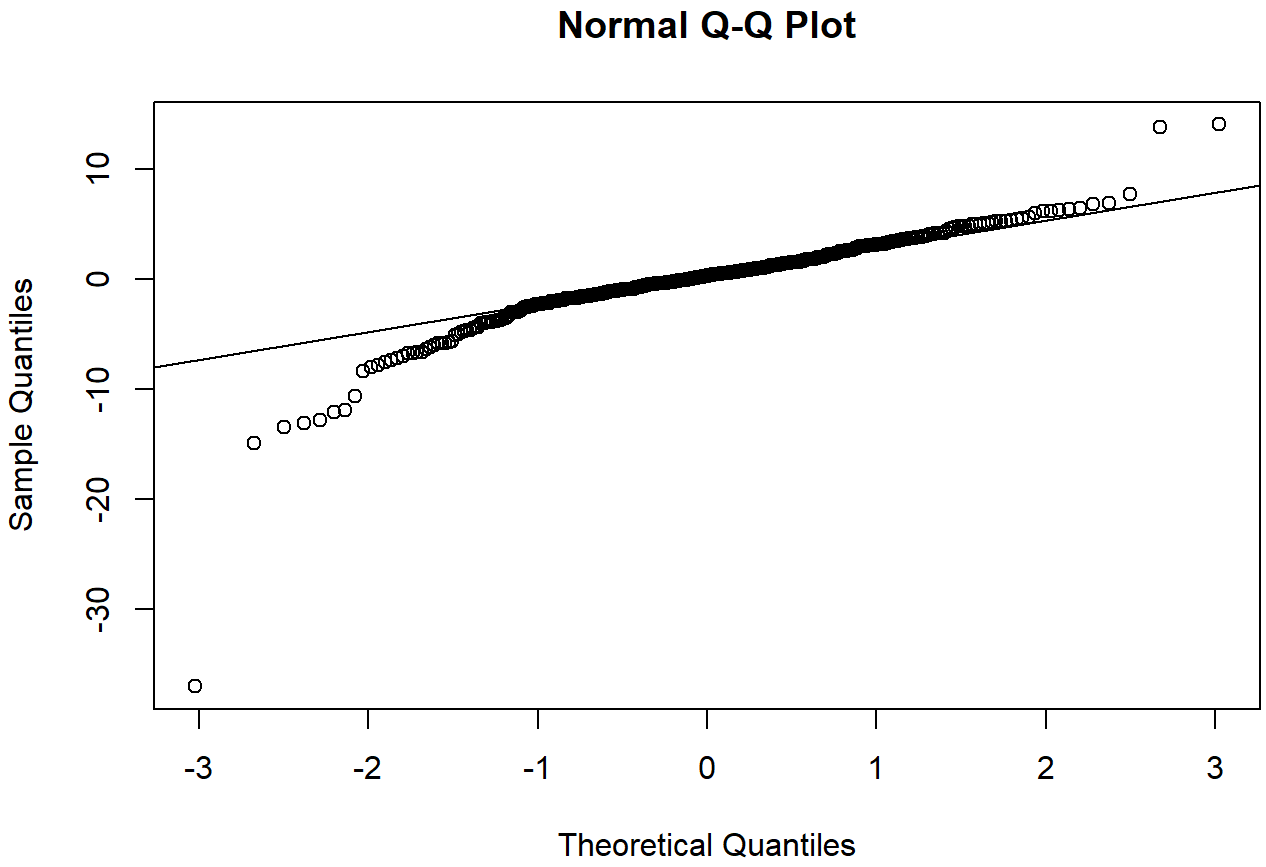

```
## Optimal Lambda for Ridge Regression: 0.5775694
```

```
## Optimal Lambda for LASSO: 0.02174512
```



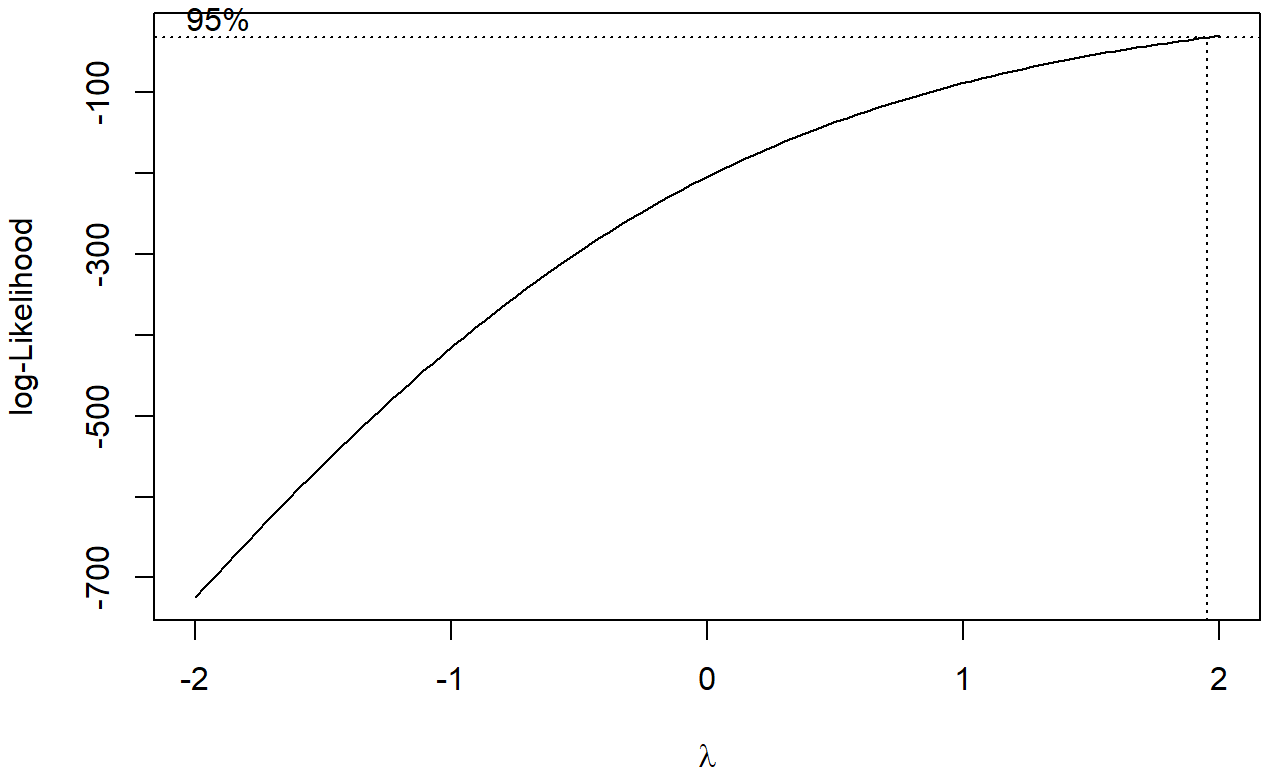
Based on the graph we can see that the Ridge and Lasso medthod are providing similar results. Based on our final MLR model from our previous project we found that the optimal Lambda for Ridge Regression: 0.5775694 and the optimal Lambda for LASSO: 0.02174512. In conclusion, it appears from the given lambda values and coefficient outputs that Ridge and LASSO produce comparable models at their respective optimal regularization strengths. This observation is consistent with the overlapping regions of the LASSO and Ridge graphs. Trade-offs between Ridge and LASSO may have to do with interpretability, model complexity, and how important sparsity is for feature selection.


Histogram of Residuals
Frequency
0
50
100
150
200
-40
-30
-20
-10
0
10
Residuals

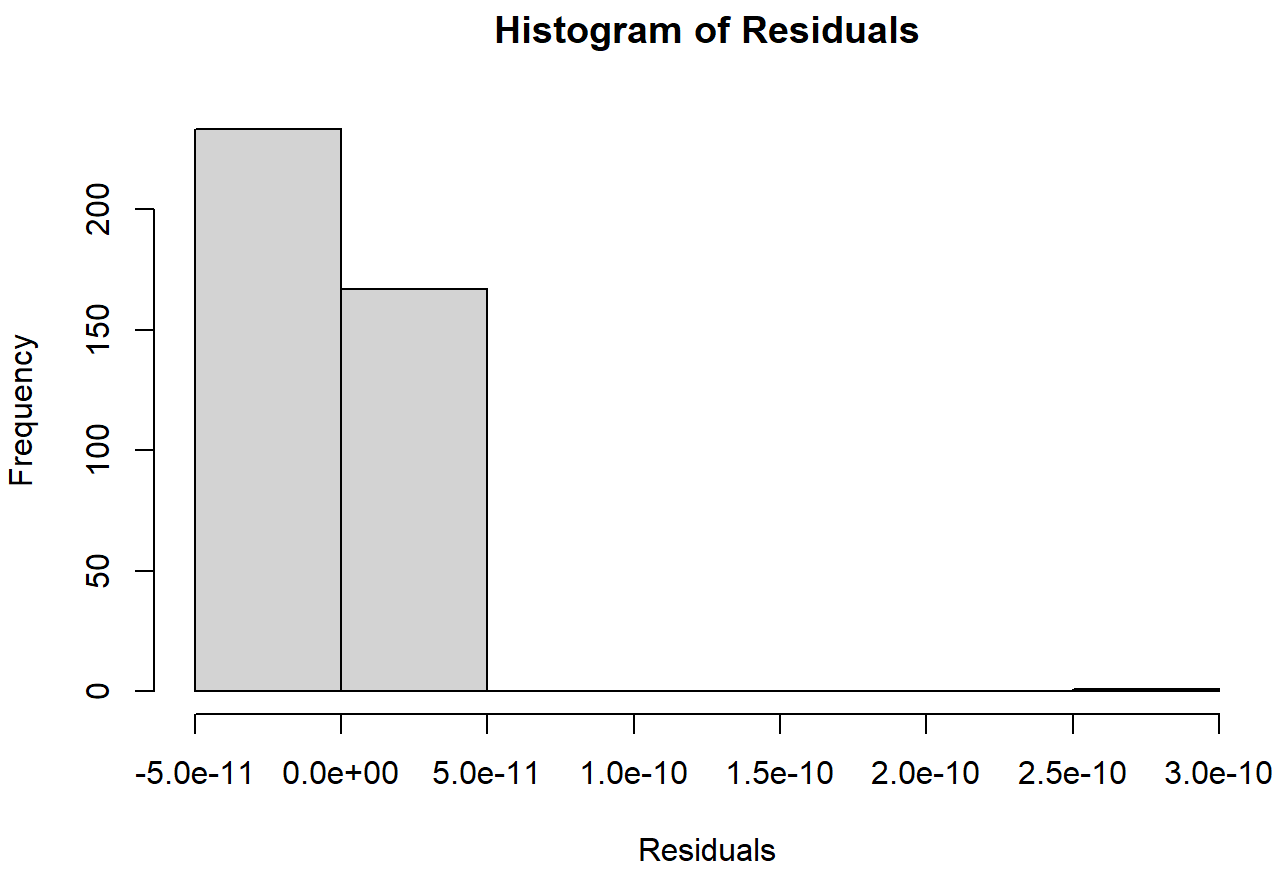

Normal Q-Q Plot
Sample Quantiles
-30
-20
-10
0
10
-3
-2
-1
0
1
2
3
Theoretical Quantiles

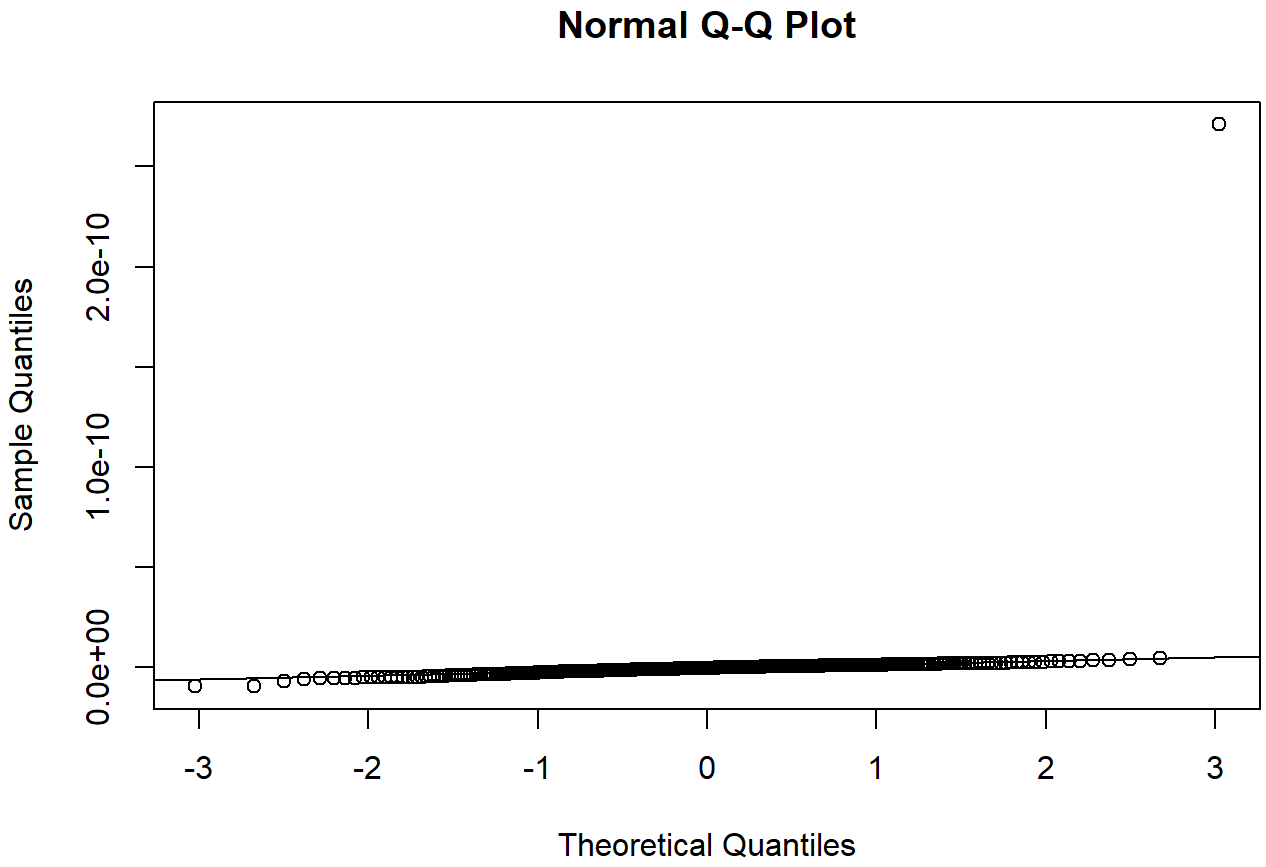
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.83853, p-value < 2.2e-16
```

Based on our original model we can see that the histogram of the residuals seems to be pretty normal. However in the QQ plot of the residuals, the residuals show a departure from normality, so we want to apply the Box Cox transformation to see if it can help the QQ plot display whole normality.The Box-Cox transformation is a method used to stabilize the variance and make a dataset more closely approximate a normal distribution




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


Histogram of Residuals
Frequency
0
50
100
150
200
-5.0e-11
0.0e+00
5.0e-11
1.0e-10
1.5e-10
2.0e-10
2.5e-10
3.0e-10
Residuals


Normal Q-Q Plot
Sample Quantiles
0.0e+00
1.0e-10
2.0e-10
-3
-2
-1
0
1
2
3
Theoretical Quantiles

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals2
## W = 0.090712, p-value < 2.2e-16
```

```
##
## Call:
## lm(formula = transformed_response ~ year + Life.Ladder + Log.GDP.per.capita +
##     Negative.affect, data = transformed_data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -9.563e-12 -1.800e-12 -3.530e-13  7.370e-13  2.710e-10
##
## Coefficients:
##                       Estimate Std. Error    t value Pr(>|t|)
## (Intercept)         1.501e+03  3.075e-10  4.881e+12   <2e-16 ***
## year                8.628e-14  1.534e-13  5.620e-01   0.5741
## Life.Ladder         1.695e-12  1.016e-12  1.668e+00   0.0960 .
## Log.GDP.per.capita -2.379e-12  9.984e-13 -2.382e+00   0.0177 *
## Negative.affect     1.177e-11  8.418e-12  1.398e+00   0.1629
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.379e-11 on 396 degrees of freedom
## Multiple R-squared:  0.5001, Adjusted R-squared:  0.4951
## F-statistic: 99.05 on 4 and 396 DF,  p-value: < 2.2e-16
```

After doing the Box Cox transformation on or linear model, we see that although the QQ plot does reflect much more normal residuals, we see that our histogram does not. Furthermore, looking at the summary of the transformed model, we can see that our variables become insignificant than our original model. We can also see that our R squared lowered by about 20%. With all of this combined we can conclude that our model is likely the best fit and does not require anymore transformation that could help make the QQ plot of the residuals look more normal. Lastly we can also see that the Box-Cox tranformed model failed the Shapiro-Wilk test as the p value is less than 0.05.

```
##       (Intercept)              year       Life.Ladder Log.GDP.per.capita
##      3.874300e+01      2.937409e-07      1.302025e-06               NaN
##    Negative.affect
##      3.430608e-06
```

The NaN result for the coefficient of Log GDP per Capita in the back-transformed model suggests a potential violation of the non-negativity condition. The Box-Cox transformation is not defined for zero or negative values, and when applied to Log GDP per Capita, it may have encountered issues with zero or negative values in the original data.