HW4 - 250620601

Ravin Lathigra

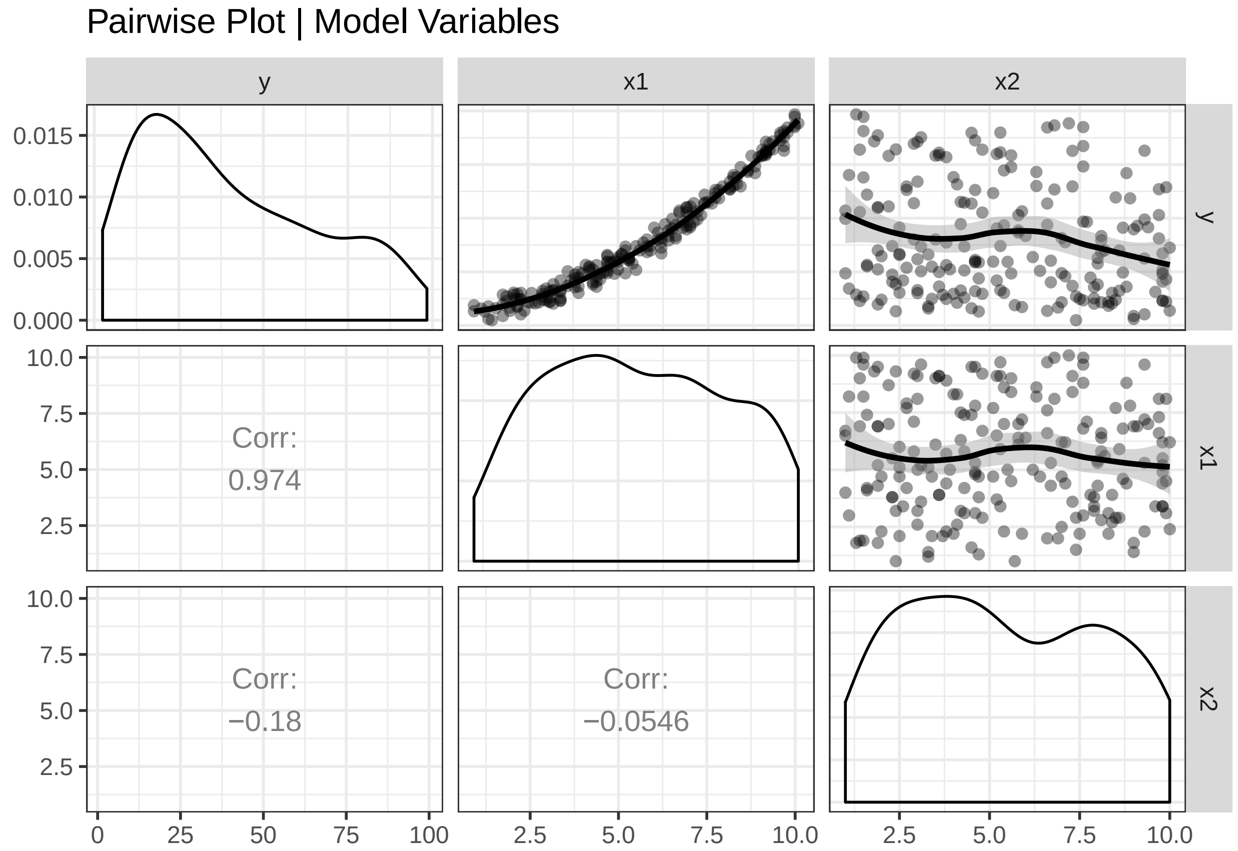
November 7, 2018

R Packages & Libraries

```
library(corrplot)    #Visualize Correlation between variables
library(kableExtra)  #Style tables
library(tidyverse)   #contains ggplot2,dplyr,tidyr, readr,purr,tibble,stringr,forcats
library(formatR)     #Improve readability of code
library(e1071)       #Functions for latent class analysis, Fourier transform ect.
library(VIM)         #Knn
library(ggfortify)   #Add on to ggplot2 to allow for more plot types
library(Rtsne)       #Dimension reduction classification
library(caret)       #streamlined model development
library(RColorBrewer)#Control colours of visualizations
library(GGally)      #Contains ggpairs plots
library(lmtest)      #Test for linear assumptions
library(MASS)
library(faraway)
```

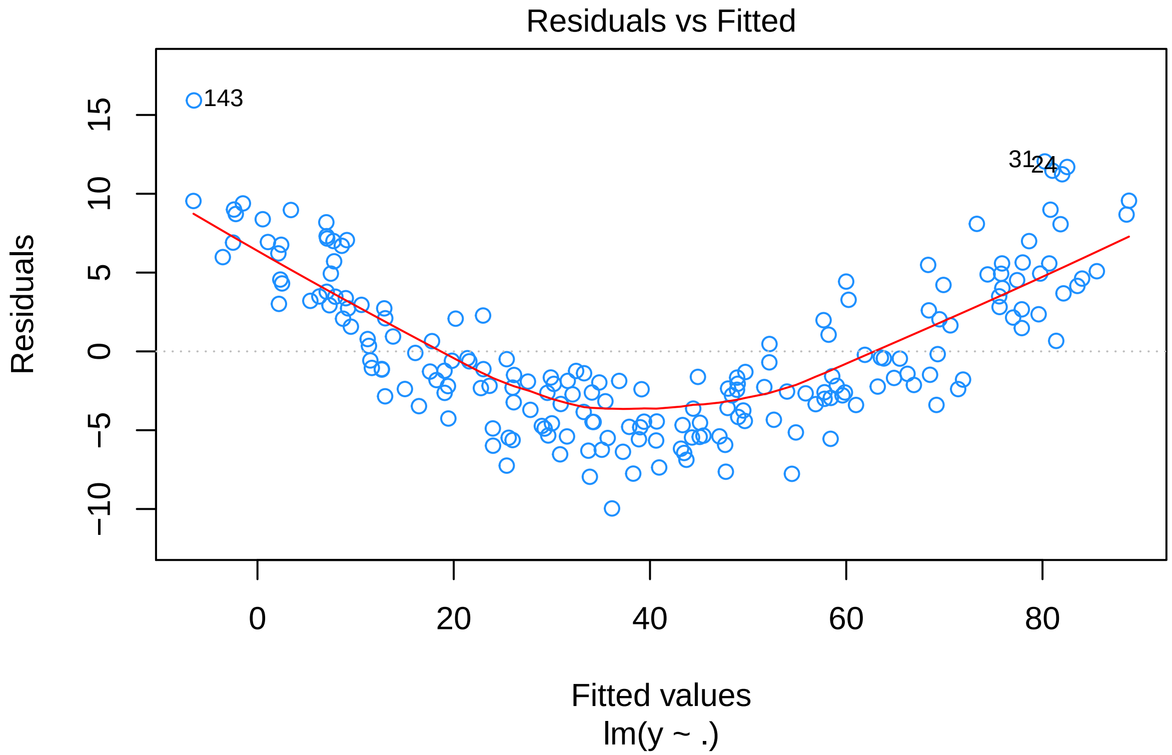
Question 1

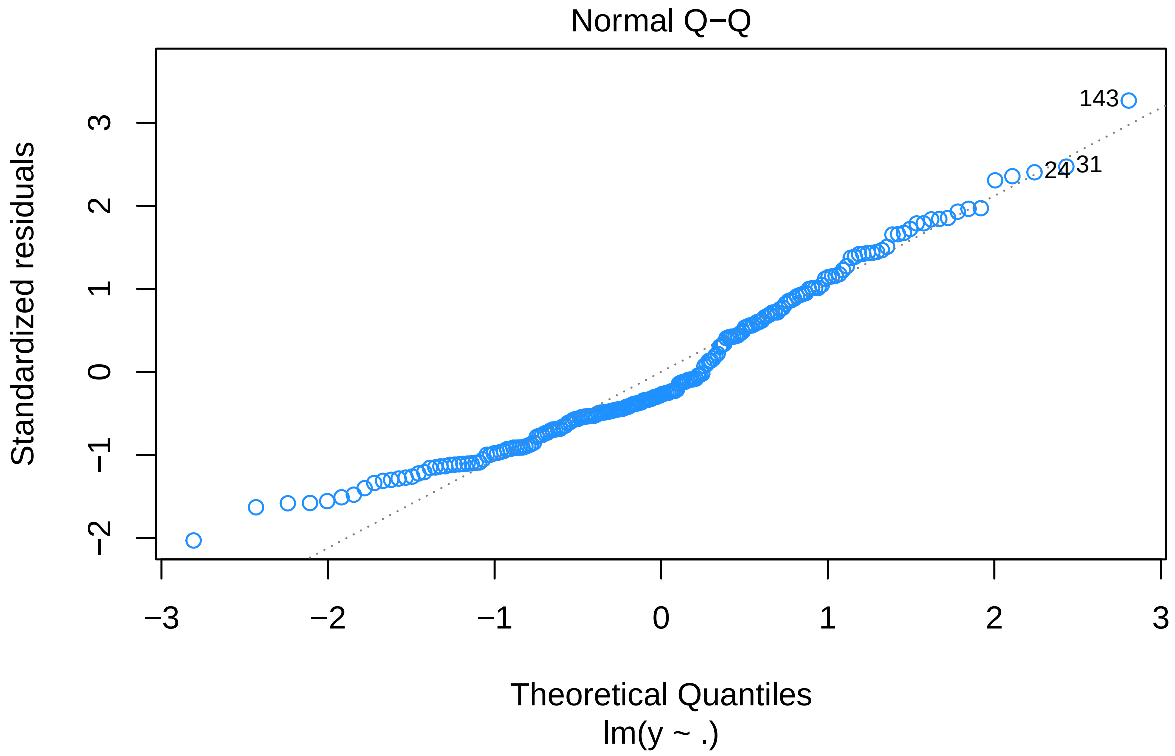
One A - Variable Relationships

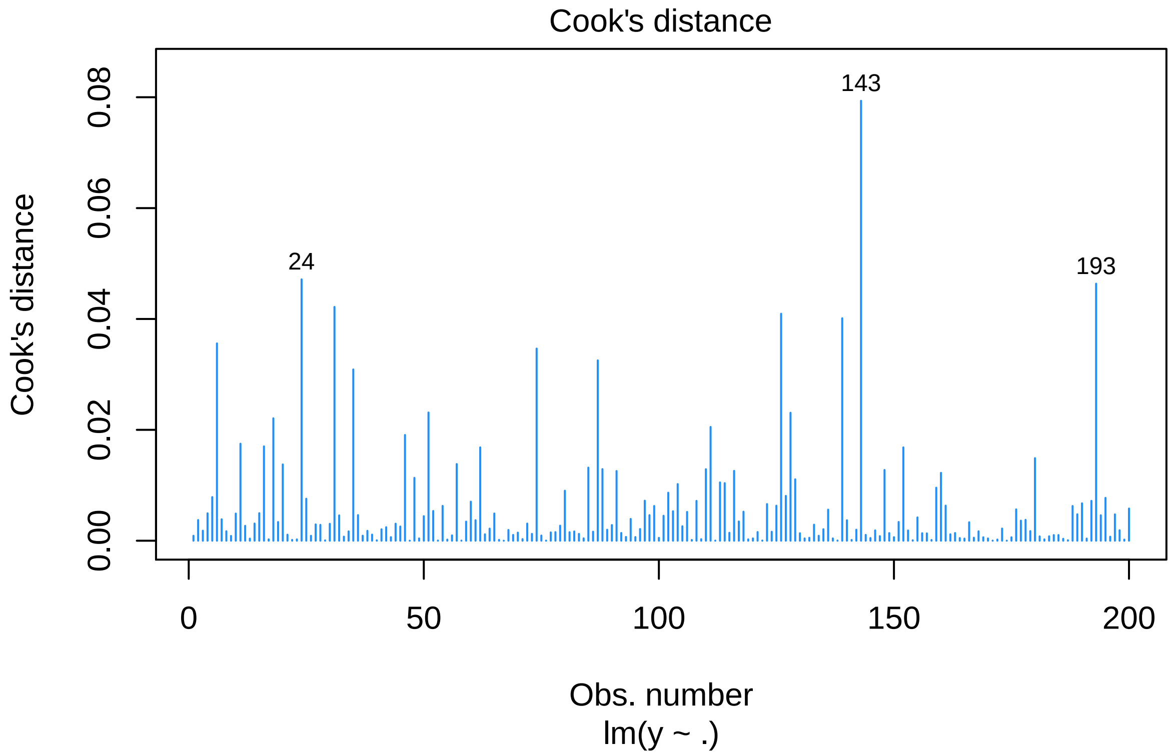


The plot *Pairwise Plot | Model Variables* shows between predictor relationships in 2 dimensions. On the **diagonal** the estimated probability density functions are illustrated. As a default in the `GGplot2 package`, a gaussian kernal is used for the estimation. The **upper** portion of the plot shows the scatterplot with the labels on right side of the plot corresponding to the variable represented on the y axis and the upper label corresponding to the x axis. A loess smoother with a 95% confidence interval is included to model any complex relationships that are difficult to capture with parametric techniques. The ploy `y vs x1` shows that there is a strong positive correlation between the two variables though the curvature in the plot may suggest that a polynomial regression may be more appropriate. The remaining plots show `y vs x2` and `x2 vs x1`. While we observed that `x1` and `y` were highly correlatied we would expect these plots to apprear similar. Both plots show that `x2` does not have an observable relationship with y or x1. This is further supported by the lower portion of the plot which shows the correlation.

One B - Model Assumptions


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ .)
−10
−5
0
5
10
15
0
20
40
60
80
143
31
24


Normal Q−Q
Standardized residuals
−2
−1
0
1
2
3
−3
−2
−1
0
1
2
3
143
24
31
Theoretical Quantiles
lm(y ~ .)



```
##
##  studentized Breusch-Pagan test
##
## data:  lm_1b
## BP = 0.094601, df = 2, p-value = 0.9538
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm_1b)
## W = 0.95915, p-value = 1.603e-05
```

The above plots display the following:

- Resiudals vs Fitted Values
- Normal qq plot
- Cook's Distance

Linear Model Appropriateness:

Linearity - Inspecting the plot "Residuals Vs Fitted" has a trend line that helps illustrate that there is a parabolic relationship between fitted values and residuals. Furthermore, the residuals do not exhibit zero mean suggesting that a linear model may not be the most appropriate model and perhaps transformations should be considered.

Equal Variance - Inspecting the plot "Residuals Vs Fitted" we see that at any subset of the fitted values, there is a constant variance.

Normality assumption - Inspecting the plot "Normal Q-Q" we that the standardized residuals moderately correspond to the theoretical quantiles of a normal distribution. To properly assess if the normality

Table 1: Influential Observations

	Index
6	6
18	18
24	24
31	31
35	35
51	51
74	74
87	87
111	111
126	126
128	128
139	139
143	143
193	193

Table 2: Table 1.0: Influential Observations with large residuals

	Index
24	24
31	31
139	139
143	143
193	193

assumption is violated the Shapiro test will be carried out. Additionally, the plot identifies 3 points that have the largest residuals.

Points of Interest - Also included is a plot of Cook's distance which is a good indicator of point that may have high influence or require further investigation.

BP Test: p-value >> 5% significance level this suggests that the equal variance assumptions holds for this model.

Shapiro Test: p-value < 5% significance level this suggests that the normality assumption doesn't hold for this model.

One C - Influential Points

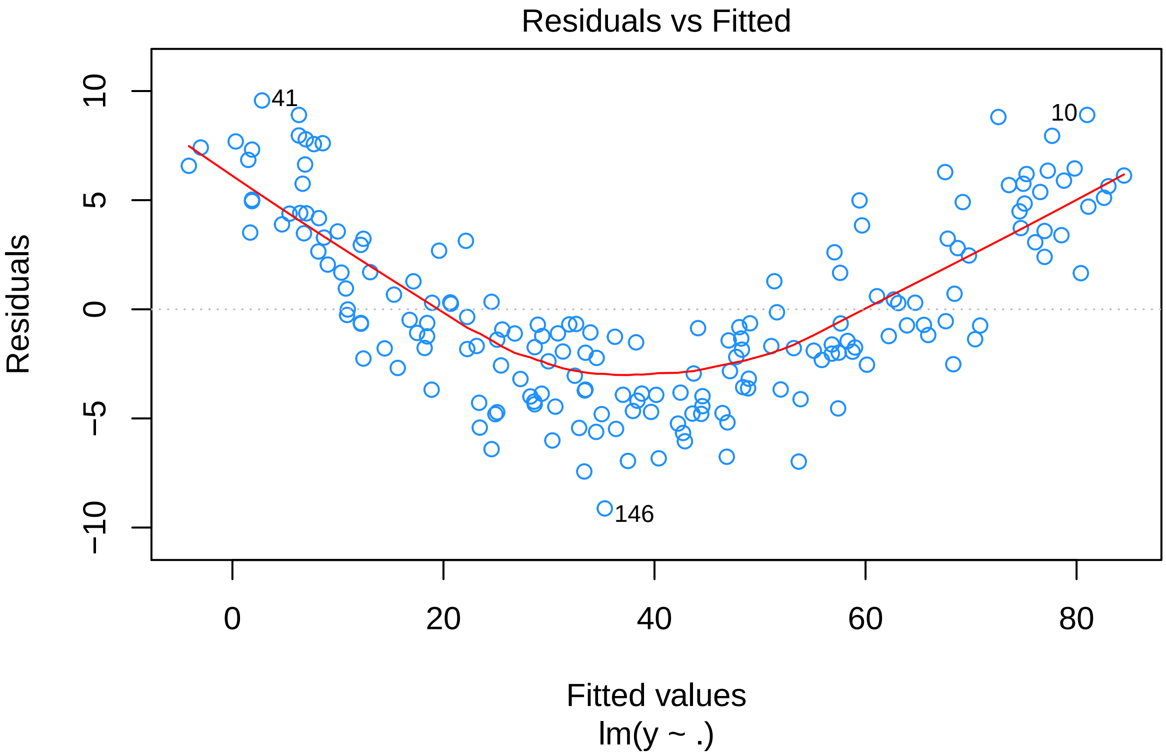
```
## Warning: package 'bindrcpp' was built under R version 3.3.3
```

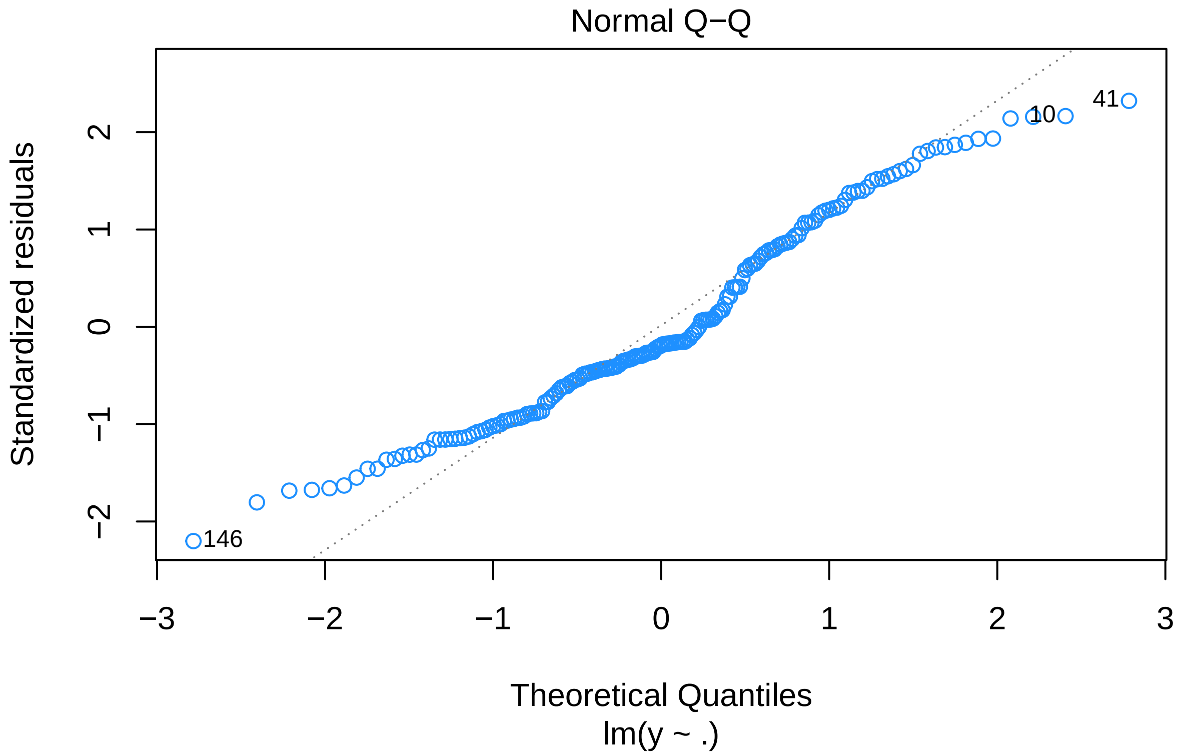
The table **Table 1.0: Influential Observations** shows that there are 14 observations whose cooks distance suggest that they are influential points.

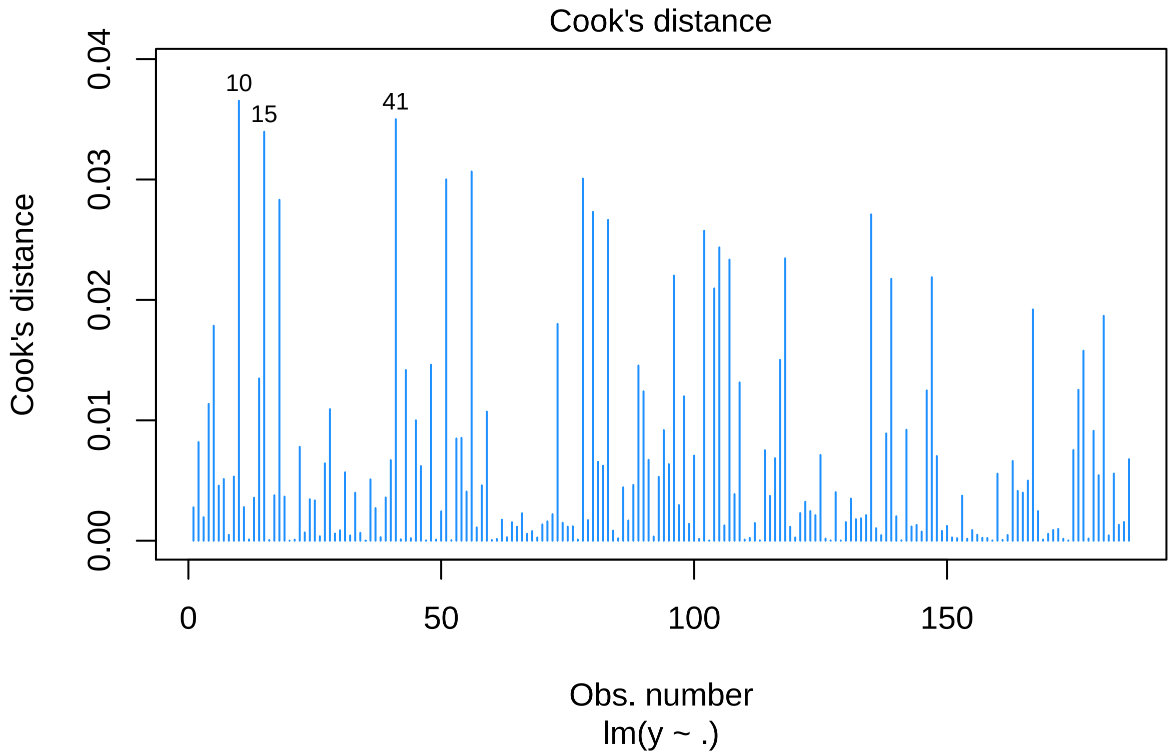
One D - Outliers

The table **Influential Observations with large residuals** shows that there are 5 observations whose cooks distance suggest that they are influential points with large residuals.

One E - Remove Influential Points


Residuals vs Fitted
Residuals
10
5
0
−5
−10
0
20
40
60
80
Fitted values
lm(y ~ .)
41
10
146


Normal Q−Q
Standardized residuals
−2
−1
0
1
2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(y ~ .)
146
10
41



```
##
##  studentized Breusch-Pagan test
##
## data:  lm_1e
## BP = 0.78179, df = 2, p-value = 0.6764
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm_1e)
## W = 0.96638, p-value = 0.0001911
```

The above plots display the following:

- Resiudals vs Fitted Values
- Normal qq plot
- Cook's Distance

Linear Model Appropriateness:

Linearity - Inspecting the plot "Residuals Vs Fitted" has a trend line that helps illustrate that there is a parabolic relationship between fitted values and residuals. Furthermore, the residuals do not exhibit zero mean suggesting that a linear model may not be the most appropriate model and perhaps transformations should be considered.

Equal Variance - Inspecting the plot "Residuals Vs Fitted" we see that at any subset of the fitted values, there is a constant variance.

Normality assumption - Inspecting the plot "Normal Q-Q" we that the standardized residuals moderately correspond to the theoretical quantiles of a normal distribution. To properly assess if the normality assumption is violated the Shapiro test will be carried out. Additionally, the plot identifies 3 points that have the largest residuals.

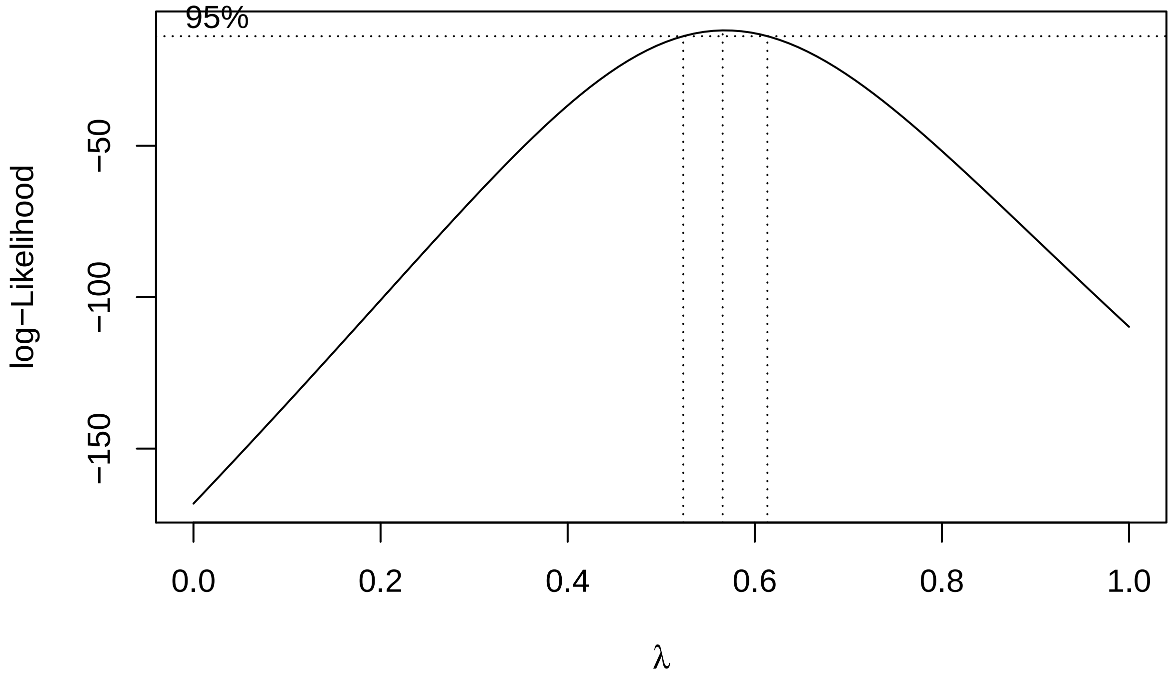
Points of Interest - Also included is a plot of Cook's distance which is a good indicator of point that may have high influence or require further investigation.

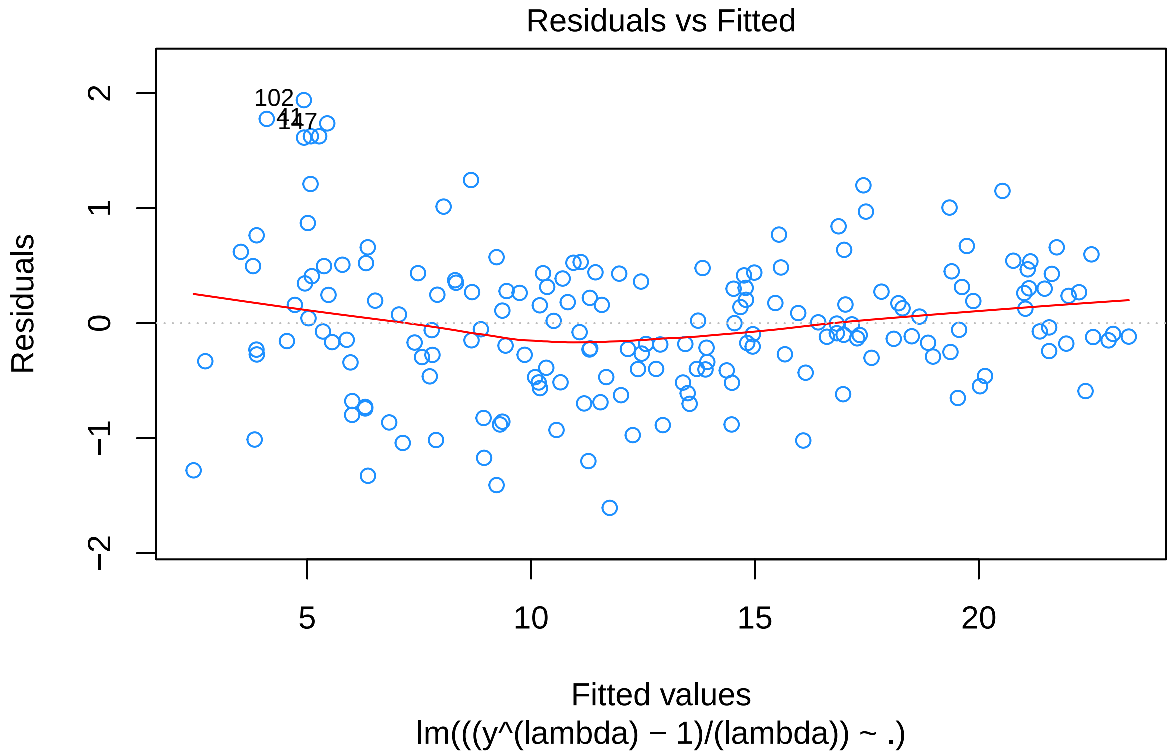
BP Test: p-value >> 5% significance level this suggests that the equal variance assumptions holds for this model.

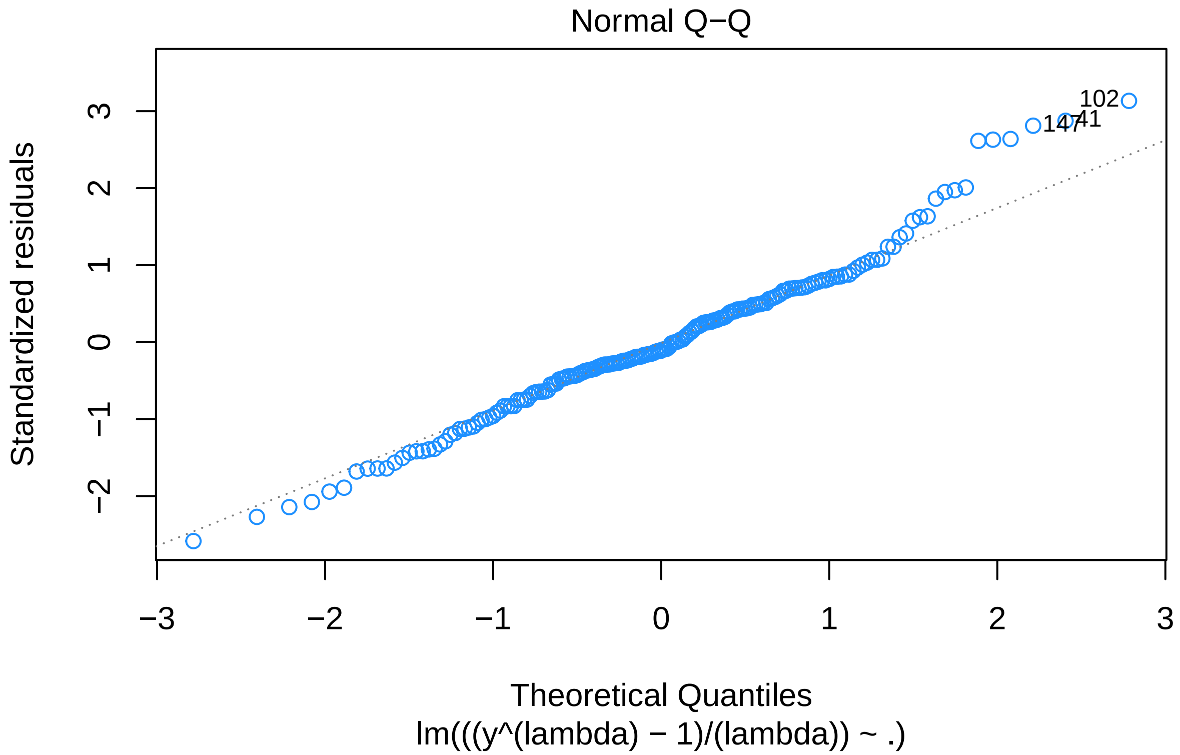
Shapiro Test: p-value < 5% significance level this suggests that the normality assumption doesn't hold for this model.

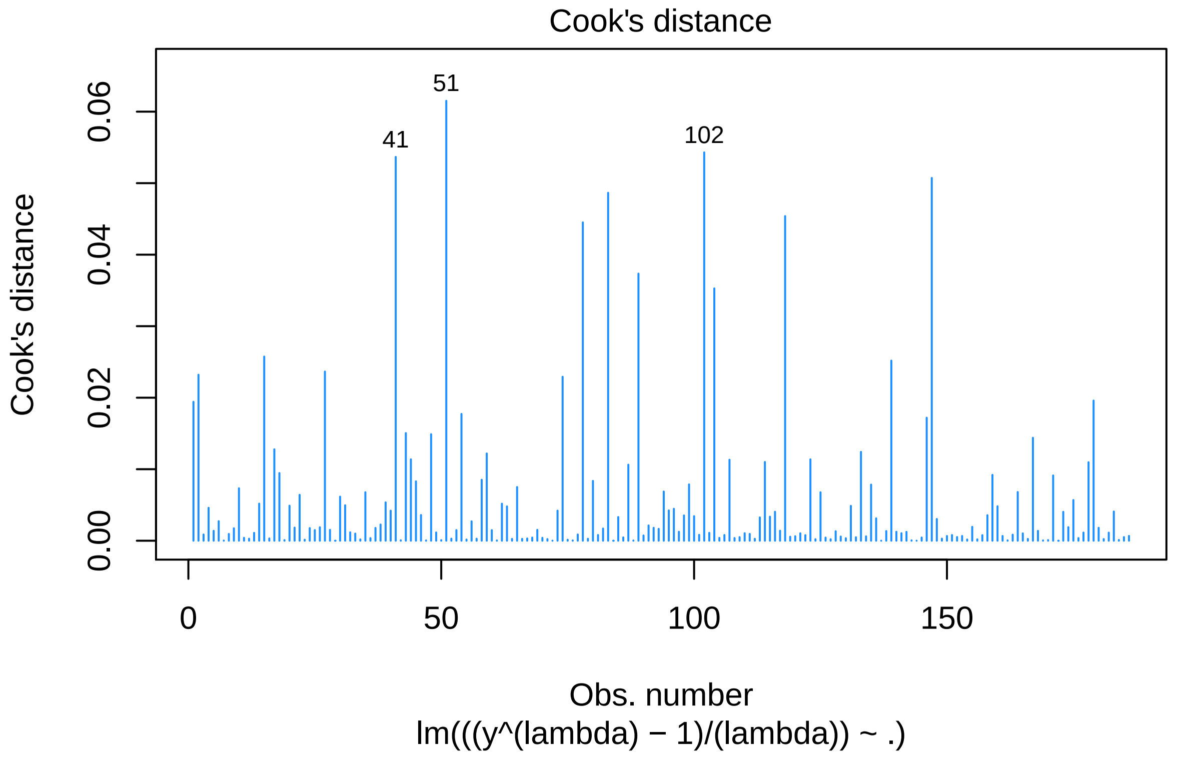
Therefore, removing influential points **did not** correct the model assumptions.

One F - Boxcox Transformation




Residuals vs Fitted
Residuals
2
1
0
−1
−2
102
41
147
5
10
15
20
Fitted values
lm(((y^(lambda) − 1)/(lambda)) ~ .)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(((y^(lambda) − 1)/(lambda)) ~ .)
−3
−2
−1
0
1
2
3
−2
−1
0
1
2
3
102
147
41



```
##
##  studentized Breusch-Pagan test
##
## data:  lm_1f
## BP = 18.947, df = 2, p-value = 7.684e-05
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm_1f)
## W = 0.98033, p-value = 0.01007
```

The above plots display the following:

- Resiudals vs Fitted Values
- Normal qq plot
- Cook's Distance

Linear Model Appropriateness:

Linearity - Using a transformed response variable (lambda = 0.6) removes the parabolic trend of the fitted values and residuals, but there still seems to be areas where the mean residual is non zero.

Equal Variance - Inspecting the plot "Residuals Vs Fitted" we see that there is evidence of non constant variance particularly decreasing as we move left to right on the plot.

Normality assumption - Inspecting the plot "Normal Q-Q" we that the standardized residuals moderately correspond to the theoretical quantiles of a normal distribution. To properly assess if the normality assumption is violated the Shapiro test will be carried out. Additionally, the plot identifies 3 points that have the largest residuals.

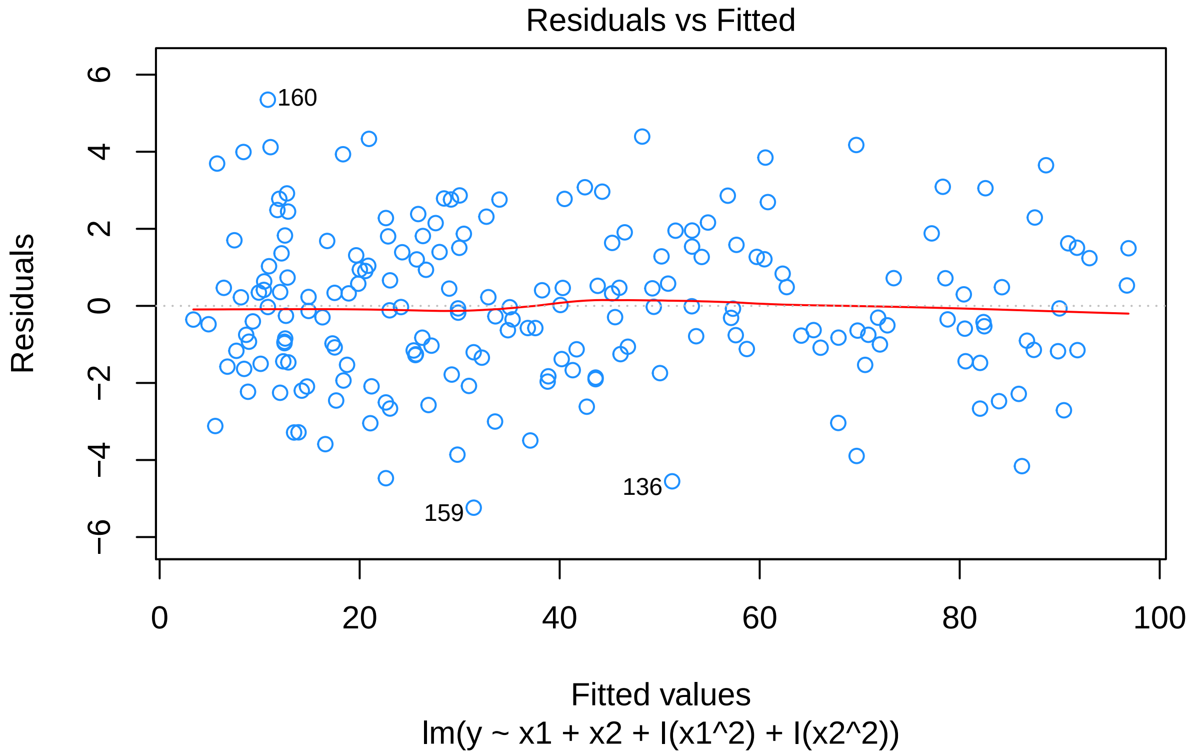
Points of Interest - Also included is a plot of Cook's distance which is a good indicator of point that may have high influence or require further investigation.

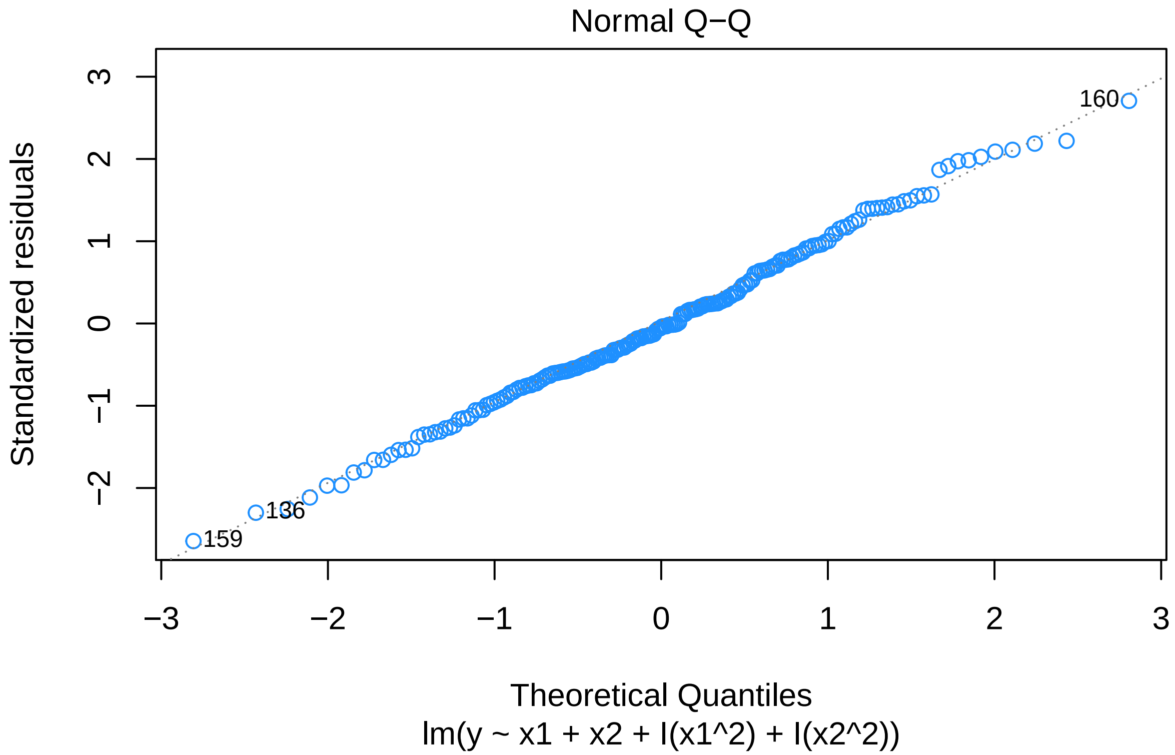
BP Test: p-value < 5% significance level this suggests that the equal variance assumptions is violated for this model.

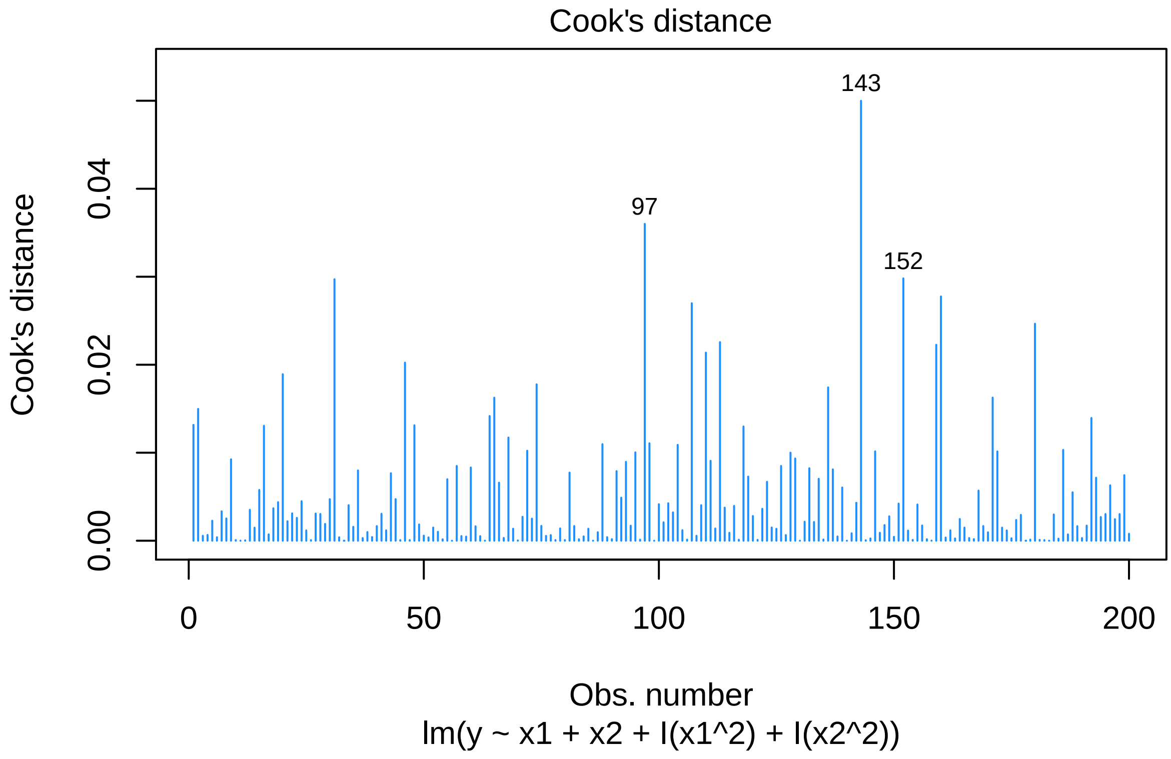
Shapiro Test: p-value < 5% significance level this suggests that the normality assumption doesn't hold for this model.

Therefore, applying a boxcox transformation with lambda of 0.6 **did not** correct the model assumptions.

One G - Quadratic Model




Normal Q−Q
Standardized residuals
−2
−1
0
1
2
3
−3
−2
−1
0
1
2
3
160
136
159
Theoretical Quantiles
lm(y ~ x1 + x2 + I(x1^2) + I(x2^2))



```
##
##  studentized Breusch-Pagan test
##
## data:  lm_1g
## BP = 2.6009, df = 4, p-value = 0.6267
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm_1g)
## W = 0.9956, p-value = 0.8331
```

```
##
## Call:
## lm(formula = y ~ ., data = q1_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.963 -3.503 -1.347  3.473 15.919
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.5112     1.1359  -8.374 1.03e-14 ***
## x1           10.0947     0.1402  71.983  < 2e-16 ***
## x2           -1.2387     0.1309  -9.461  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.927 on 197 degrees of freedom
## Multiple R-squared:  0.9646, Adjusted R-squared:  0.9642
## F-statistic:  2681 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = ((y^(lambda) - 1)/(lambda)) ~ ., data = q1_data_rm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60604 -0.37609 -0.06595  0.35961  1.93982
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.98863    0.15814   6.251 2.79e-09 ***
## x1           2.37331    0.01998 118.783  < 2e-16 ***
## x2          -0.28171    0.01719 -16.384  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6241 on 183 degrees of freedom
## Multiple R-squared:  0.9878, Adjusted R-squared:  0.9877
## F-statistic:  7406 on 2 and 183 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + I(x1^2) + I(x2^2), data = q1_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2370 -1.2533 -0.0942  1.3701  5.3505
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.65216    0.93122   9.291  < 2e-16 ***
## x1           1.30413    0.28367   4.597 7.68e-06 ***
## x2          -0.72887    0.25617  -2.845  0.00491 **
## I(x1^2)      0.77857    0.02463  31.614  < 2e-16 ***
## I(x2^2)     -0.02560    0.02259  -1.133  0.25854
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.995 on 195 degrees of freedom
## Multiple R-squared:  0.9942, Adjusted R-squared:  0.9941
## F-statistic:  8422 on 4 and 195 DF,  p-value: < 2.2e-16
```

The above plots display the following:

- Resiudals vs Fitted Values
- Normal qq plot
- Cook's Distance

Linear Model Appropriateness:

Linearity - Using a quadratic model, the "Residuals Vs Fitted" plot suggests that the linearity assumption holds, though there are a few regions where the mean of residuals is not zero, but is close to zero.

Equal Variance - Inspecting the plot "Residuals Vs Fitted" we see that there is generally constant variance.

Normality assumption - Inspecting the plot "Normal Q-Q" we that the standardized residuals closely correspond to the theoretical quantiles of a normal distribution.

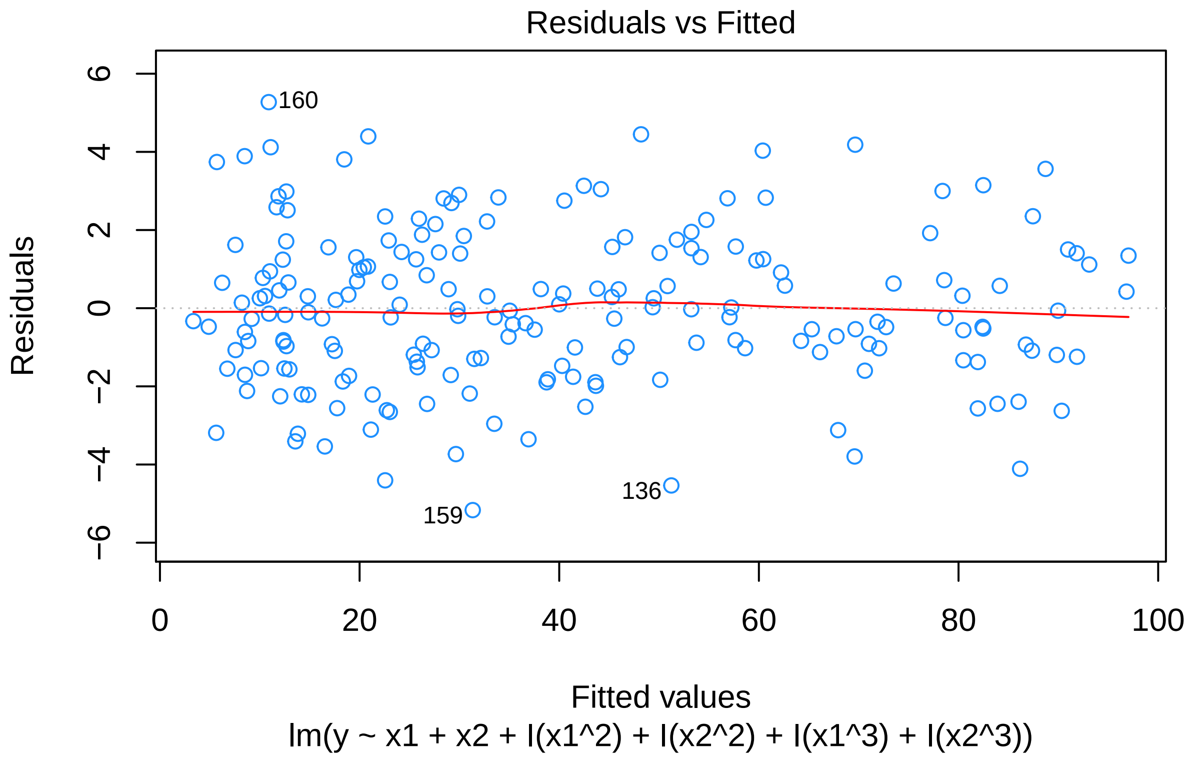
Points of Interest - Also included is a plot of Cook's distance which is a good indicator of point that may have high influence or require further investigation.

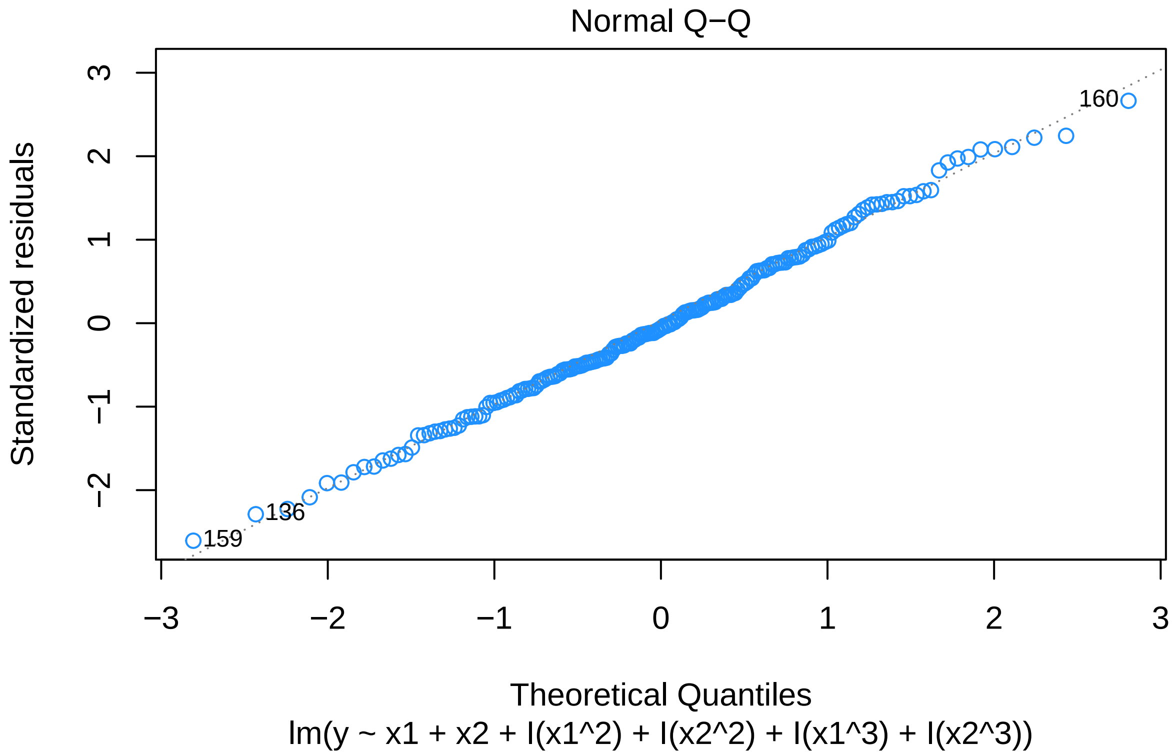
BP Test: p-value > 5% significance level this suggests that the equal variance assumptions holds for this model.

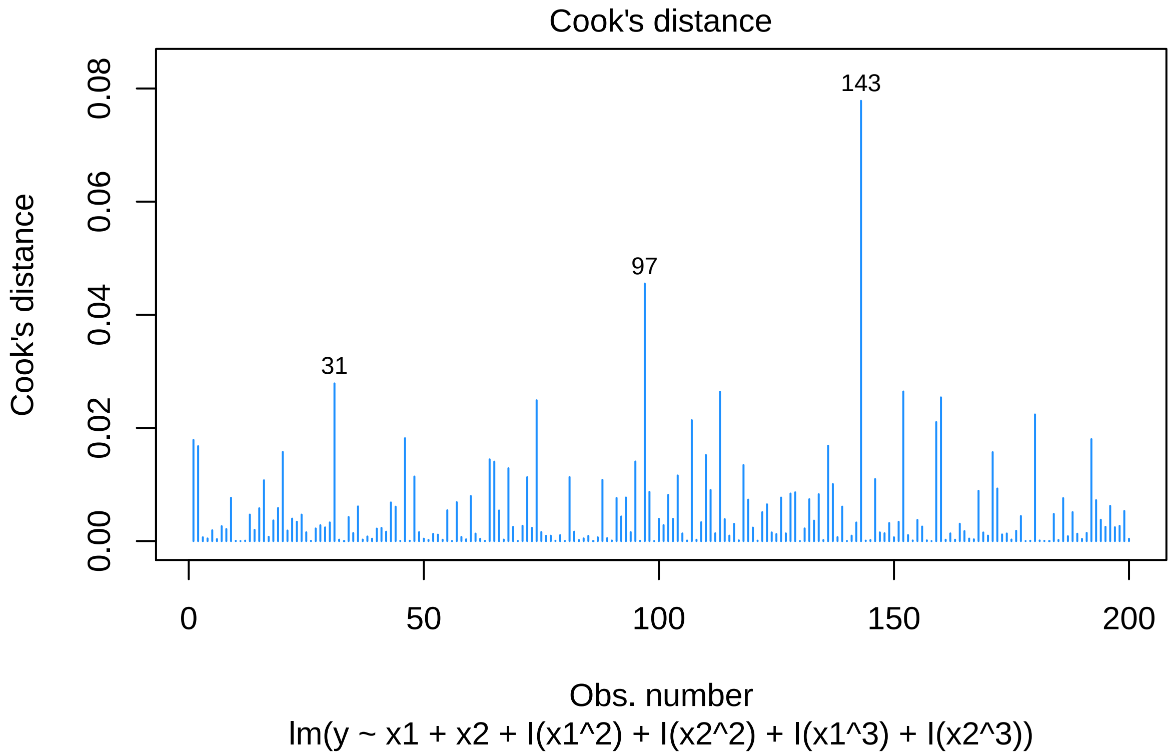
Shapiro Test: p-value > 5% significance level this suggests that the normality assumption holds for this model.

Fitting a polynomial model to the data corrects the model assumptions and even comparing adjusted R-squared values - despite the other models not meeting linearity assumptions - suggest that **this model is more appropriate than those of B and F**.

One H - Cubic Model




Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
160
136
159
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(y ~ x1 + x2 + I(x1^2) + I(x2^2) + I(x1^3) + I(x2^3))



```
##
##  studentized Breusch-Pagan test
##
## data:  lm_1h
## BP = 4.2839, df = 6, p-value = 0.6383
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm_1h)
## W = 0.99579, p-value = 0.8581
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + I(x1^2) + I(x2^2), data = q1_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2370 -1.2533 -0.0942  1.3701  5.3505
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.65216    0.93122   9.291  < 2e-16 ***
## x1           1.30413    0.28367   4.597 7.68e-06 ***
## x2          -0.72887    0.25617  -2.845  0.00491 **
## I(x1^2)      0.77857    0.02463  31.614  < 2e-16 ***
## I(x2^2)     -0.02560    0.02259  -1.133  0.25854
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.995 on 195 degrees of freedom
## Multiple R-squared:  0.9942, Adjusted R-squared:  0.9941
## F-statistic:  8422 on 4 and 195 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + I(x1^2) + I(x2^2) + I(x1^3) + I(x2^3),
##     data = q1_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.166 -1.281 -0.122  1.359  5.273
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.065491   1.810742   5.007 1.25e-06 ***
## x1           1.420580   0.929225   1.529    0.128
## x2          -1.182477   0.801651  -1.475    0.142
## I(x1^2)      0.755965   0.182125   4.151 4.97e-05 ***
## I(x2^2)      0.069683   0.161015   0.433    0.666
## I(x1^3)      0.001279   0.010753   0.119    0.905
## I(x2^3)     -0.005755   0.009623  -0.598    0.551
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.004 on 193 degrees of freedom
## Multiple R-squared:  0.9943, Adjusted R-squared:  0.9941
## F-statistic:  5568 on 6 and 193 DF,  p-value: < 2.2e-16
```

The above plots display the following:

- Resiudals vs Fitted Values
- Normal qq plot
- Cook's Distance

Linear Model Appropriateness:

Linearity - Using a cubic model, the "Residuals Vs Fitted" plot suggests that the linearity assumption holds, though there are a few regions where the mean of residuals is not zero, but is close to zero.

Equal Variance - Inspecting the plot "Residuals Vs Fitted" we see that there is generally constant variance.

Normality assumption - Inspecting the plot "Normal Q-Q" we that the standardized residuals closely correspond to the theoretical quantiles of a normal distribution.

Points of Interest - Also included is a plot of Cook's distance which is a good indicator of point that may have high influence or require further investigation.

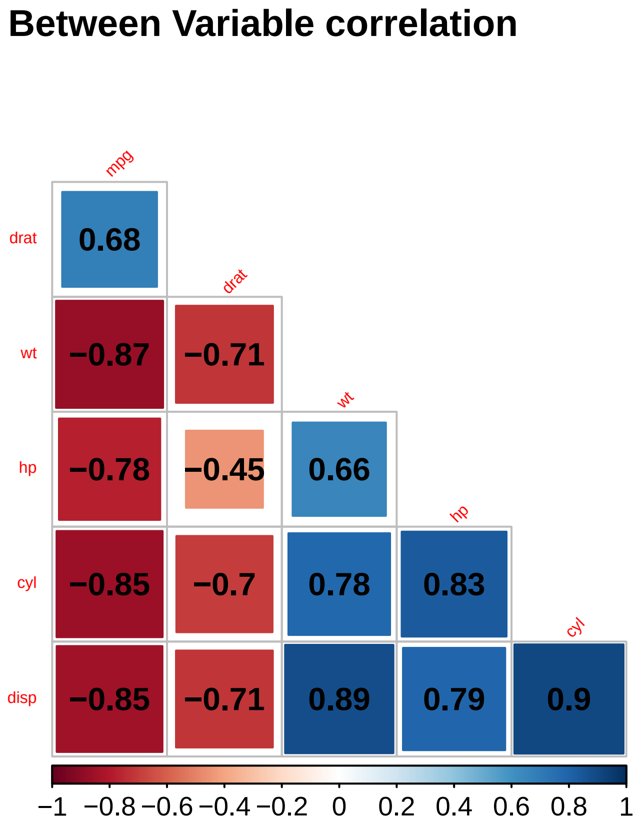
BP Test: p-value > 5% significance level this suggests that the equal variance assumptions holds for this model.

Shapiro Test: p-value > 5% significance level this suggests that the normality assumption holds for this model.

Though linearity assumptions hold for both the quadratic and cubic models, there is some evidence that suggests that the quadratic model is more appropriate. Inspecting the adjusted R-squared model, to account for the difference in number of predictors between the two modes, there is almost no difference between the two models. To avoid potential for overfitting the data, it should be preferred to use the model with few predictors i.e quadratic.

Question 2

Two A - Analysis of Collinearity



```
##       cyl      disp        hp        wt      drat
##  7.869010 10.463957  3.990380  5.168795  2.662298
```

Using a subset of the Mtcars dataset, we would like to assess the data to see if there is any collinearity that may influence our model. A preliminary visualization can be done with a correlogram of the variables. The correlogram **Between predictor correlation** shows that there is high correlation among the variables. We can look further into this by inspecting the VIF of the various predictors. Calculating VIF for the predictors shows that `cyl`, `disp` and `wt` have VIF values greater than 5 suggesting that collinearity does exist, the variable with the highest being `disp`.

Two B - VIF Part I

```
##      cyl       hp       wt     drat
## 6.173560 3.784670 3.076225 2.639229
```

Without using built-in R functions, VIF is calculated using the following steps:

1. Model variable of interest by the other model predictors
2. Determine the R-squared values
3. $$VIF = \frac{1}{1 - R^2}$$

Table 3: VIF by variable

	VIF
cyl	**6.173560**
hp	3.784670
wt	3.076225
drat	2.639229

Table 4: VIF by variable

	VIF
hp	1.769308
wt	2.869445
drat	2.033837

4. Repeat for all remaining variables

The table **Table 3: VIF by variable** shows the VIF by predictor after `dist` was removed. Highlighted in `light blue` is `cyl` which is indicitave of a VIF greater than 5. This suggests that collinearity still exists.

Two C - VIF Part II

After removing `cyl` and `dist`, the VIF for all predictors are below 5. Table **Table 4: VIF by variable** displays the updaed VIFs for each remaining predictor.

Two D - AIC Feature Selection

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt, data = q2_data)
##
## Coefficients:
## (Intercept)          cyl           hp           wt
##    38.75179     -0.94162     -0.01804     -3.16697
```

Using AIC and backward selection, it can be shown that the best subset to consider for the model are `cyl`, `hp` and `wt`.

Two D - Model Selection

Comparing the following models:

Model C: `mpg` ~ `hp` + `wt` + `drat` **Model D**: `mpg` ~ `cyl` + `hp` + `wt`

table **Table 5: Model Selection** shows that the better model, using adj r-squared as the criteria, is model D. This is supported by a larger r squared value.

Question 3

Three A - Best Model

Table 5: Model Selection

	Model	Adj R-Squared
C	Model C	0.8194018
D	**Model D**	**0.8263446**

Table 6: Model Selection

	RSME LOOCV
Model A	3.201673
Model B	**2.688538**
Model C	2.747478

Table 7: Model Selection

	R squared
Model A	0.7528328
Model B	0.8496636
Model C	**0.8667078**

```
##         df      AIC
## model_a  3 166.0294
## model_b  5 154.1194
## model_c  8 156.2687
```

```
##         df      BIC
## model_a  3 170.4266
## model_b  5 161.4481
## model_c  8 167.9946
```

```
##         Adj R-Square
## Model A    0.7445939
## Model B    0.8335561
## Model C    0.8347177
```

From the 2 criteria to select the models from i.e AIC, BIC and adj. R squared, the best model is model B. Model B has the loweset AIC, BIC and nearly the same adj r-squared as C.

Three B - Best Model Part II

The table **Table 6: Model Selection** shows the RSME using LOOCV and shows that model B has the lowest cost function therefore it is the most appropriate model.

Three C - Compare RSME LOOCV to Rsquared

The table **Table 7:Model Selection** shows the Rsquared value and shows that model C has the greatest Rsquared value suggesting that it is the most appropriate model. The rsquared value however, does not take into consideration the increased number of predictors. If models are being compared of varying predictor totals then a measurement that considers the number of preditors should be considered.

Three D - 2 Fold CV

The table **Table 8:Model Selection** shows the RSME value considering 2-fold cross validation shows that model B has the loweset average RSME value suggesting that it is the most appropriate model.

Table 8: Model Selection

	RMSE
Model A	3.406207
Model B	**2.825478**
Model C	3.932775