Bike Dataset Analysis Final Project

Maisy Song

12/11/2021

Bike Dataset Analysis

```
library(ggplot2)
library(faraway)
library(ISLR)
library(MASS)
```

```
directory = ("Documents/uiuc/finalproject2/")
bikesdata = read.csv("bike_clean.csv")
```

Cleaning Dataset

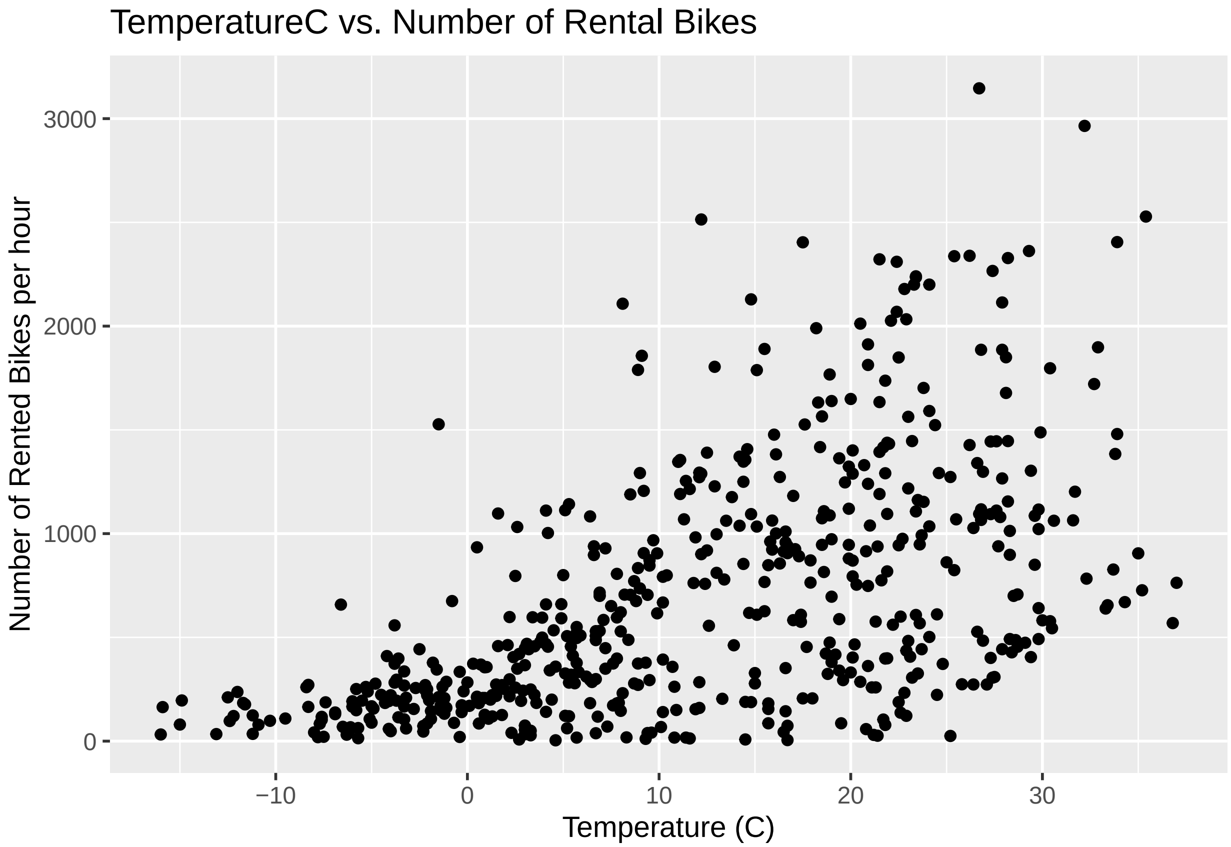
Getting Rid of unneccessary variables and choosing the ones I want to focus in this analysis.

```
bikes = subset(bikesdata, select = -c(X))
bikes = subset(bikes, select = c(Month, TemperatureC, Rented.Bike.Count, Hour,
                                 Humidity, Wind.speed.ms, Rainfallmm,
                                 Solar.Radiation..MJ.m2., Seasons))
names(bikes)
```

```
## [1] "Month"                     "TemperatureC"
## [3] "Rented.Bike.Count"         "Hour"
## [5] "Humidity"                  "Wind.speed.ms"
## [7] "Rainfallmm"                "Solar.Radiation..MJ.m2."
## [9] "Seasons"
```

I first wanted to visualize my data points with some of the variables. Decided to visualize TemperatureC vs. Rented.Bike.Count.

```
ggplot(bikes, mapping = aes(x = TemperatureC, y = Rented.Bike.Count)) +
  geom_point() +
  labs(title = "TemperatureC vs. Number of Rental Bikes",) +
  xlab("Temperature (C)") +
  ylab("Number of Rented Bikes per hour")
```



Models

Fitting a saturated model of all first order terms of all the variables.

```
firstmodel = lm(Rented.Bike.Count ~ ., bikes)
summary(firstmodel)
```

```
##
## Call:
## lm(formula = Rented.Bike.Count ~ ., data = bikes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -969.51 -243.30  -36.66  189.31 1706.60
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             846.2787   105.3258   8.035 6.06e-15 ***
## Month                    -0.6368     6.1374  -0.104 0.917401
## TemperatureC             22.6314     3.1608   7.160 2.69e-12 ***
## Hour                     27.8967     2.6639  10.472  < 2e-16 ***
## Humidity                 -8.6859     1.0606  -8.190 1.94e-15 ***
## Wind.speed.ms             5.6301    19.1779   0.294 0.769201
## Rainfallmm              -48.8568    12.7694  -3.826 0.000146 ***
## Solar.Radiation..MJ.m2.  -66.5336    26.2379  -2.536 0.011504 *
## SeasonsSpring           -189.7555    62.5571  -3.033 0.002537 **
## SeasonsSummer           -125.8787    67.4191  -1.867 0.062435 .
## SeasonsWinter           -464.4512    70.4962  -6.588 1.07e-10 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 393.6 on 533 degrees of freedom
## Multiple R-squared:  0.5745, Adjusted R-squared:  0.5665
## F-statistic: 71.97 on 10 and 533 DF,  p-value: < 2.2e-16
```

Here we can see that some of the variables are not significant at the alpha level of a = 0.1.

The second model is the following model with all the significant variables.

```
secondmodel = lm(Rented.Bike.Count ~ Hour + TemperatureC + Humidity +
                   Solar.Radiation..MJ.m2. + Rainfallmm + Seasons, bikes)
summary(secondmodel)
```

```
##
## Call:
## lm(formula = Rented.Bike.Count ~ Hour + TemperatureC + Humidity +
##     Solar.Radiation..MJ.m2. + Rainfallmm + Seasons, data = bikes)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -979.0 -246.5  -34.9  189.5 1709.0
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              850.747     86.645   9.819  < 2e-16 ***
## Hour                      28.004      2.616  10.706  < 2e-16 ***
## TemperatureC              22.573      3.142   7.183 2.29e-12 ***
## Humidity                  -8.753      1.035  -8.453 2.70e-16 ***
## Solar.Radiation..MJ.m2.  -64.545     25.314  -2.550 0.011057 *
## Rainfallmm               -48.523     12.700  -3.821 0.000149 ***
## SeasonsSpring           -184.118     49.450  -3.723 0.000217 ***
## SeasonsSummer           -122.498     63.082  -1.942 0.052678 .
## SeasonsWinter           -460.032     65.659  -7.006 7.38e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 392.9 on 535 degrees of freedom
## Multiple R-squared:  0.5744, Adjusted R-squared:  0.5681
## F-statistic: 90.27 on 8 and 535 DF,  p-value: < 2.2e-16
```

Since we will only be using the variables we have chosen for our second model, we need a new dataset. Here we create a new dataset called ‘newbikes’ with the variables from the second model. When deciding which interaction terms to use, I made multiple models to see which variables were the most significant and relevant when predicting Rented.Bike.Count.

```
newbikes = subset(bikes, select = c(Rented.Bike.Count, Hour, TemperatureC, Humidity,
                                    Solar.Radiation..MJ.m2., Rainfallmm, Seasons))
```

```r
second_order_model = lm(Rented.Bike.Count ~ .^2, newbikes)
summary(second_order_model)
```

```
##
## Call:
## lm(formula = Rented.Bike.Count ~ .^2, data = newbikes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1052.21  -164.21   -42.01   109.82  1531.63
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           3.730e+02  2.232e+02   1.671 0.095241
## Hour                                  3.540e+01  1.102e+01   3.213 0.001395
## TemperatureC                          3.140e+01  1.533e+01   2.048 0.041067
## Humidity                              9.579e-01  3.238e+00   0.296 0.767508
## Solar.Radiation..MJ.m2.              -2.794e+02  1.526e+02  -1.830 0.067759
## Rainfallmm                           -1.267e+03  1.160e+03  -1.092 0.275253
## SeasonsSpring                        -1.713e+02  2.312e+02  -0.741 0.459019
## SeasonsSummer                         9.606e+02  4.131e+02   2.325 0.020441
## SeasonsWinter                        -4.278e+02  2.611e+02  -1.639 0.101914
## Hour:TemperatureC                     1.752e+00  4.450e-01   3.937 9.41e-05
## Hour:Humidity                        -4.665e-01  1.478e-01  -3.157 0.001690
## Hour:Solar.Radiation..MJ.m2.          1.249e+01  8.347e+00   1.497 0.135047
## Hour:Rainfallmm                       4.795e+00  3.705e+00   1.294 0.196168
## Hour:SeasonsSpring                   -9.113e-02  7.367e+00  -0.012 0.990135
## Hour:SeasonsSummer                    8.755e-01  8.908e+00   0.098 0.921749
## Hour:SeasonsWinter                    2.518e+00  9.215e+00   0.273 0.784729
## TemperatureC:Humidity                -3.700e-01  1.908e-01  -1.939 0.052996
## TemperatureC:Solar.Radiation..MJ.m2. -8.305e+00  3.745e+00  -2.218 0.027008
## TemperatureC:Rainfallmm              -1.455e+01  1.487e+01  -0.978 0.328319
## TemperatureC:SeasonsSpring            5.719e+00  7.572e+00   0.755 0.450434
## TemperatureC:SeasonsSummer           -3.721e+01  9.727e+00  -3.825 0.000147
## TemperatureC:SeasonsWinter           -1.908e+01  8.383e+00  -2.276 0.023237
## Humidity:Solar.Radiation..MJ.m2.      4.844e+00  1.480e+00   3.273 0.001136
## Humidity:Rainfallmm                   1.424e+01  1.150e+01   1.239 0.216009
## Humidity:SeasonsSpring               -1.342e+00  2.888e+00  -0.465 0.642366
## Humidity:SeasonsSummer               -1.544e+00  4.260e+00  -0.362 0.717204
## Humidity:SeasonsWinter                1.472e+00  3.515e+00   0.419 0.675504
## Solar.Radiation..MJ.m2.:Rainfallmm   -3.073e+02  1.926e+02  -1.595 0.111323
## Solar.Radiation..MJ.m2.:SeasonsSpring -1.407e+01  6.741e+01  -0.209 0.834761
## Solar.Radiation..MJ.m2.:SeasonsSummer -2.372e+01  7.886e+01  -0.301 0.763684
## Solar.Radiation..MJ.m2.:SeasonsWinter -5.279e+01  1.154e+02  -0.458 0.647487
## Rainfallmm:SeasonsSpring             -8.520e+01  1.203e+02  -0.708 0.479075
## Rainfallmm:SeasonsSummer              7.428e+01  1.242e+02   0.598 0.550203
## Rainfallmm:SeasonsWinter              8.906e+02  2.551e+03   0.349 0.727093
##
## (Intercept)                          .
## Hour                                 **
## TemperatureC                         *
## Humidity
## Solar.Radiation..MJ.m2.              .
## Rainfallmm
## SeasonsSpring
## SeasonsSummer                         *
## SeasonsWinter
## Hour:TemperatureC                     ***
## Hour:Humidity                         **
## Hour:Solar.Radiation..MJ.m2.
## Hour:Rainfallmm
## Hour:SeasonsSpring
## Hour:SeasonsSummer
## Hour:SeasonsWinter
## TemperatureC:Humidity                 .
## TemperatureC:Solar.Radiation..MJ.m2.  *
## TemperatureC:Rainfallmm
## TemperatureC:SeasonsSpring
## TemperatureC:SeasonsSummer            ***
## TemperatureC:SeasonsWinter            *
## Humidity:Solar.Radiation..MJ.m2.      **
## Humidity:Rainfallmm
## Humidity:SeasonsSpring
## Humidity:SeasonsSummer
## Humidity:SeasonsWinter
## Solar.Radiation..MJ.m2.:Rainfallmm
## Solar.Radiation..MJ.m2.:SeasonsSpring
## Solar.Radiation..MJ.m2.:SeasonsSummer
## Solar.Radiation..MJ.m2.:SeasonsWinter
## Rainfallmm:SeasonsSpring
## Rainfallmm:SeasonsSummer
## Rainfallmm:SeasonsWinter
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 349 on 510 degrees of freedom
## Multiple R-squared:  0.6799, Adjusted R-squared:  0.6592
## F-statistic: 32.83 on 33 and 510 DF,  p-value: < 2.2e-16
```

```
test_model = lm(Rented.Bike.Count ~ . + Hour * TemperatureC, newbikes)
summary(test_model)
```

```
##
## Call:
## lm(formula = Rented.Bike.Count ~ . + Hour * TemperatureC, data = newbikes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1063.16  -195.52   -41.75   150.03  1600.60
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1022.7749    84.7441  12.069  < 2e-16 ***
## Hour                       9.2553     3.4100   2.714 0.006859 **
## TemperatureC               4.8216     3.7109   1.299 0.194396
## Humidity                  -8.3772     0.9806  -8.543  < 2e-16 ***
## Solar.Radiation..MJ.m2.  -72.7473    23.9682  -3.035 0.002521 **
## Rainfallmm               -53.0559    12.0269  -4.411 1.24e-05 ***
## SeasonsSpring           -172.9787    46.7991  -3.696 0.000241 ***
## SeasonsSummer           -104.8042    59.7154  -1.755 0.079822 .
## SeasonsWinter           -429.7717    62.2267  -6.907 1.42e-11 ***
## Hour:TemperatureC          1.5819     0.1980   7.991 8.31e-15 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 371.7 on 534 degrees of freedom
## Multiple R-squared:  0.6199, Adjusted R-squared:  0.6135
## F-statistic: 96.76 on 9 and 534 DF,  p-value: < 2.2e-16
```

```
test_model2 = lm(Rented.Bike.Count ~ . + TemperatureC * Humidity, newbikes)
summary(test_model2)
```

```
##
## Call:
## lm(formula = Rented.Bike.Count ~ . + TemperatureC * Humidity,
##     data = newbikes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1099.41  -245.74   -34.77   176.04  1749.58
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              696.62669   92.68814   7.516 2.40e-13 ***
## Hour                      27.54280    2.57738  10.686  < 2e-16 ***
## TemperatureC              39.95215    5.12819   7.791 3.50e-14 ***
## Humidity                  -5.16734    1.32328  -3.905 0.000106 ***
## Solar.Radiation..MJ.m2.  -86.97521   25.47278  -3.414 0.000688 ***
## Rainfallmm               -41.34330   12.61563  -3.277 0.001117 **
## SeasonsSpring           -180.18073   48.68891  -3.701 0.000237 ***
## SeasonsSummer            -43.81814   64.80220  -0.676 0.499217
## SeasonsWinter           -460.54216   64.63663  -7.125 3.39e-12 ***
## TemperatureC:Humidity     -0.35610    0.08381  -4.249 2.53e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 386.8 on 534 degrees of freedom
## Multiple R-squared:  0.5884, Adjusted R-squared:  0.5814
## F-statistic:  84.8 on 9 and 534 DF,  p-value: < 2.2e-16
```

```
test_model3 = lm(Rented.Bike.Count ~ . + TemperatureC * Humidity +
                   Hour * TemperatureC, newbikes)
summary(test_model3)
```

```
##
## Call:
## lm(formula = Rented.Bike.Count ~ . + TemperatureC * Humidity +
##     Hour * TemperatureC, data = newbikes)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -1057.6  -187.1   -41.6   151.1  1607.4
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              911.24125   93.33528   9.763  < 2e-16 ***
## Hour                      10.39633    3.41375   3.045 0.002438 **
## TemperatureC              17.27258    5.81145   2.972 0.003091 **
## Humidity                  -6.11757    1.27048  -4.815 1.92e-06 ***
## Solar.Radiation..MJ.m2.  -86.43477   24.32595  -3.553 0.000414 ***
## Rainfallmm               -48.12671   12.08389  -3.983 7.76e-05 ***
## SeasonsSpring           -171.31882   46.51271  -3.683 0.000254 ***
## SeasonsSummer            -55.94422   61.90701  -0.904 0.366573
## SeasonsWinter           -432.41390   61.84814  -6.992 8.17e-12 ***
## TemperatureC:Humidity     -0.22727    0.08198  -2.772 0.005763 **
## Hour:TemperatureC          1.46076    0.20152   7.249 1.49e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 369.3 on 533 degrees of freedom
## Multiple R-squared:  0.6253, Adjusted R-squared:  0.6183
## F-statistic: 88.94 on 10 and 533 DF,  p-value: < 2.2e-16
```

We can see that from the model with all the second order interaction variables only a few of the variables are very significant. From this model I chose to see if adding these interactions variables are helpful to the second additive model I already had. For the two test models, I decided to try and adding two interaction terms: TemperatureC * Humidity and TemperatureC * Hour. However, the third test model shows that the adjusted R-squared value is greater than both of them, so I decided to use the test_model3.

Model Selection

```
anova(lm(Rented.Bike.Count ~ 1, bikes), secondmodel)
```

```
## Analysis of Variance Table
##
## Model 1: Rented.Bike.Count ~ 1
## Model 2: Rented.Bike.Count ~ Hour + TemperatureC + Humidity + Solar.Radiation..MJ.m2. +
##     Rainfallmm + Seasons
##   Res.Df       RSS Df Sum of Sq     F    Pr(>F)
## 1    543 194049073
## 2    535  82580031  8 111469042 90.27 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
# # one interaction term added:
anova(secondmodel, test_model2)
```

```
## Analysis of Variance Table
##
## Model 1: Rented.Bike.Count ~ Hour + TemperatureC + Humidity + Solar.Radiation..MJ.m2. +
##     Rainfallmm + Seasons
## Model 2: Rented.Bike.Count ~ Hour + TemperatureC + Humidity + Solar.Radiation..MJ.m2. +
##     Rainfallmm + Seasons + TemperatureC * Humidity
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1    535 82580031
## 2    534 79879251  1   2700780 18.055 2.533e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
# # two interaction terms added:
anova(secondmodel, test_model3)
```

```
## Analysis of Variance Table
##
## Model 1: Rented.Bike.Count ~ Hour + TemperatureC + Humidity + Solar.Radiation..MJ.m2. +
##     Rainfallmm + Seasons
## Model 2: Rented.Bike.Count ~ Hour + TemperatureC + Humidity + Solar.Radiation..MJ.m2. +
##     Rainfallmm + Seasons + TemperatureC * Humidity + Hour * TemperatureC
##   Res.Df      RSS Df Sum of Sq     F    Pr(>F)
## 1    535 82580031
## 2    533 72711459  2   9868572 36.17 1.862e-15 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The first anova test is to check if we want to use the second model over the null model. From the second model, we can see that I would want to use test_model2 over the secondmodel. However, comparing the last two anova tests, we would want to use the model with both interaction terms.

AIC as metric

```
aic_backwards = step(test_model3, direction = 'backward')
```

```
## Start:  AIC=6442.86
## Rented.Bike.Count ~ Hour + TemperatureC + Humidity + Solar.Radiation..MJ.m2. +
##     Rainfallmm + Seasons + TemperatureC * Humidity + Hour * TemperatureC
##
##                           Df Sum of Sq      RSS    AIC
## <none>                                 72711459 6442.9
## - TemperatureC:Humidity    1   1048384 73759844 6448.7
## - Solar.Radiation..MJ.m2.  1   1722317 74433776 6453.6
## - Rainfallmm               1   2163887 74875346 6456.8
## - Seasons                  3   7213791 79925251 6488.3
## - Hour:TemperatureC        1   7167791 79879251 6492.0
```

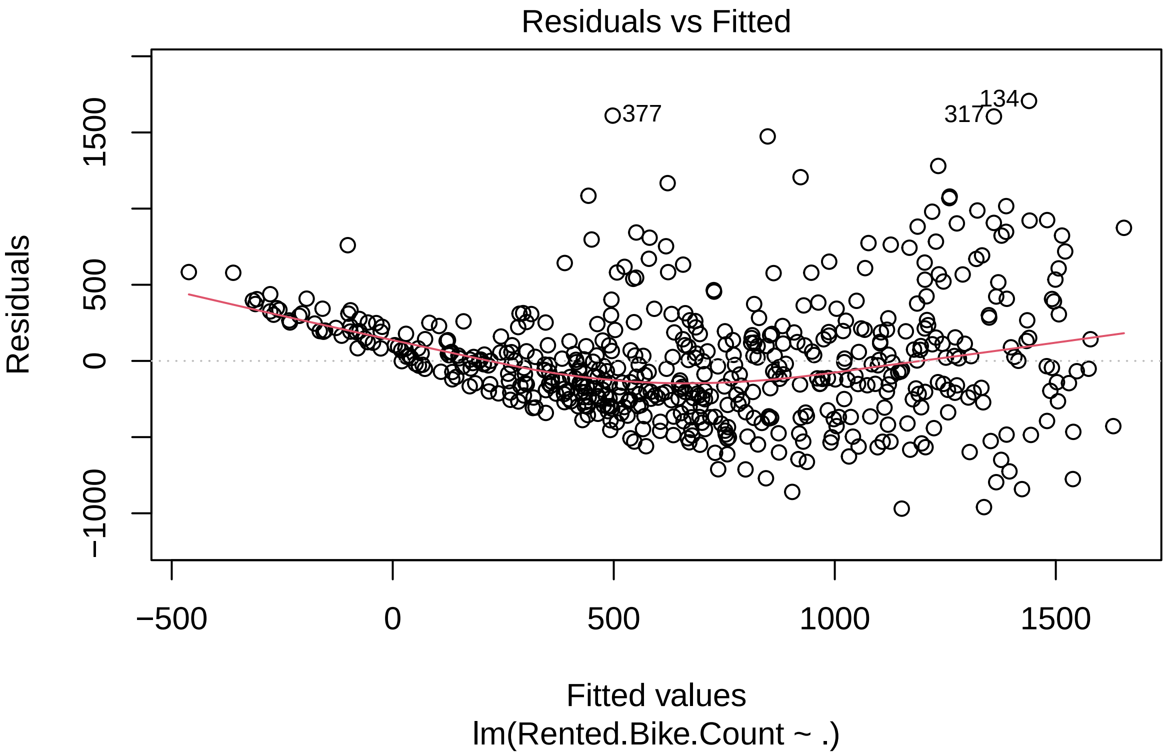
```
finalmodel = lm(Rented.Bike.Count ~ Hour + TemperatureC + Humidity +
                  Solar.Radiation..MJ.m2. + Rainfallmm + Seasons +
                  TemperatureC * Humidity + Hour * TemperatureC, newbikes)
summary(finalmodel)
```

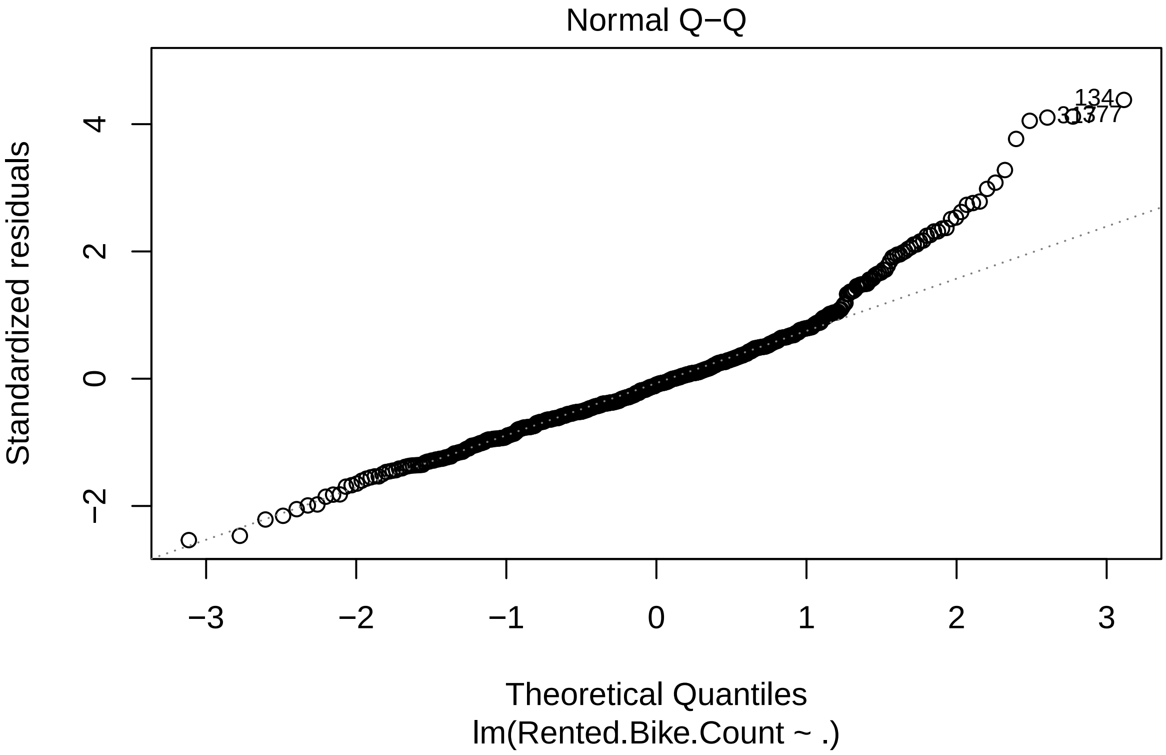
```
##
## Call:
## lm(formula = Rented.Bike.Count ~ Hour + TemperatureC + Humidity +
##     Solar.Radiation..MJ.m2. + Rainfallmm + Seasons + TemperatureC *
##     Humidity + Hour * TemperatureC, data = newbikes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1057.6  -187.1   -41.6   151.1  1607.4
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             911.24125   93.33528   9.763  < 2e-16 ***
## Hour                     10.39633    3.41375   3.045 0.002438 **
## TemperatureC             17.27258    5.81145   2.972 0.003091 **
## Humidity                 -6.11757    1.27048  -4.815 1.92e-06 ***
## Solar.Radiation..MJ.m2.  -86.43477   24.32595  -3.553 0.000414 ***
## Rainfallmm              -48.12671   12.08389  -3.983 7.76e-05 ***
## SeasonsSpring          -171.31882   46.51271  -3.683 0.000254 ***
## SeasonsSummer           -55.94422   61.90701  -0.904 0.366573
## SeasonsWinter          -432.41390   61.84814  -6.992 8.17e-12 ***
## TemperatureC:Humidity    -0.22727    0.08198  -2.772 0.005763 **
## Hour:TemperatureC         1.46076    0.20152   7.249 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 369.3 on 533 degrees of freedom
## Multiple R-squared:  0.6253, Adjusted R-squared:  0.6183
## F-statistic: 88.94 on 10 and 533 DF,  p-value: < 2.2e-16
```

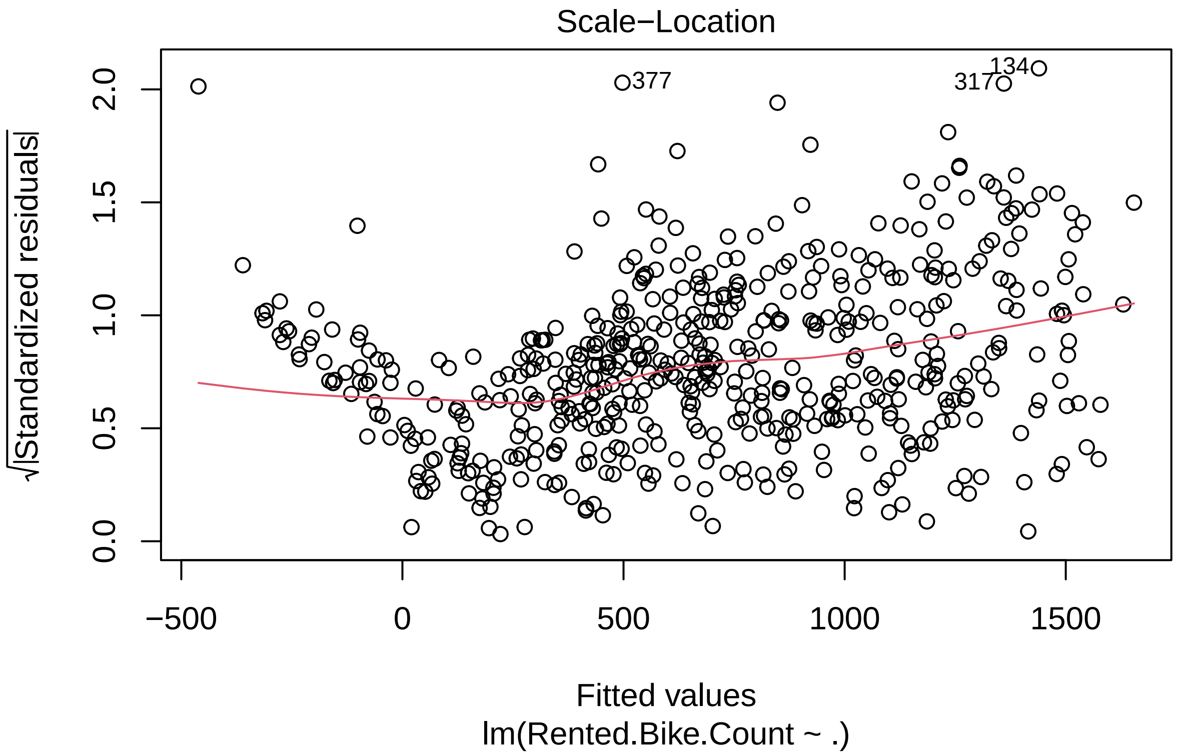
I also performed another model selection method to double check. I used the AIC as metric and backward direction. I started with test_model3, which is the model with both interaction terms. With the AIC value of 6442.86, I decided that the test_model3 was useful. Then I assigned the model to the variable 'finalmodel' before I realized I needed to transform the model. I can also see the result of the summary of the model where the adjusted R-squared value is 0.6183 and the p-value of 2.2e-16.

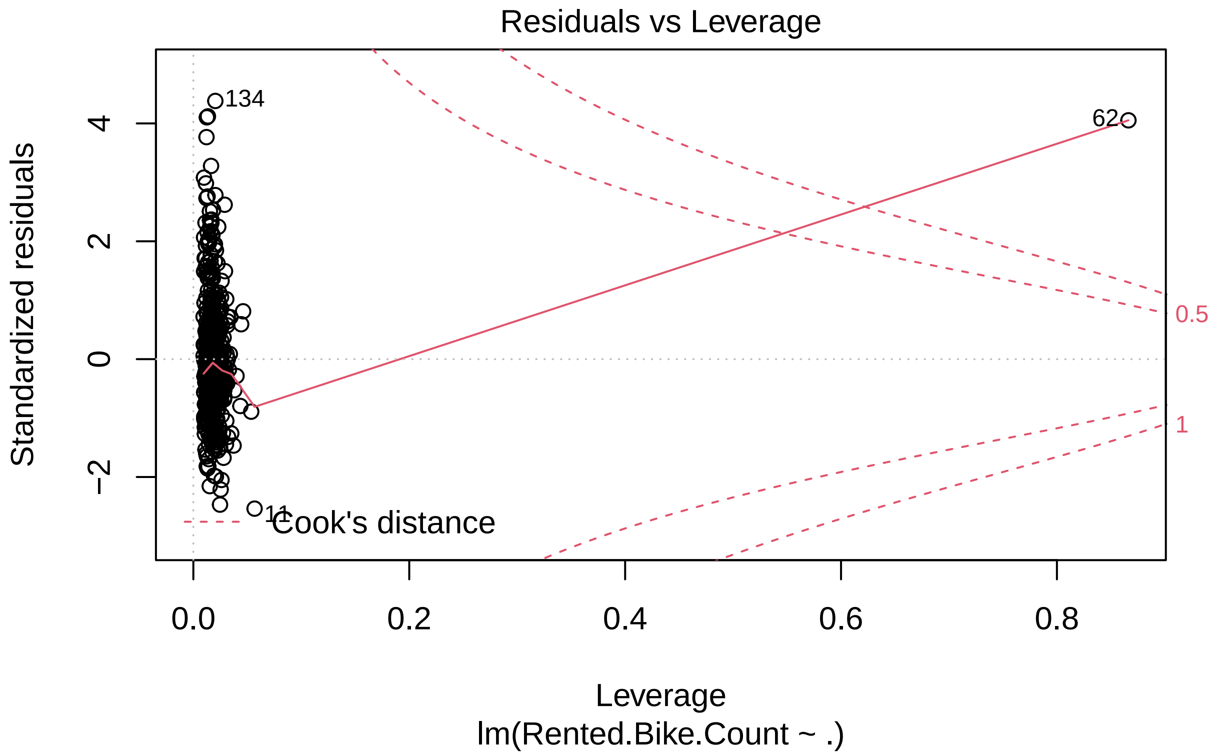
Unusual Points

```
plot(firstmodel)
```

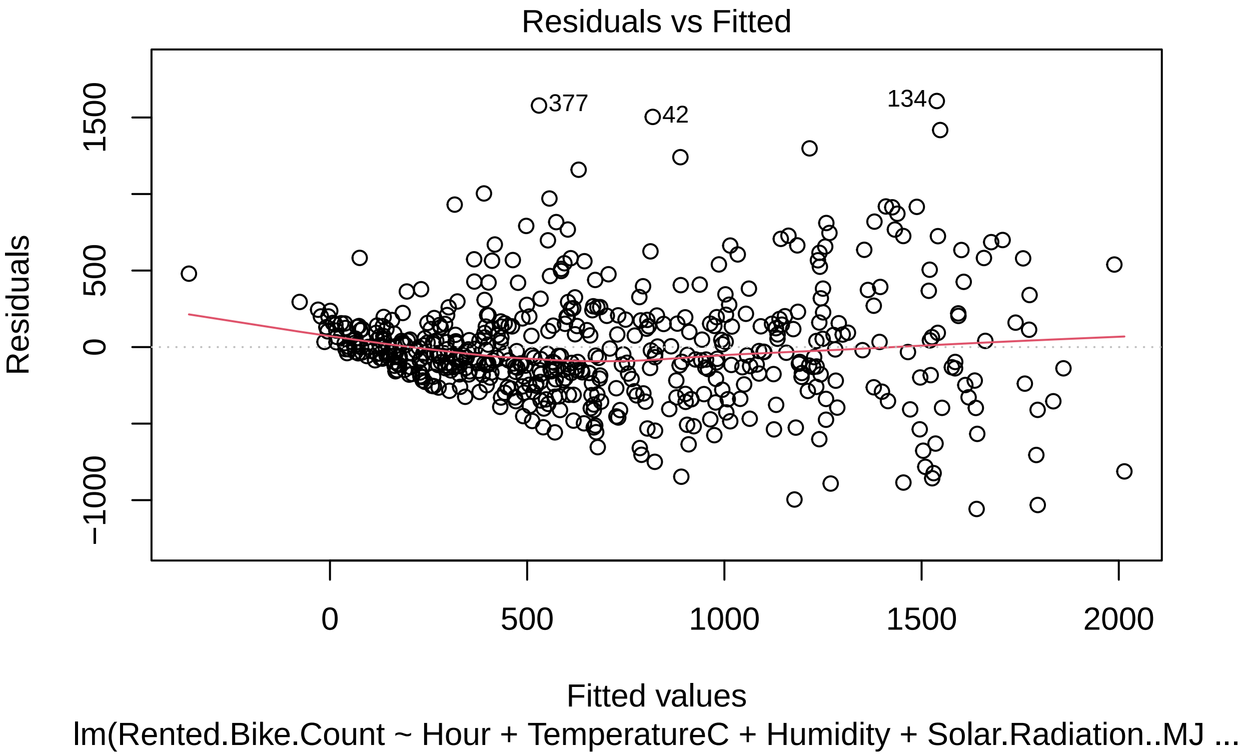

Residuals vs Fitted
Residuals
Fitted values
lm(Rented.Bike.Count ~ .)
−1000
0
500
1500
−500
0
500
1000
1500
377
317
134

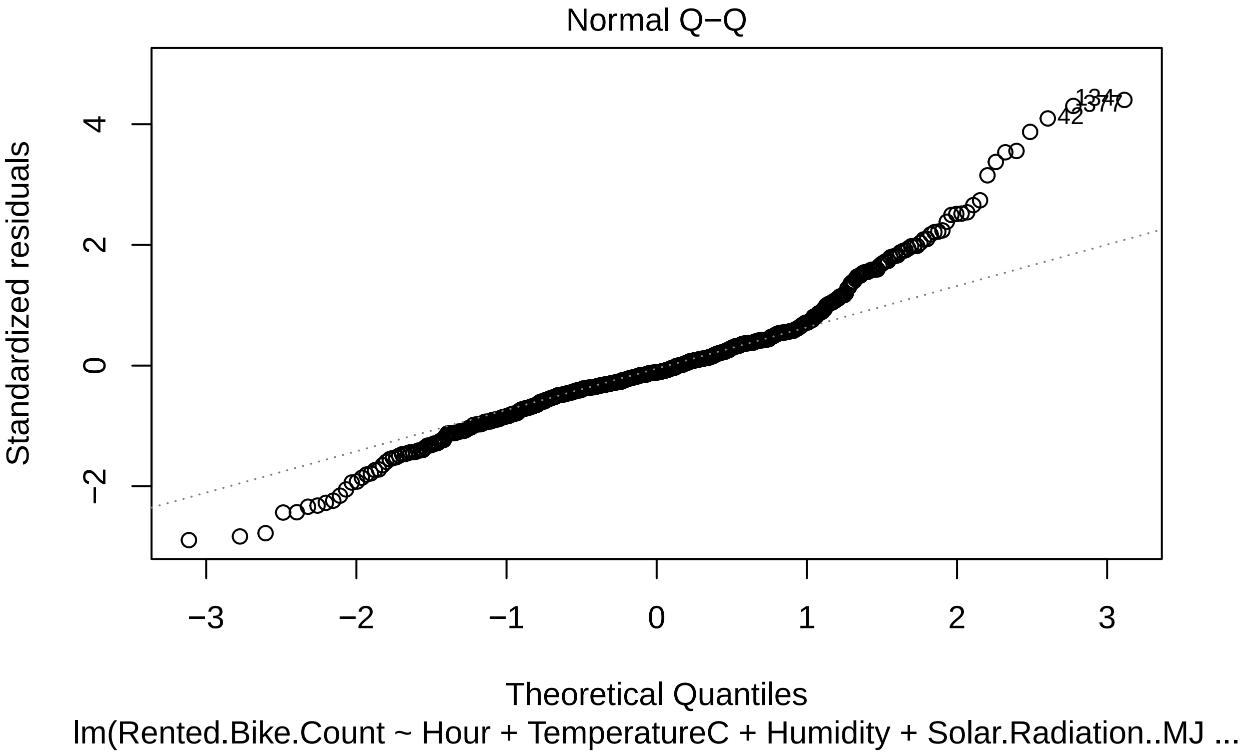

Normal Q−Q
Standardized residuals
−2
0
2
4
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Rented.Bike.Count ~ .)
134
313
77

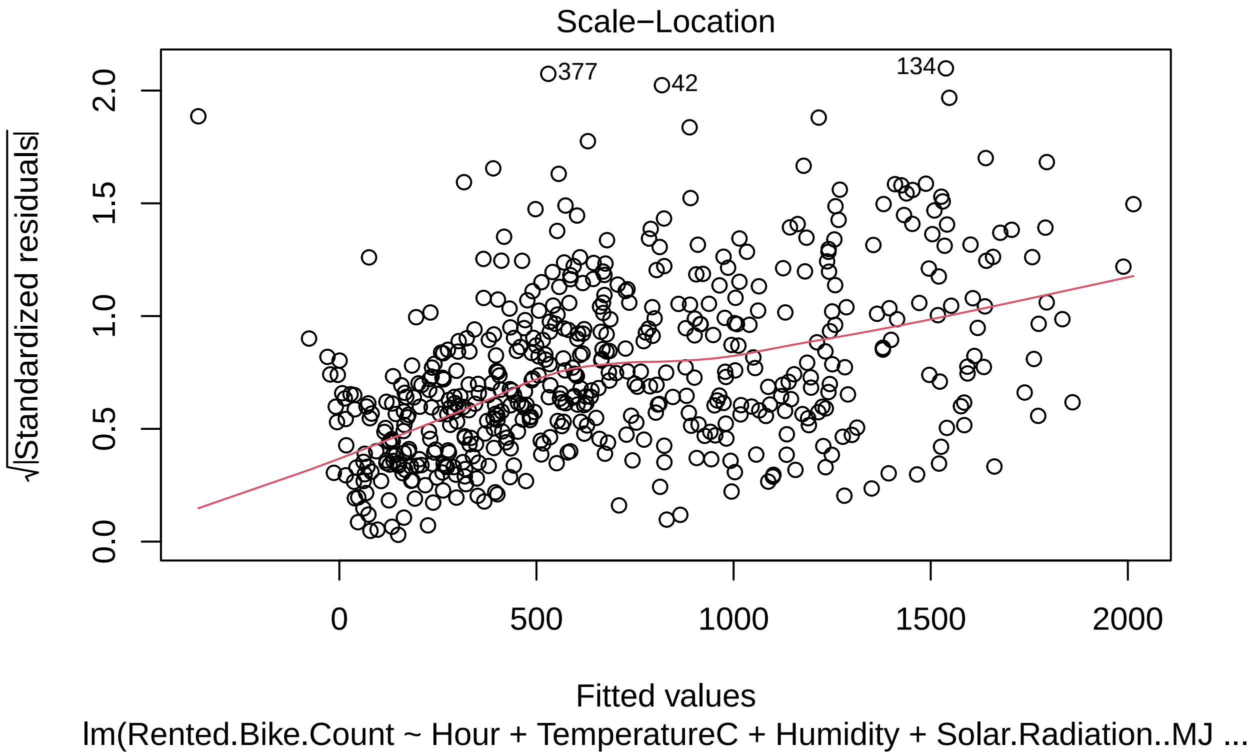

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
−500
0
500
1000
1500
377
134
317
Fitted values
lm(Rented.Bike.Count ~ .)

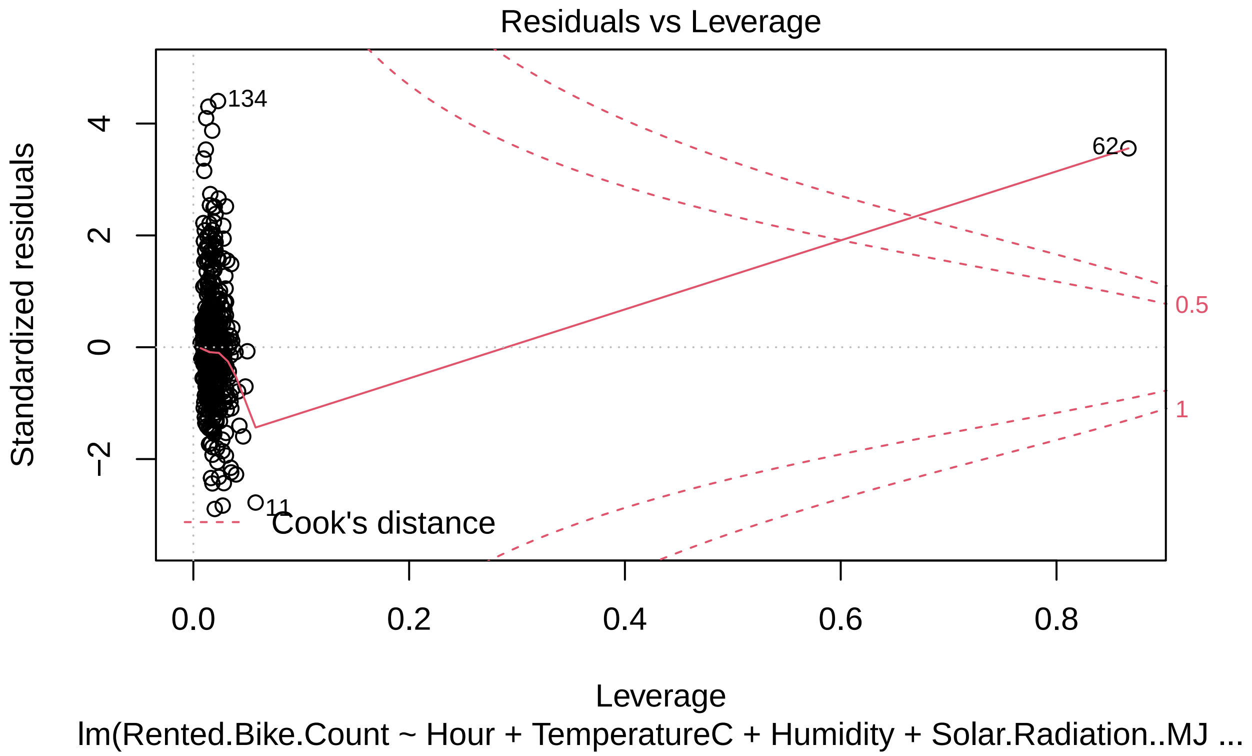


```
plot(finalmodel)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Rented.Bike.Count ~ Hour + TemperatureC + Humidity + Solar.Radiation..MJ ...
1500
500
0
−1000
0
500
1000
1500
2000
377
42
134


Normal Q−Q
Standardized residuals
4
2
0
−2
134
42
377
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Rented.Bike.Count ~ Hour + TemperatureC + Humidity + Solar.Radiation..MJ ...

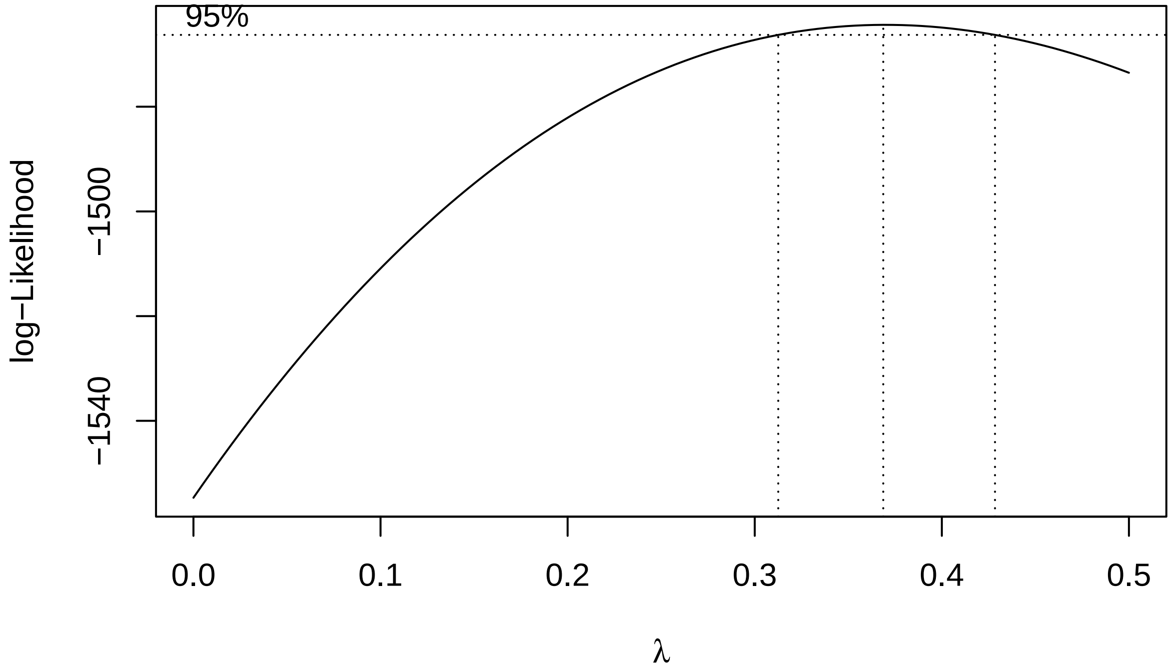

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
500
1000
1500
2000
377
42
134
Fitted values
lm(Rented.Bike.Count ~ Hour + TemperatureC + Humidity + Solar.Radiation..MJ ...



Here we can see the four default plots of the first model and the finalmodel and see how the all the graphs are much better for the final model than the first model. However, I can see that the Residuals vs. Fitted plot is not necessarily randomly scattered, Normal Q-Q plot seems bit off at the end of the tail, and the Scale-Location graph shows how the red line was not flat enough.

Boxcox

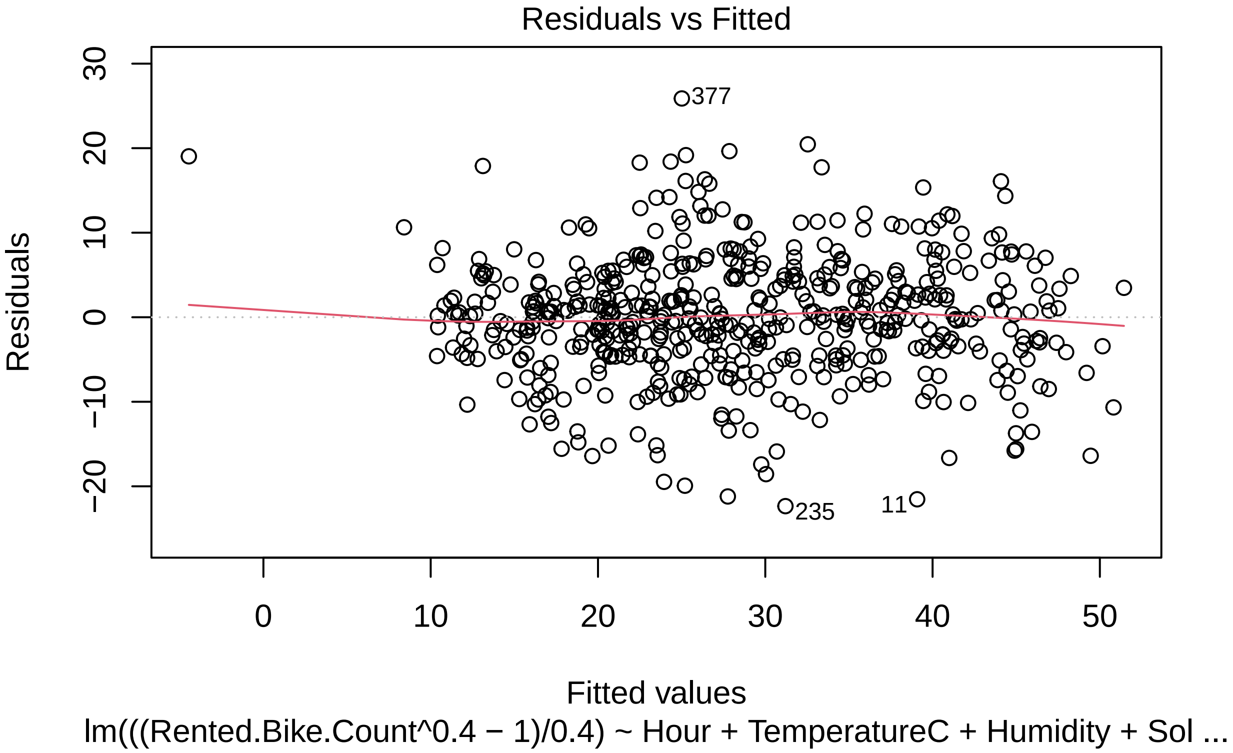
```
boxcox(finalmodel, plotit = TRUE, lambda = seq(0, 0.5, 0.1))
```

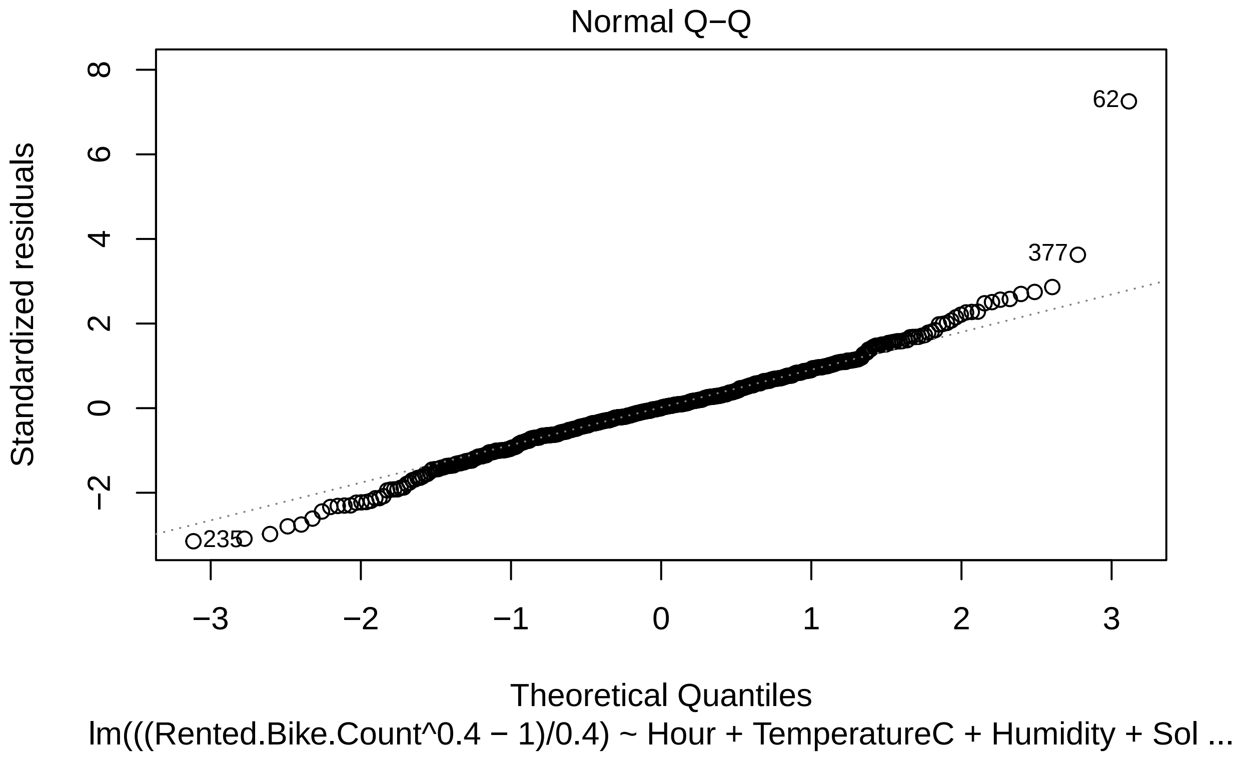


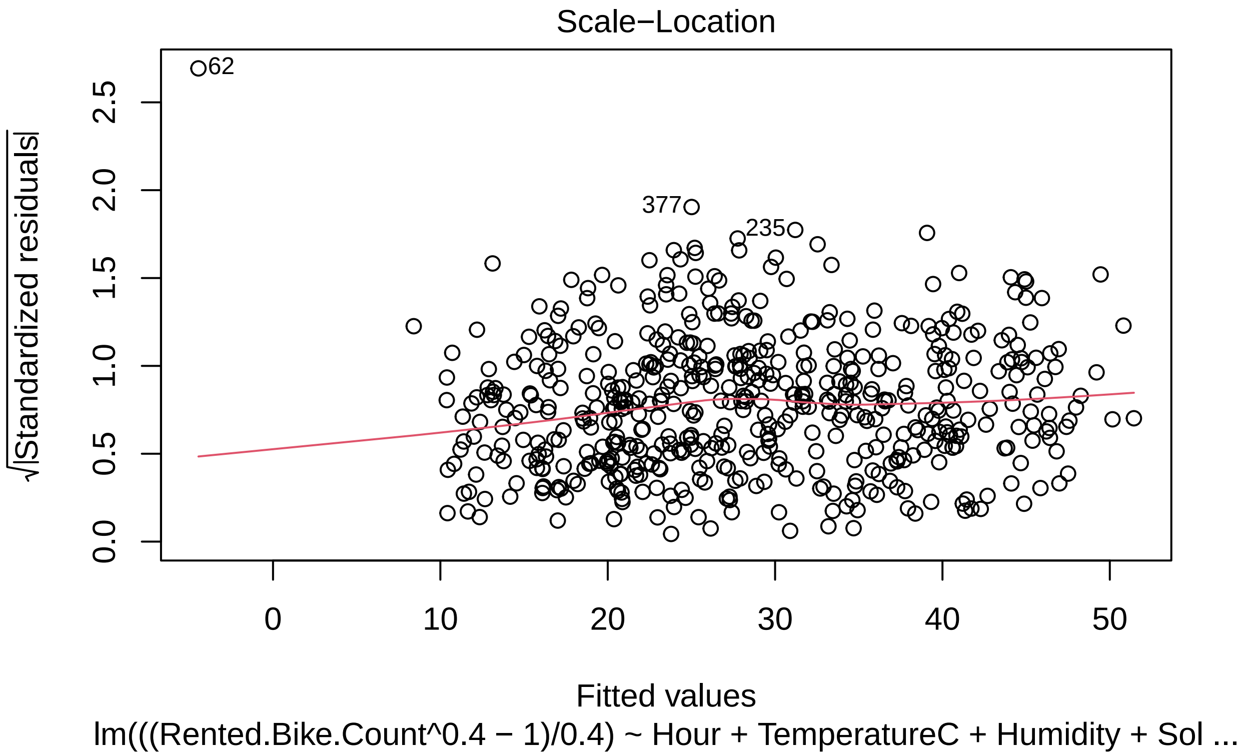
```
transformed_mod = lm(((Rented.Bike.Count ^0.4 - 1)/0.4) ~ Hour + TemperatureC + Humidity + Solar.Radiati
summary(transformed_mod)
```

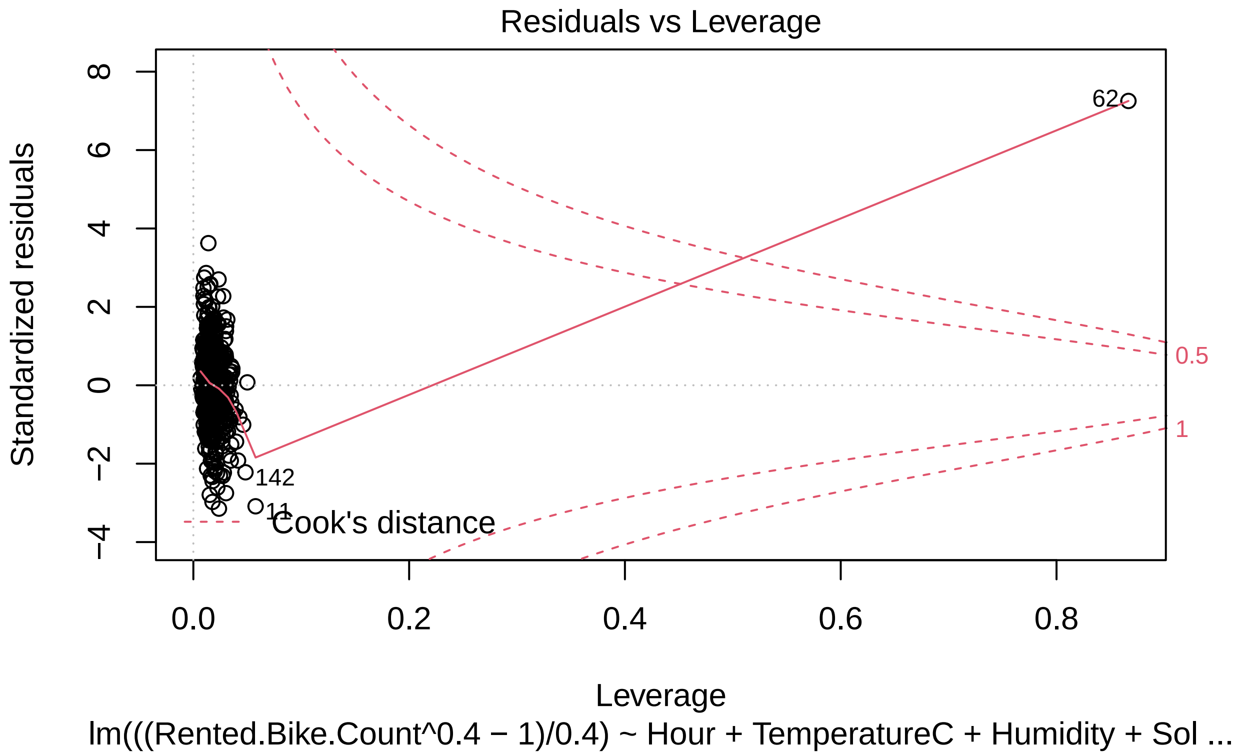
```
##
## Call:
## lm(formula = ((Rented.Bike.Count^0.4 - 1)/0.4) ~ Hour + TemperatureC +
##     Humidity + Solar.Radiation..MJ.m2. + Rainfallmm + Seasons +
##     TemperatureC * Humidity + Hour * TemperatureC, data = newbikes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.3468  -4.1299   0.0141   4.3931  25.8809
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             33.492523   1.816418  18.439  < 2e-16 ***
## Hour                     0.384364   0.066436   5.786 1.23e-08 ***
## TemperatureC             0.408831   0.113098   3.615 0.000329 ***
## Humidity                -0.158969   0.024725  -6.429 2.85e-10 ***
## Solar.Radiation..MJ.m2. -0.937375   0.473412  -1.980 0.048213 *
## Rainfallmm              -1.308152   0.235167  -5.563 4.21e-08 ***
## SeasonsSpring           -3.795978   0.905194  -4.194 3.22e-05 ***
## SeasonsSummer           -1.177376   1.204785  -0.977 0.328889
## SeasonsWinter          -11.082665   1.203640  -9.208  < 2e-16 ***
## TemperatureC:Humidity   -0.002863   0.001595  -1.794 0.073326 .
## Hour:TemperatureC         0.013284   0.003922   3.387 0.000758 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.188 on 533 degrees of freedom
## Multiple R-squared:  0.6588, Adjusted R-squared:  0.6524
## F-statistic: 102.9 on 10 and 533 DF,  p-value: < 2.2e-16
```

```
plot(transformed_mod)
```




Normal Q−Q
Standardized residuals
8
6
4
2
0
−2
62
377
235
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(((Rented.Bike.Count^0.4 − 1)/0.4) ~ Hour + TemperatureC + Humidity + Sol ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
0
10
20
30
40
50
62
377
235
Fitted values
lm(((Rented.Bike.Count^0.4 − 1)/0.4) ~ Hour + TemperatureC + Humidity + Sol ...



```
finalmodel = transformed_mod
```

Here I performed the Box-Cox to see if any transformation was necessary for the final model. The lambda value seems to be close to 0.4 so I decided to transform my Rented.Bike.Count, or my response variable, accordingly with lambda = 0.4. The following transformed_mod shows the transformed linear model. The summary of the transformed model shows how most of the variables are pretty significant and a bit higher adjusted R-squared. The most significant difference shows from the four default plots. The Residuals vs. Fitted shows much more randomly scattered values with the average being close to the 0 line, the Normal Q-Q plot shows how the tail is much closer to the line, and the Scale-Location graph shows that the red line is much flatter than before. I decided that this transformed model will be my final model, so I assigned it to the variable "finalmodel".

Collinearity

```
# here we want to omit categorical variable
firstmodel_nocat = lm(Rented.Bike.Count ~ . - Seasons, bikes)
finalmodel_nocat = lm(((Rented.Bike.Count ^0.4 - 1)/0.4) ~ Hour + TemperatureC +
                   Humidity + Solar.Radiation..MJ.m2. + Rainfallmm +
                   TemperatureC * Humidity + Hour * TemperatureC, newbikes)

vif(firstmodel_nocat)
```

```
##                     Month          TemperatureC                     Hour
##                1.084704                1.358481                1.120060
##                Humidity           Wind.speed.ms              Rainfallmm
##                1.695711                1.328152                1.057349
## Solar.Radiation..MJ.m2.
##                1.703043
```

```
vif(finalmodel_nocat)
```

```
##                    Hour            TemperatureC                Humidity
##                2.113449               16.797243                2.676192
## Solar.Radiation..MJ.m2.              Rainfallmm   TemperatureC:Humidity
##                1.692868                1.080575               13.574262
##       Hour:TemperatureC
##                5.347797
```

The collinearity for the first model seems fine, but the collinearity in the finalmodel model seems to be concerning for the variable 'TemperatureC' and 'TemperatureC:Humidity'.

R-squared

```
summary(firstmodel)$adj.r.squared
```

```
## [1] 0.5665355
```

```
summary(finalmodel)$adj.r.squared
```

```
## [1] 0.6523626
```

##RMSE

```
sqrt(mean(resid(firstmodel) ^ 2))
```

```
## [1] 389.5801
```

```
sqrt(mean(resid(finalmodel) ^ 2))
```

```
## [1] 7.11495
```

Try fitting a new point

I wanted to see if the model was somewhat reasonable so I made up a scenario, assigning each variable a certain number. With my experience, I know Summer gets very, (extremely) hot and humid in Seoul so I wanted to try a scenario in the Summer. My scenario was Hour = 15, TemperatureC = 30, Humidity = 90, Solar.Radiation..MJ.m2. = 1.8, Rainfallmm = 0, Seasons = "Summer".

```
transformed_mod$coefficients
```

```
##             (Intercept)                    Hour            TemperatureC 
##            33.492523097             0.384363947             0.408831369 
##                Humidity Solar.Radiation..MJ.m2.              Rainfallmm 
##            -0.158968974            -0.937375254            -1.308151968 
##           SeasonsSpring           SeasonsSummer           SeasonsWinter 
##            -3.795977977            -1.177376054           -11.082664910 
##   TemperatureC:Humidity       Hour:TemperatureC 
##            -0.002862839             0.013283877
```

```
point1 = data.frame(Hour = 15, TemperatureC = 30, Humidity = 90, Solar.Radiation..MJ.m2. = 1.8, Rainfal

y = 33.492523097 + 0.384363947*point1$Hour + 0.408831369*point1$TemperatureC + -0.158968974*point1$Humic
((y*0.4) + 1)**(1/0.4)
```

```
## [1] 738.5693
```

I would predict approximately 739 bikes rented in Seoul.