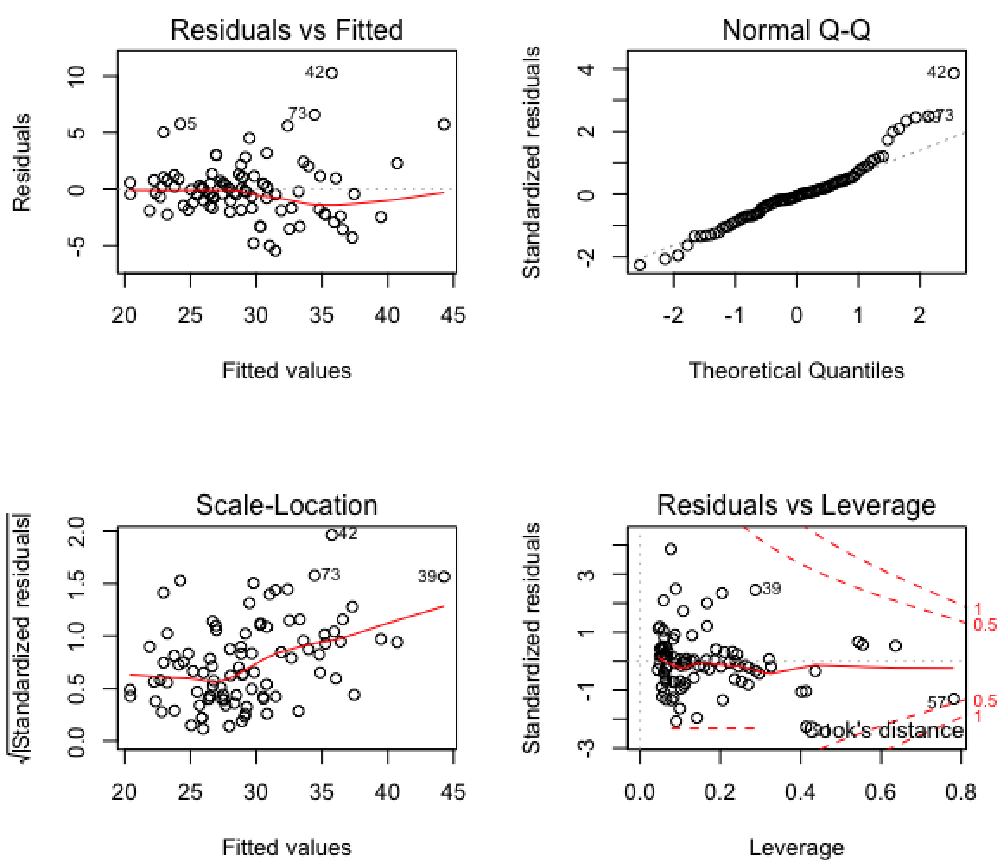
Larce Blake Project 3

**1A)**

Here we are showing how certain variables within a car affects the highway MPG. The interactions with the origin are the Price, Engine Size, Horsepower, Length, and Weight. After plotting the full model we can see that the Mazda RX was an outlier because it had very high leverage, as shown in the model below.

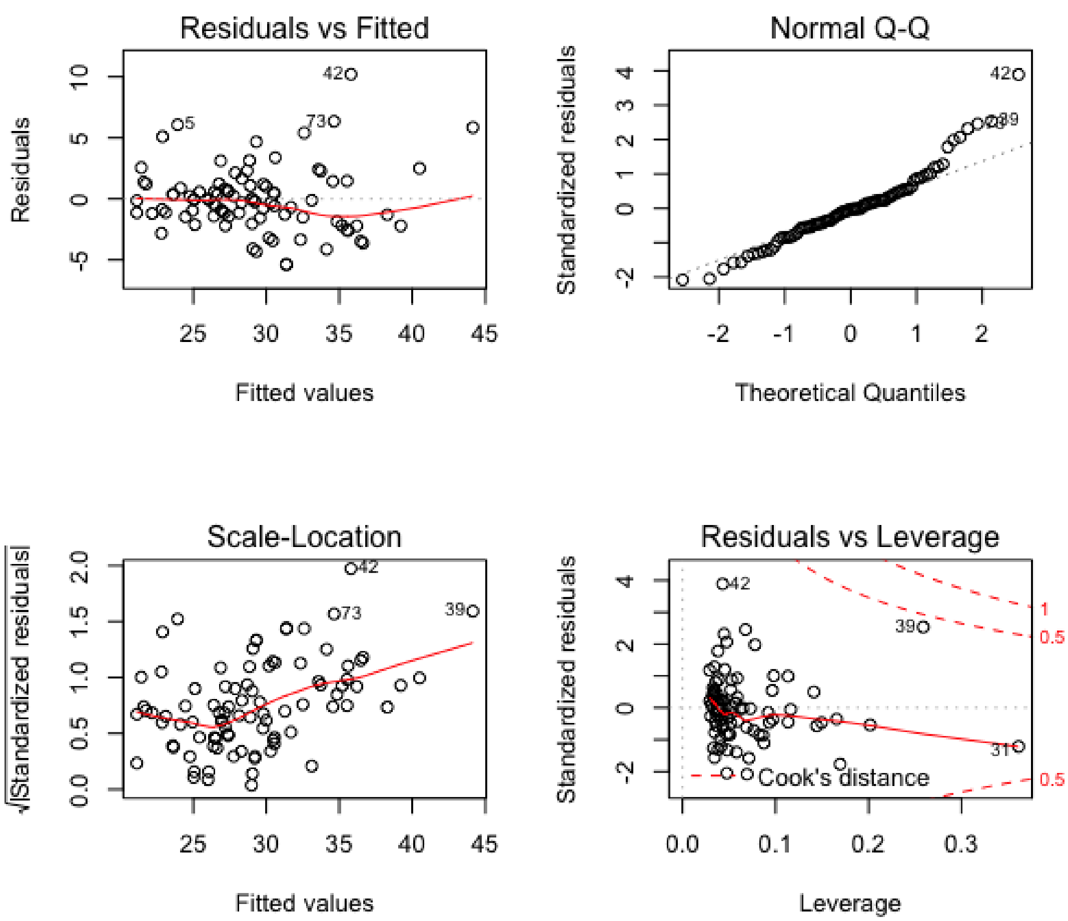
Multiple R-squared is 0.7719




```
Call:
lm(formula = MPG.highway ~ . * Origin - DriveTrain:Origin + I(Weight^2),
    data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4610 -1.6722 -0.1783  0.9504 10.2599

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             4.603e+01  9.760e+00   4.717 1.03e-05 ***
  Price                -4.988e-02  9.643e-02  -0.517 0.606474
  EngineSize           -6.316e-01  9.703e-01  -0.651 0.517042
  Horsepower            1.602e-02  1.450e-02   1.104 0.272792
  DriveTrainFront       1.253e+00  1.135e+00   1.105 0.272725
  DriveTrainRear        1.072e+00  1.533e+00   0.699 0.486470
Length                  1.738e-01  4.850e-02   3.584 0.000587 ***
```

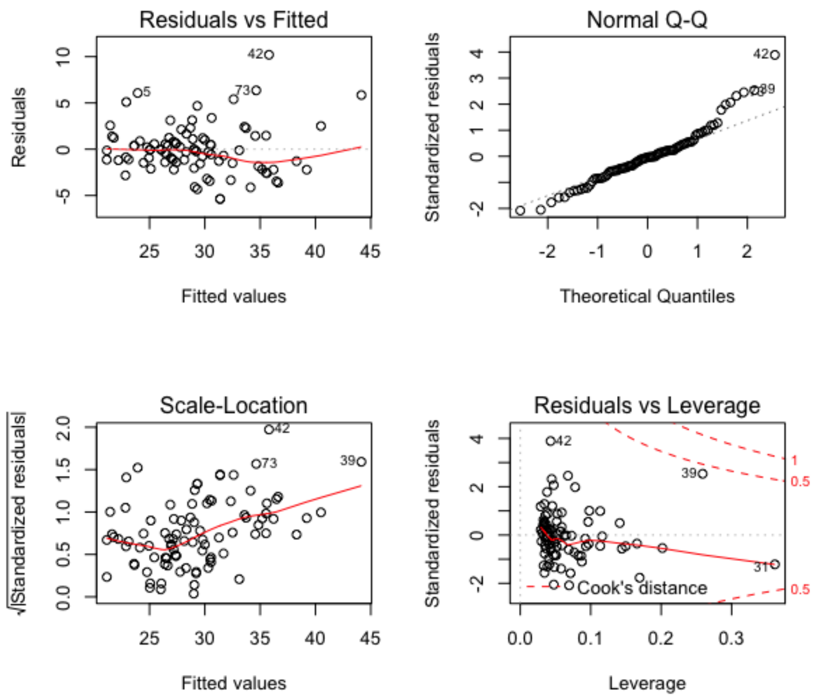
```
Weight                      -2.448e-02  5.939e-03  -4.122 9.29e-05 ***
Originnon-USA                2.419e+01  1.176e+01   2.057 0.043025 *
I(Weight^2)                  2.567e-06  9.096e-07   2.822 0.006044 **
Price:Originnon-USA          2.273e-02  1.252e-01   0.182 0.856376
EngineSize:Originnon-USA     1.578e+00  1.671e+00   0.945 0.347738
Horsepower:Originnon-USA    -2.756e-02  2.298e-02  -1.199 0.234134
Length:Originnon-USA        -1.251e-01  8.520e-02  -1.468 0.146149
Weight:Originnon-USA        -3.122e-04  2.494e-03  -0.125 0.900718
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.765 on 78 degrees of freedom
Multiple R-squared:  0.7719,   Adjusted R-squared:  0.731
F-statistic: 18.86 on 14 and 78 DF,  p-value: < 2.2e-16
```

We then removed the Mazda RX for it was the only significant outlier in the model, and refit the model. Here the only interaction involved was the length and Origin. When we plotted the refit model we saw there were no significant outliers, however the Geo Metro and Honda Civic did stand out as shown below.



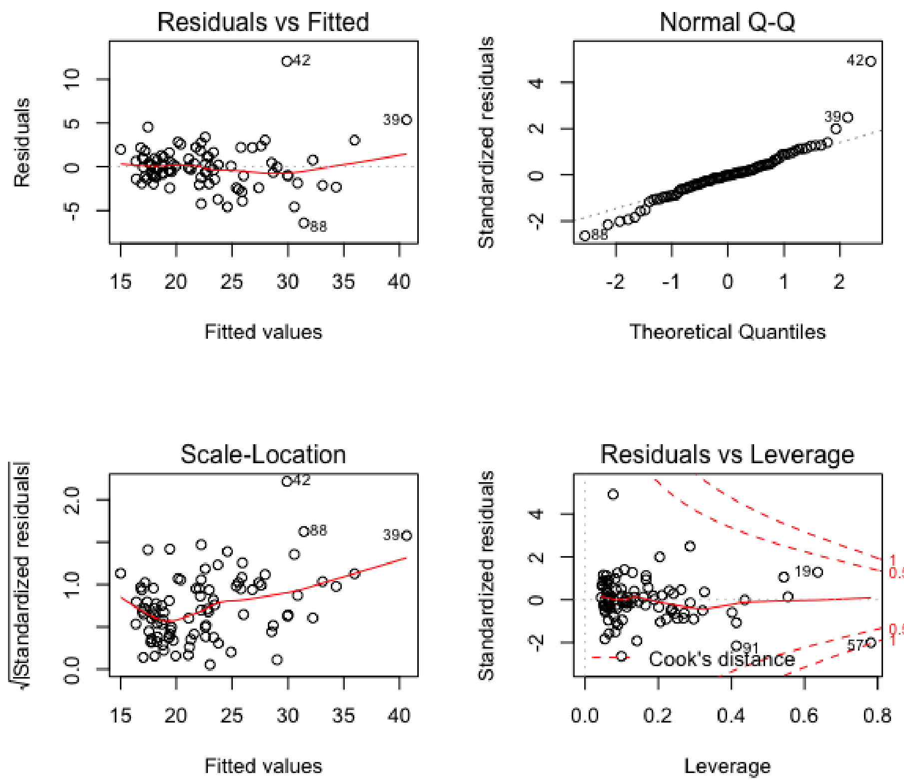
```
Call:
lm(formula = MPG.highway ~ Length + Weight + Origin + I(Weight^2) +
    Length:Origin, data = cars2)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.3855 -1.4247 -0.1144  1.0959 10.1985 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            4.881e+01  8.881e+00   5.496 3.82e-07 ***
Length                 1.781e-01  3.568e-02   4.992 3.05e-06 ***
Weight                -2.562e-02  5.314e-03  -4.822 6.00e-06 ***
Originnon-USA          2.114e+01  8.437e+00   2.506  0.01408 *  
I(Weight^2)            2.628e-06  8.285e-07   3.173  0.00209 ** 
Length:Originnon-USA -1.114e-01  4.651e-02  -2.395  0.01875 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.68 on 87 degrees of freedom
Multiple R-squared:  0.761,   Adjusted R-squared:  0.7473 
F-statistic:  55.4 on 5 and 87 DF,  p-value: < 2.2e-16
```

After refitting the model we could reduce the model using BIC and summarize the results. When doing this the Multiple r-squared= 0.761 which is greater than 0.5 so assumptions are reasonable. Essential the plots came out very similar to the refit model. The Ford Festiva had very high leverage while the Honda Civic and Geo Metro had high residuals. So overall we can say that the Geo Metro had is the highest influence amongst highway MPG.



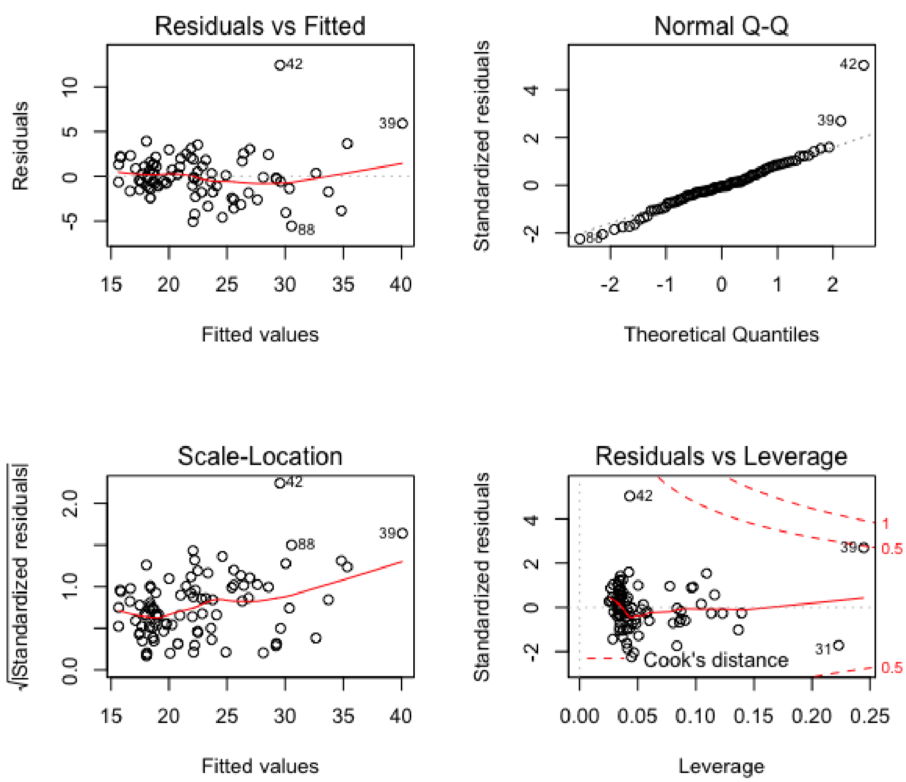

**1B)**

Instead of comparing all of the variable's to the MPG on the highway, we are comparing the variable's to the MPG in the city. The interactions with the origin are the Price, Engine Size, Horsepower, Length, and Weight. After plotting the full model we can see that the Geo Metro is causing a relatively large change in the effect for the origin due to its high leverage. So it has a high influence for MPG city.




```
                         Call:
lm(formula = MPG.city ~ . * Origin - DriveTrain:Origin + I(Weight^2),
                         data = cars3)

                         Residuals:
          Min       1Q  Median       3Q      Max
      -6.4249 -1.1522 -0.0375   1.0168 12.0882

                         Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)               6.970e+01  9.046e+00   7.705 3.50e-11 ***
  Price                  -5.695e-02  8.938e-02  -0.637    0.526
  EngineSize             -1.168e+00  8.994e-01  -1.298    0.198
  Horsepower              9.085e-03  1.344e-02   0.676    0.501
  DriveTrainFront         9.947e-01  1.052e+00   0.946    0.347
  DriveTrainRear          5.264e-01  1.420e+00   0.371    0.712
  Length                  7.436e-02  4.496e-02   1.654    0.102
Weight                  -3.355e-02  5.505e-03  -6.094 3.93e-08 ***
  Originnon-USA           7.380e+00  1.090e+01   0.677    0.500
I(Weight^2)               4.461e-06  8.431e-07   5.291 1.08e-06 ***
  Price:Originnon-USA     4.027e-02  1.160e-01   0.347    0.729
  EngineSize:Originnon-USA  1.790e+00  1.549e+00   1.156    0.251
  Horsepower:Originnon-USA -2.348e-02  2.130e-02  -1.102    0.274
  Length:Originnon-USA    2.781e-03  7.897e-02   0.035    0.972
  Weight:Originnon-USA   -2.797e-03  2.312e-03  -1.210    0.230
                              ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Residual standard error: 2.563 on 78 degrees of freedom
    Multiple R-squared:  0.8236,     Adjusted R-squared:  0.792
     F-statistic: 26.02 on 14 and 78 DF,  p-value: < 2.2e-16
```

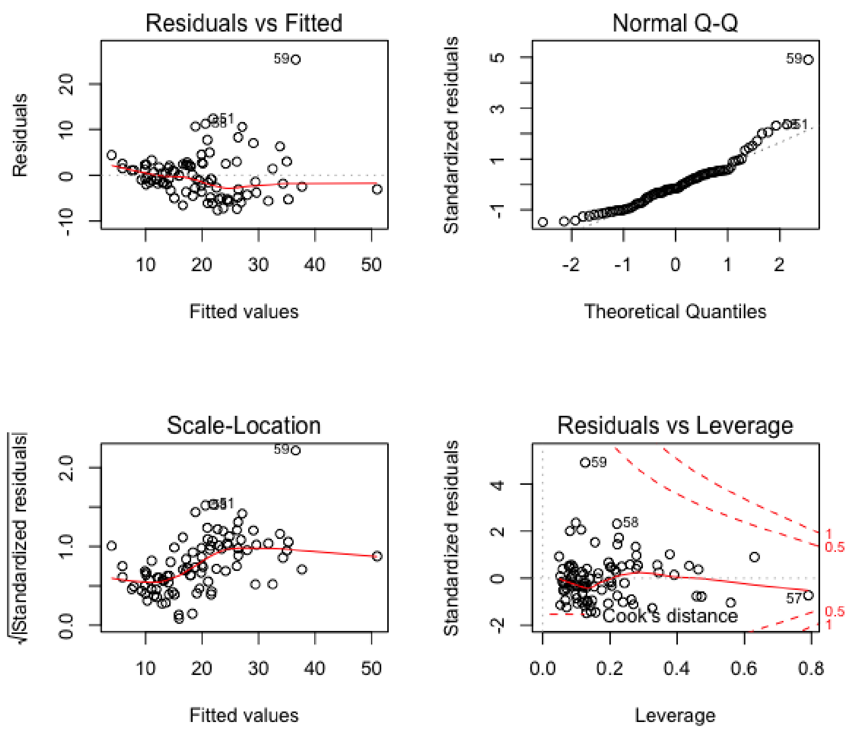
After that we remove the Mazda RX7 because of its rotary engine since it is different from everything else. So when we refit the model it appears the Ford Festiva has high leverage. We also notice that the Corvette has such high leverage. Since the Corvette is such a light car but has high engine size and horsepower, we can can assume that's the reasoning for its high leverage. So we filter the horse powers > 250 and reduce the model, where we can state that assumptions of normaility are good..




```
                          Call:
lm(formula = MPG.city ~ Length + Weight + Origin + I(Weight^2),
                     data = cars3)

                        Residuals:
                Min      1Q  Median      3Q     Max
            -5.5520 -1.2969 -0.1041  1.3235 12.4423

                      Coefficients:
                Estimate Std. Error t value Pr(>|t|)
    (Intercept)    7.714e+01  7.182e+00  10.741  < 2e-16 ***
     Length        7.881e-02  3.300e-02   2.388   0.0191 *
    Weight        -3.752e-02  4.679e-03  -8.018 4.25e-12 ***
     Originnon-USA  1.311e+00  5.692e-01   2.303   0.0236 *
    I(Weight^2)    4.644e-06  7.398e-07   6.277 1.27e-08 ***
                               ---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 2.527 on 88 degrees of freedom
  Multiple R-squared:  0.8066,    Adjusted R-squared:  0.7978
    F-statistic: 91.76 on 4 and 88 DF,  p-value: < 2.2e-16
```

**1C)**

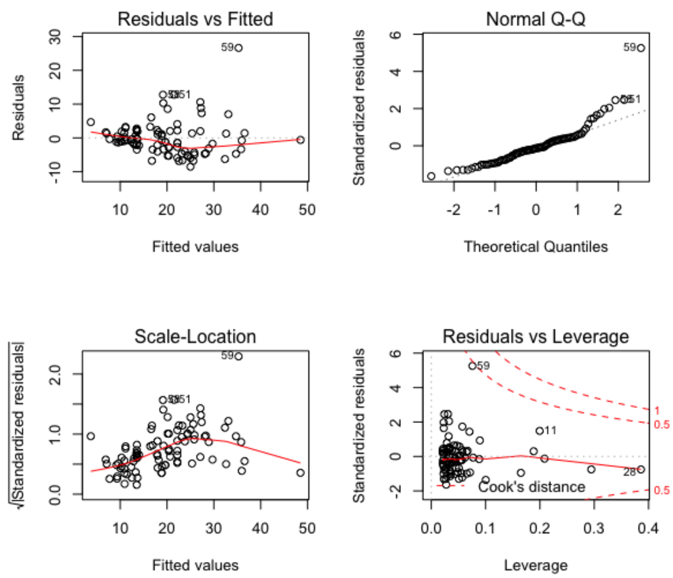
Here we are going to using the variable's MPG.highway, MPG.city, EngineSize, Horsepower, DriveTrain, Length, Weight, Origin to predict price. When looking at the full model we can see that the Mercedes Benz 300 has high residuals and the Mazda RX7 has high leverage. The Mercedes more than likely has such a high residual due to it being a lot more expensive compared to most cars, but it still posses the same quality of the more reasonably priced cars.




```
Call:
lm(formula = Price ~ . * Origin - DriveTrain:Origin + I(Weight^2),
    data = cars5)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6296 -3.1225 -0.7763  2.3285 25.3339

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.854e+01  3.560e+01   0.521  0.60410
MPG.highway              8.744e-02  5.436e-01   0.161  0.87265
MPG.city                -5.975e-01  8.223e-01  -0.727  0.46968
EngineSize               4.504e-01  1.951e+00   0.231  0.81803
Horsepower               8.252e-02  2.539e-02   3.249  0.00172 **
DriveTrainFront          8.708e-01  2.323e+00   0.375  0.70874
DriveTrainRear           3.157e+00  3.018e+00   1.046  0.29879
Length                   7.313e-02  1.186e-01   0.616  0.53942
Weight                  -9.859e-03  1.607e-02  -0.613  0.54150
Originnon-USA           -1.880e+01  3.500e+01  -0.537  0.59280
I(Weight^2)              1.242e-06  2.276e-06   0.545  0.58703
MPG.highway:Originnon-USA -3.453e-01  8.585e-01  -0.402  0.68862
MPG.city:Originnon-USA   7.117e-01  9.587e-01   0.742  0.46020
EngineSize:Originnon-USA 3.155e+00  3.255e+00   0.969  0.33537
Horsepower:Originnon-USA 5.423e-02  4.086e-02   1.327  0.18846
Length:Originnon-USA     1.392e-02  1.848e-01   0.075  0.94015
Weight:Originnon-USA    -6.236e-05  6.782e-03  -0.009  0.99269
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.524 on 76 degrees of freedom
Multiple R-squared:  0.7299,    Adjusted R-squared:  0.673
F-statistic: 12.83 on 16 and 76 DF,  p-value: 1.343e-15
```

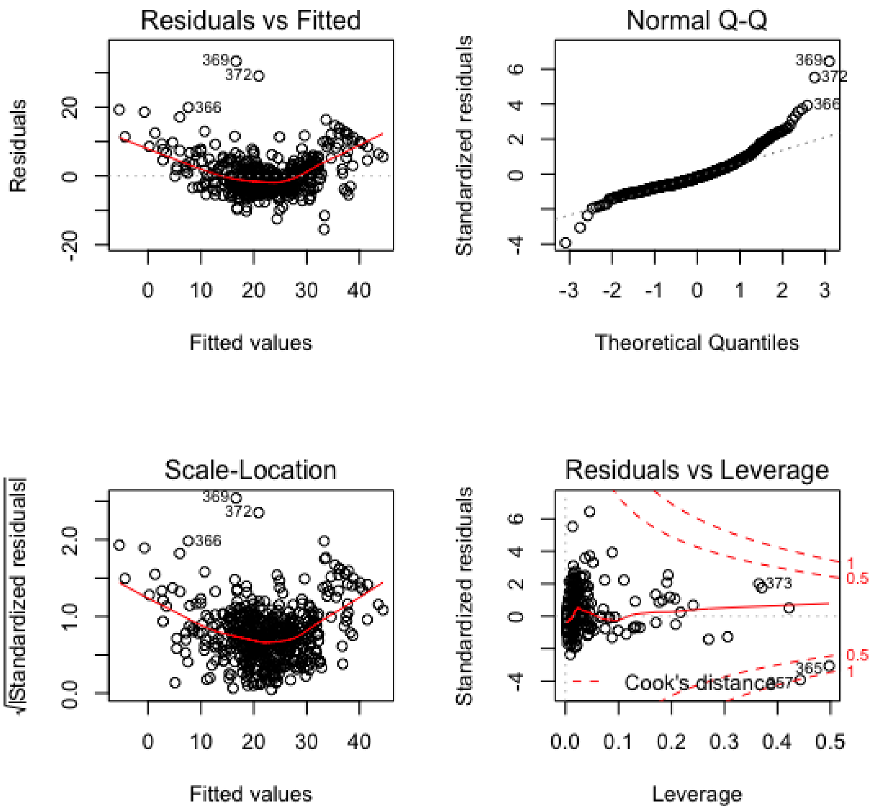
We can then reduce the model and see that the assumptions of normality are good.




```
Call:
lm(formula = Price ~ EngineSize + Horsepower + Origin + Horsepower:Origin,
    data = cars5)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5078 -3.3283 -0.6437  1.9145 26.6249

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              -0.30545    2.30211  -0.133 0.894746
EngineSize                2.83455    0.92403   3.068 0.002867 **
Horsepower                0.06905    0.02058   3.355 0.001173 **
Originnon-USA            -7.40398    3.22361  -2.297 0.024004 *
Horsepower:Originnon-USA  0.08724    0.02165   4.029 0.000119 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.268 on 88 degrees of freedom
Multiple R-squared:  0.7155,  Adjusted R-squared:  0.7026
F-statistic: 55.34 on 4 and 88 DF,  p-value: < 2.2e-16
```

**2A)**

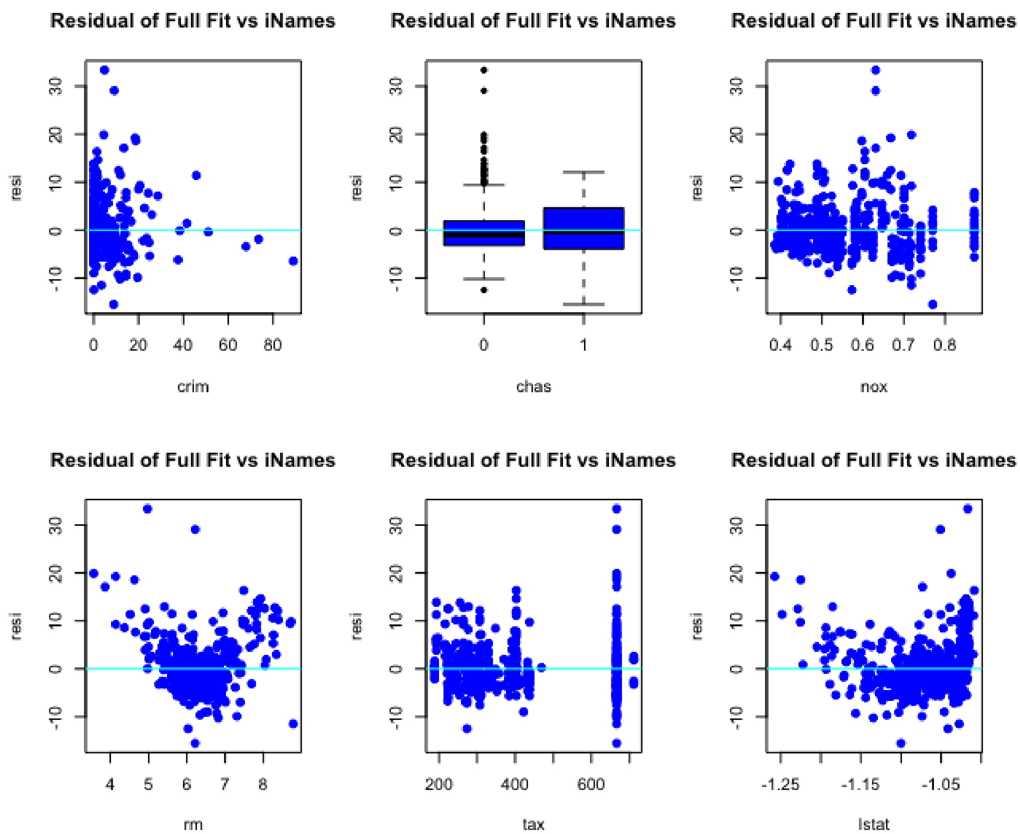
When fitting our model, we can see that the residual vs fitted is in a bowl shape we applied the quadratic term. Also we can note that #357 & #365 are outside of cook's D




```
      medv             crim              chas         nox                rm
 Min.   : 5.00   Min.   : 0.00632   0:471    Min.   :0.3850   Min.   :3.561
 1st Qu.:17.02   1st Qu.: 0.08204   1: 35    1st Qu.:0.4490   1st Qu.:5.886
 Median :21.20   Median : 0.25651            Median :0.5380   Median :6.208
 Mean   :22.53   Mean   : 3.61352            Mean   :0.5547   Mean   :6.285
 3rd Qu.:25.00   3rd Qu.: 3.67708            3rd Qu.:0.6240   3rd Qu.:6.623
 Max.   :50.00   Max.   :88.97620            Max.   :0.8710   Max.   :8.780
                           tax             lstat
                      Min.   :187.0   Min.   :-1.258
                      1st Qu.:279.0   1st Qu.:-1.096
                      Median :330.0   Median :-1.062
                      Mean   :408.2   Mean   :-1.072
                      3rd Qu.:666.0   3rd Qu.:-1.036
                      Max.   :711.0   Max.   :-1.009
```

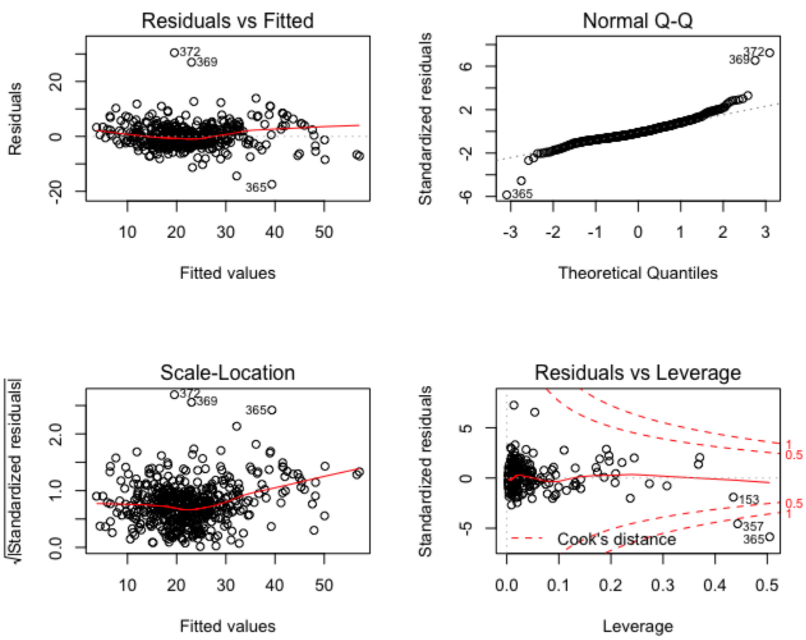
**2B)**

First we look at the diagnostic plots, and we see that they look OK.




```
                    Call:
lm(formula = medv ~ . * chas, data = boston)

                   Residuals:
    Min      1Q  Median      3Q     Max
-15.505  -3.237  -0.970   2.013  33.367

                  Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   63.570375  10.118488   6.283 7.31e-10 ***
 crim         -0.064699   0.034787  -1.860  0.06349 .
 chas1       140.558910  45.976214   3.057  0.00236 **
  nox          2.797029   3.175809   0.881  0.37889
rm             5.569485   0.444026  12.543  < 2e-16 ***
tax           -0.008696   0.002211  -3.932 9.61e-05 ***
lstat         69.121899   8.377214   8.251 1.44e-15 ***
crim:chas1     3.462350   0.882693   3.922  0.00010 ***
 chas1:nox   -32.968826  10.191978  -3.235  0.00130 **
  chas1:rm    -3.331025   1.643954  -2.026  0.04328 *
   chas1:tax  -0.017987   0.015193  -1.184  0.23702
  chas1:lstat 89.297373  35.842001   2.491  0.01305 *
                       ---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

   Residual standard error: 5.296 on 494 degrees of freedom
  Multiple R-squared:  0.6756,     Adjusted R-squared:  0.6684
   F-statistic: 93.55 on 11 and 494 DF,  p-value: < 2.2e-16
```

To find out which one, we check each predictor by refitting. Once we refit we can see that that the residual vs. fitted has a significant curve. Therefore the assumption is not reasonable. Although the QQ-Plot looks good.



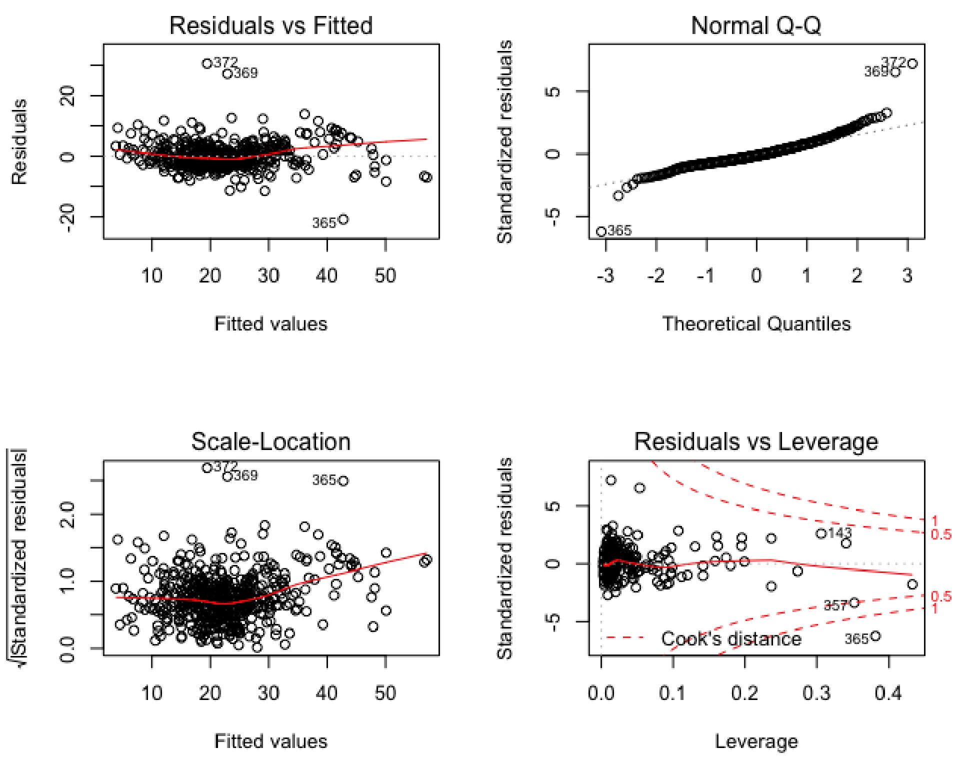

To figure out why there are two very high leverages we converted the varibles into a factor. The coeffeicient for those varibles are changed by the removal or addition of that observation. So now we identify those observations and can indicate where its taking effect when we look at the first row and first column.

```
> boston[1,]
  medv    crim chas   nox    rm tax     lstat
1   24 0.00632    0 0.538 6.575 296 -1.025759
> boston[,1]
  [1] 24.0 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 15.0 18.9 21.7 20.4
 [15] 18.2 19.9 23.1 17.5 20.2 18.2 13.6 19.6 15.2 14.5 15.6 13.9 16.6 14.8
 [29] 18.4 21.0 12.7 14.5 13.2 13.1 13.5 18.9 20.0 21.0 24.7 30.8 34.9 26.6
 [43] 25.3 24.7 21.2 19.3 20.0 16.6 14.4 19.4 19.7 20.5 25.0 23.4 18.9 35.4
 [57] 24.7 31.6 23.3 19.6 18.7 16.0 22.2 25.0 33.0 23.5 19.4 22.0 17.4 20.9
 [71] 24.2 21.7 22.8 23.4 24.1 21.4 20.0 20.8 21.2 20.3 28.0 23.9 24.8 22.9
 [85] 23.9 26.6 22.5 22.2 23.6 28.7 22.6 22.0 22.9 25.0 20.6 28.4 21.4 38.7
 [99] 43.8 33.2 27.5 26.5 18.6 19.3 20.1 19.5 19.5 20.4 19.8 19.4 21.7 22.8
[113] 18.8 18.7 18.5 18.3 21.2 19.2 20.4 19.3 22.0 20.3 20.5 17.3 18.8 21.4
[127] 15.7 16.2 18.0 14.3 19.2 19.6 23.0 18.4 15.6 18.1 17.4 17.1 13.3 17.8
[141] 14.0 14.4 13.4 15.6 11.8 13.8 15.6 14.6 17.8 15.4 21.5 19.6 15.3 19.4
[155] 17.0 15.6 13.1 41.3 24.3 23.3 27.0 50.0 50.0 50.0 22.7 25.0 50.0 23.8
[169] 23.8 22.3 17.4 19.1 23.1 23.6 22.6 29.4 23.2 24.6 29.9 37.2 39.8 36.2
[183] 37.9 32.5 26.4 29.6 50.0 32.0 29.8 34.9 37.0 30.5 36.4 31.1 29.1 50.0
[197] 33.3 30.3 34.6 34.9 32.9 24.1 42.3 48.5 50.0 22.6 24.4 22.5 24.4 20.0
[211] 21.7 19.3 22.4 28.1 23.7 25.0 23.3 28.7 21.5 23.0 26.7 21.7 27.5 30.1
[225] 44.8 50.0 37.6 31.6 46.7 31.5 24.3 31.7 41.7 48.3 29.0 24.0 25.1 31.5
[239] 23.7 23.3 22.0 20.1 22.2 23.7 17.6 18.5 24.3 20.5 24.5 26.2 24.4 24.8
[253] 29.6 42.8 21.9 20.9 44.0 50.0 36.0 30.1 33.8 43.1 48.8 31.0 36.5 22.8
```

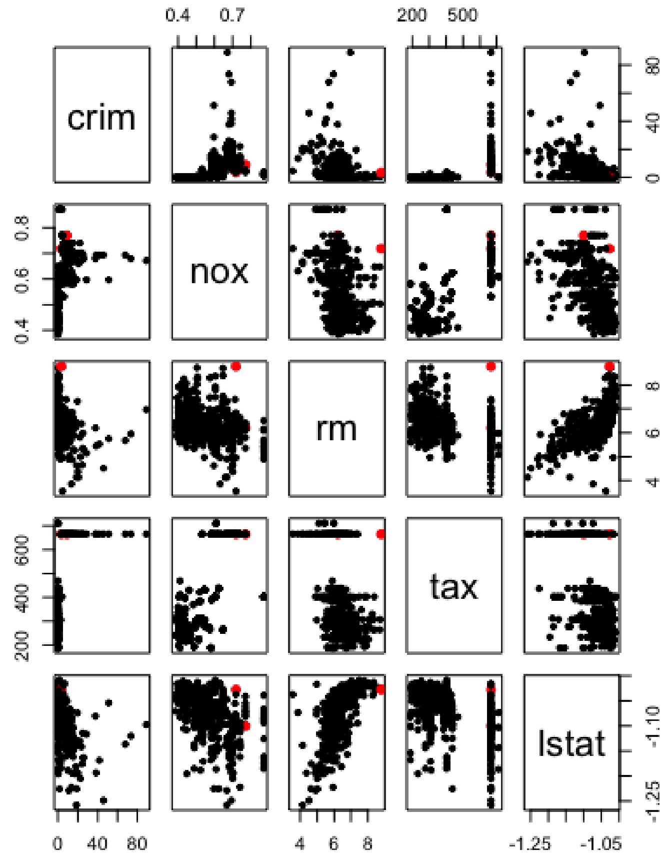
```
[267] 30.7 50.0 43.5 20.7 21.1 25.2 24.4 35.2 32.4 32.0 33.2 33.1 29.1 35.1
[281] 45.4 35.4 46.0 50.0 32.2 22.0 20.1 23.2 22.3 24.8 28.5 37.3 27.9 23.9
[295] 21.7 28.6 27.1 20.3 22.5 29.0 24.8 22.0 26.4 33.1 36.1 28.4 33.4 28.2
[309] 22.8 20.3 16.1 22.1 19.4 21.6 23.8 16.2 17.8 19.8 23.1 21.0 23.8 23.1
[323] 20.4 18.5 25.0 24.6 23.0 22.2 19.3 22.6 19.8 17.1 19.4 22.2 20.7 21.1
[337] 19.5 18.5 20.6 19.0 18.7 32.7 16.5 23.9 31.2 17.5 17.2 23.1 24.5 26.6
[351] 22.9 24.1 18.6 30.1 18.2 20.6 17.8 21.7 22.7 22.6 25.0 19.9 20.8 16.8
[365] 21.9 27.5 21.9 23.1 50.0 50.0 50.0 50.0 50.0 13.8 13.8 15.0 13.9 13.3
[379] 13.1 10.2 10.4 10.9 11.3 12.3  8.8  7.2 10.5  7.4 10.2 11.5 15.1 23.2
[393]  9.7 13.8 12.7 13.1 12.5  8.5  5.0  6.3  5.6  7.2 12.1  8.3  8.5  5.0
[407] 11.9 27.9 17.2 27.5 15.0 17.2 17.9 16.3  7.0  7.2  7.5 10.4  8.8  8.4
[421] 16.7 14.2 20.8 13.4 11.7  8.3 10.2 10.9 11.0  9.5 14.5 14.1 16.1 14.3
[435] 11.7 13.4  9.6  8.7  8.4 12.8 10.5 17.1 18.4 15.4 10.8 11.8 14.9 12.6
[449] 14.1 13.0 13.4 15.2 16.1 17.8 14.9 14.1 12.7 13.5 14.9 20.0 16.4 17.7
[463] 19.5 20.2 21.4 19.9 19.0 19.1 19.1 20.1 19.9 19.6 23.2 29.8 13.8 13.3
[477] 16.7 12.0 14.6 21.4 23.0 23.7 25.0 21.8 20.6 21.2 19.1 20.6 15.2  7.0
[491]  8.1 13.6 20.1 21.8 24.5 23.1 19.7 18.3 21.2 17.5 16.8 22.4 20.6 23.9
                                [505] 22.0 11.9
```

**2C)**

Here we want to reduce using BIC and interpret the model.




We can see that numbers 372, 369, and 365 have residuals on the Residual vs Fitted Line. Along with that, they also have high leverages. We then checked 357 and 365 to determine that 357 has a relatively high influence and because chas= 1 it is close to the river.So we need to look at the pair plots but with the high leverage points.




**2D)**

Here we need to construct a 95% confidence interval model for predicted medv at the mean. We are selecting the houses by the river and not by the river. Looking at the residual vs. fitted gives you an overall understanding.

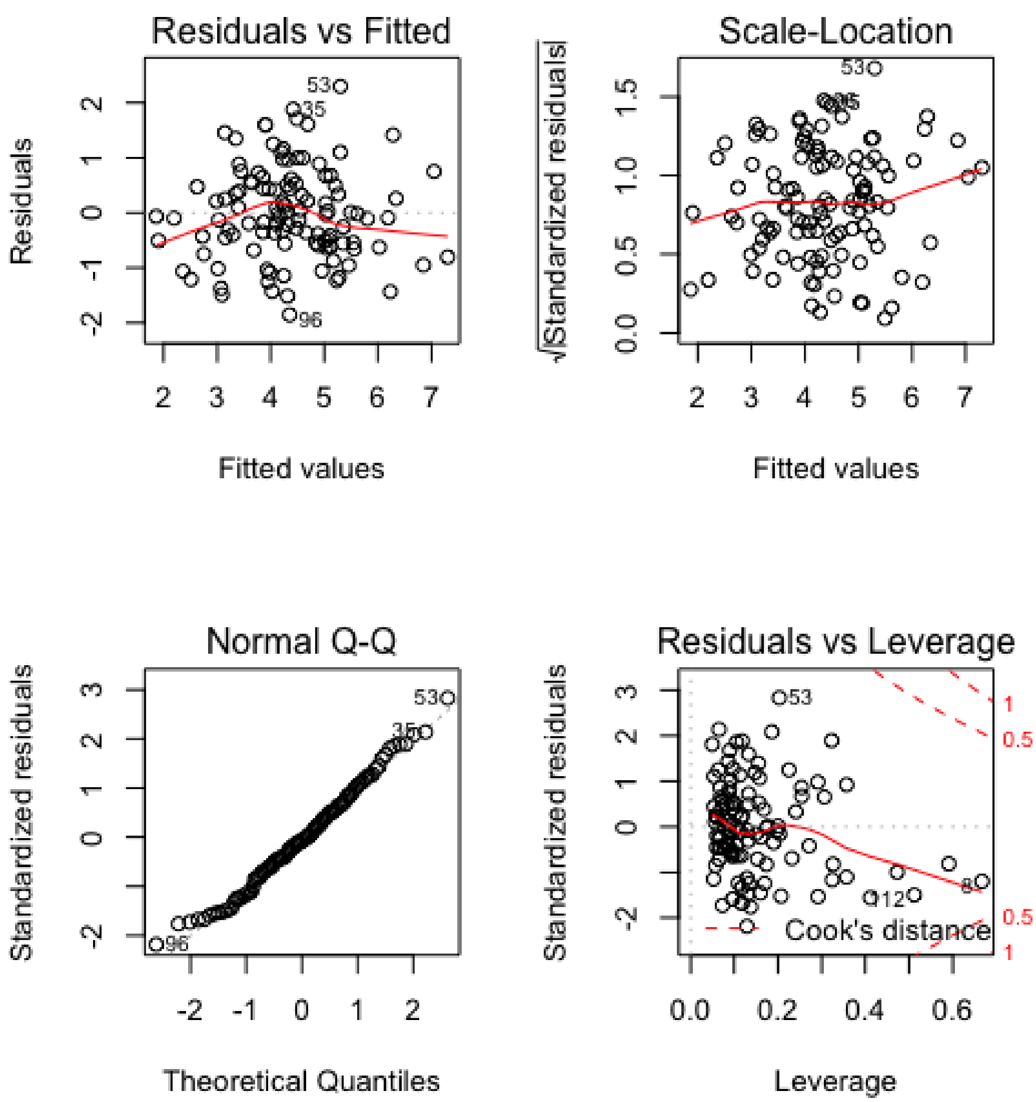
| | *fit* | *lwr* | *upr* |
|---|---|---|---|
| *1* | *23.22864* | *22.60524* | *23.85205* |
| *2* | *37.87515* | *35.12868* | *40.62162* |

**3A)**

Larce Blake Project 3

In this problem we are looking at infection rates amongst the data set. So we are fitting an appropriate model to predict infection risks with interactions between MedSchool and the variables Culturing, Xray, and Nurses.  While making the model we will add in the interaction from the question to the model. Based on the QQ-Plot below we can say that the normality is almost exact. Along with that, only 8 has high leverage and 53 has high residual.

Multiple R-squared= 0.6052



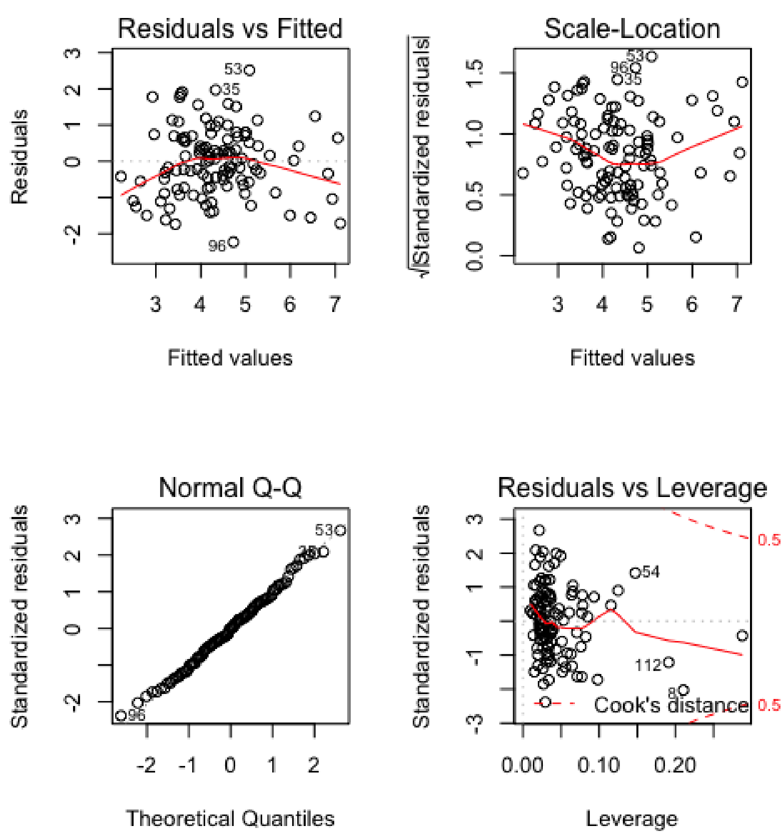
**3B)**

Here we are reducing the model to the most important varibles to predict infection rate. We will then observe to see if there are any major changes.

```
             Df Sum of Sq     RSS     AIC
<none>                     97.441  6.8976
- Xray        1    4.5775 102.019  7.3577
- Facilities  1    9.3791 106.821 12.5548
- Stay        1   10.6987 108.140 13.9421
- Culturing   1   19.7651 117.207 23.0397
```

We can see that after using BIC that the only variables left are X-ray, Facilities, Stay, and Culturing.

```
                     Coefficients:
               Estimate Std. Error t value Pr(>|t|)
  (Intercept) -0.063581   0.533207  -0.119 0.905305
Stay         0.188411    0.054714   3.444 0.000818 ***
Culturing    0.046446    0.009923   4.680 8.35e-06 ***
 Xray         0.012052    0.005351   2.252 0.026316 *
 Facilities   0.020465    0.006347   3.224 0.001671 **
                          ---

Residual standard error: 0.9499 on 108 degrees of freedom
Multiple R-squared:  0.5161,    Adjusted R-squared:  0.4982
F-statistic:  28.8 on 4 and 108 DF,  p-value: 2.728e-16
```



We can see that the highest infection rate comes from the longer you stay in the hospital. Facilities and Stay both have positive correlations since they are byproducts of infections. Everything looks normal with no outliers.

**3C)**
Here we had to add new values to the varibles in the study. We then had to obtain a 95% prediction interval for the infection risk amongst all the hospitals. When we attempted to the predict the infection rate of the hospital, we got the following values.

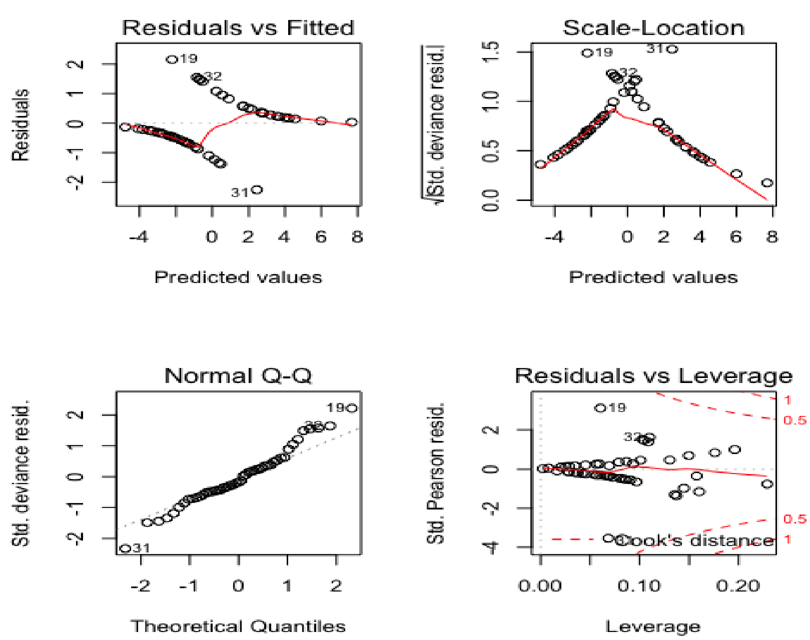
| fit | lwr | upr |
|---|---|---|
| 4.988875 | 3.066227 | 6.911523 |

The mean infection rate is seen to be between 3.1, and 6.9. So following this we must compare the mean of this certain hospital to the mean of all of the hospitals which comes out too:

4.354867

**4A)**

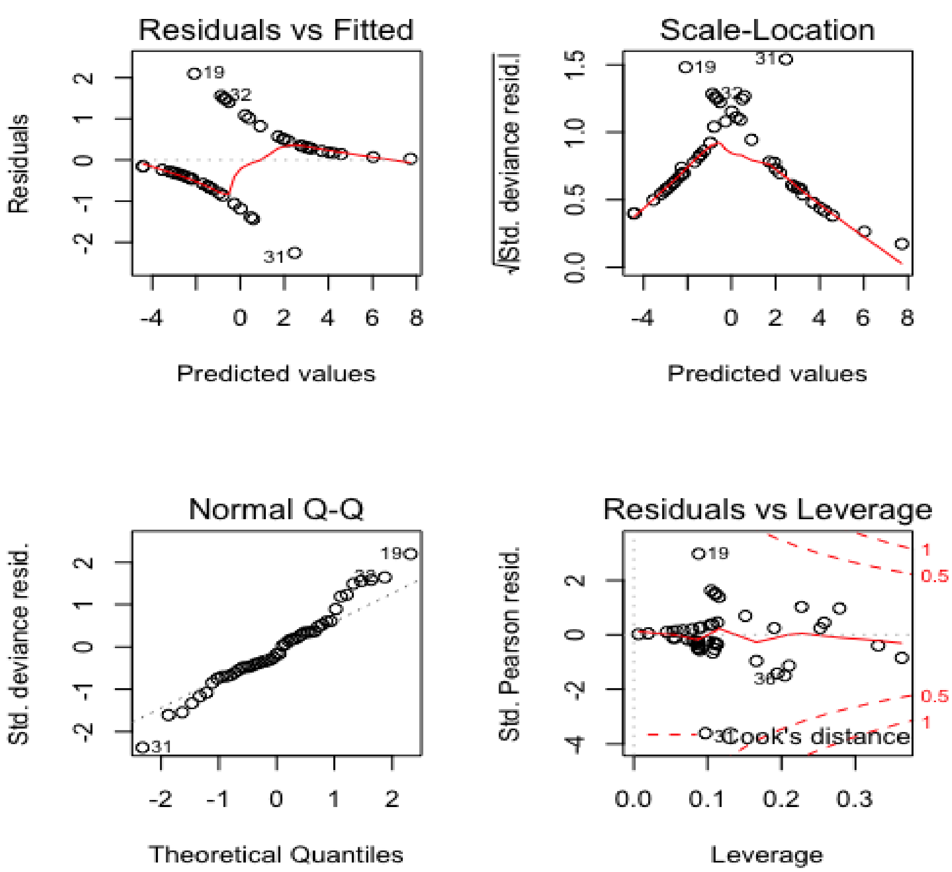
To determine weather there is a significant interaction between severity and hospital we fit the model and reduced it down to its final model. Intially we will factor the Outcome and Hosptial into the data without the interactions. Where all plots look reasonable.



Now we fit the model with the missing interactions and reduce the model using BIC where we are left with the following data.

```
Start:  AIC=58.03
Outcome ~ Severity * Hospital

                    Df Deviance    AIC
- Severity:Hospital  2   34.742 50.309
<none>                   34.676 58.027

Step:  AIC=50.31
Outcome ~ Severity + Hospital

           Df Deviance    AIC
<none>          34.742 50.309
- Hospital  2   45.994 53.778
- Severity  1   53.490 65.165
> summary(ocBIC.lm)
```

We can assume that this is accurate because we are comparing the 3 hospitals and the severity rates. The other variables and interactions would be useful for this test. We only kept additive model we can assume the interactions between severity and hospital isn’t significant.



**4B)**

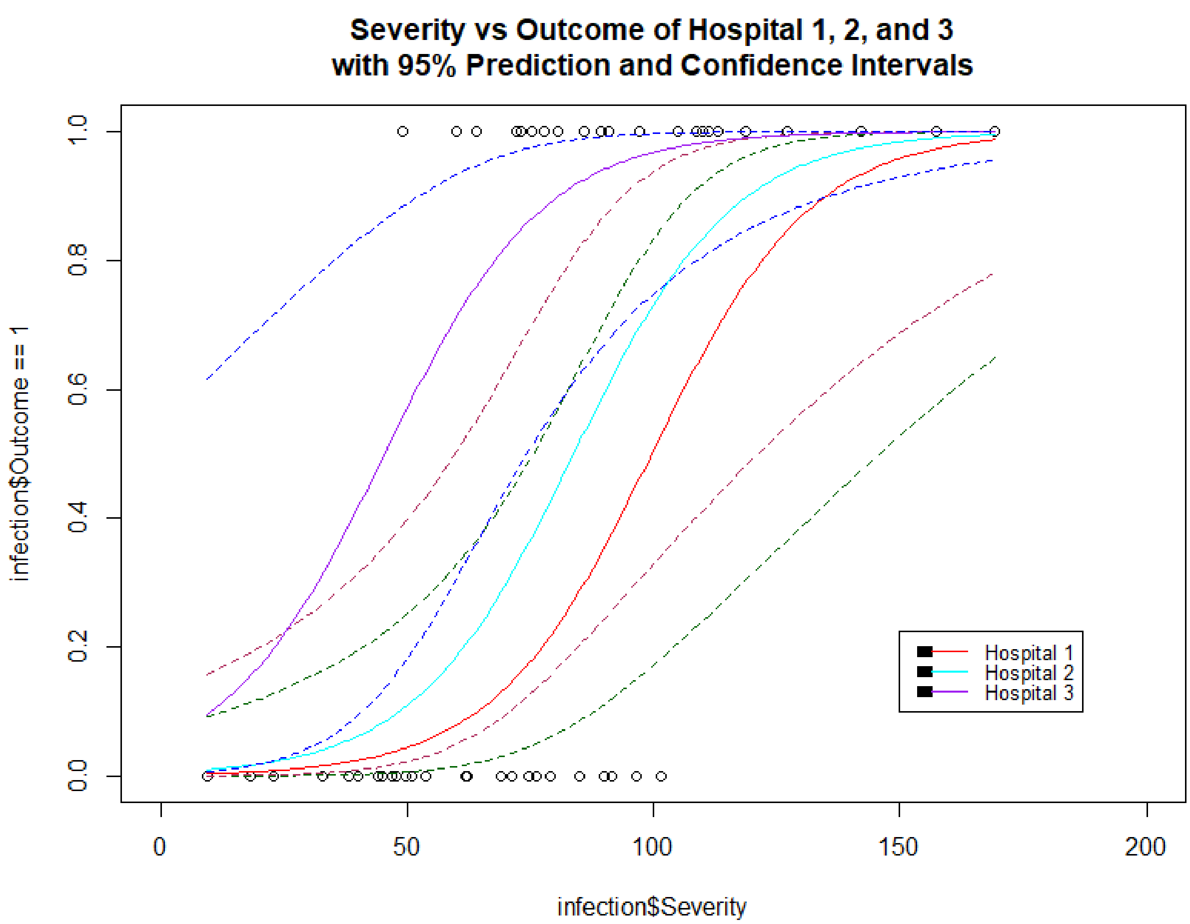
Here we are comparing the 95% confidence intervals for the positive treatment outcome for each hospital. We want to see if one hospital is better/worse than others. We obtained the following data between the three hospitals:

```
Hospital 1: [1] 0.05 0.51
Hospital 2: [1] 0.12 0.73
Hospital 3: [1] 0.50 0.95
```

We can see Hospital 1 is the best because its chance of having a bad outcome is 51% which is the lowest out of the three hospitals.

**4C)**

We then created a graph that has combined the 95% Confidence Intervals of each hospital.



Larce Blake Project 3