HW5

8.1 - One example I can think of is a regression model to predict the price of homes. The attributes that can be useful are 1) no. of bedrooms, 2) no. of bathrooms 3) sq. ft. 4) average median income of the household in that neighborhood 5) ZIP code 6) amount of yard space in terms of sq. ft. 7) no. of times it has been sold in the last 5 years 8) Crime rate per thousand 9) school ratings (elementary, middle and high) - each could be an attribute.

Another example that is suited for Regression is in the healthcare industry to predict the cardiovascular disease level/indicator. The attributes that can be used are 1) Age 2) BMI 3) High blood pressure 4) Low blood pressure 5) A1C level 6) Good cholesterol level 7) Bad cholesterol level 8) Triglyceride level 9) heartbeats per minute 10) Sex (0 or 1 for M and F). The cardiovascular disease level can be a numerical value calculated based on the values of the above mentioned attributes.

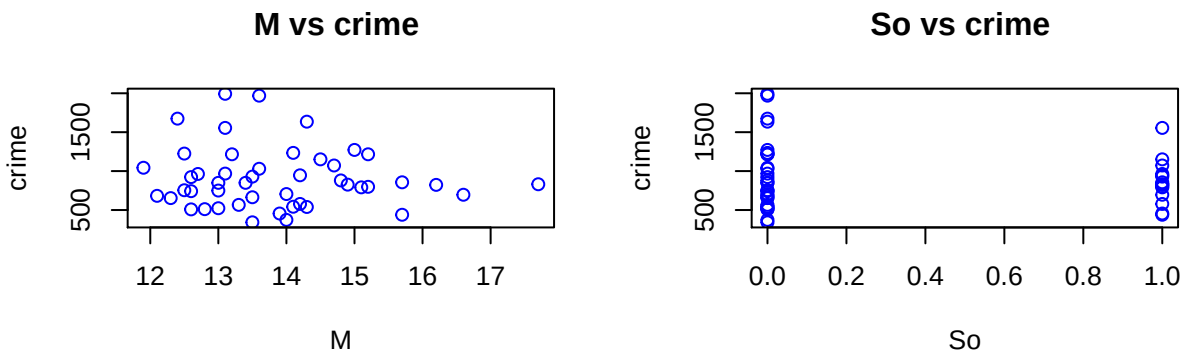
```
library("cluster")
```

```
## Warning: package 'cluster' was built under R version 3.4.4
```

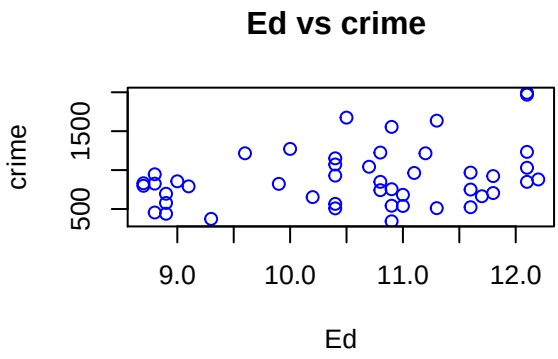
```
crime_data = read.table("uscrime.txt", header=TRUE)


test_data = crime_data[48,1:15]
crime_data = crime_data[1:47,]
```

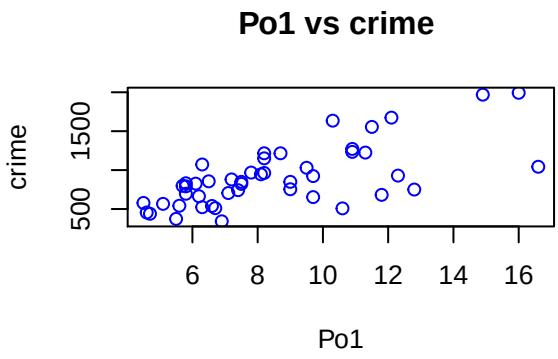
8.2 - First, I plotted each independent variable against the dependent variable to conclude that Regression would indeed be a good tool to analyze the data. However, I do have mention that some variables such as So - Southern State is just a binary variable hence not a perfect candidate for Regression.

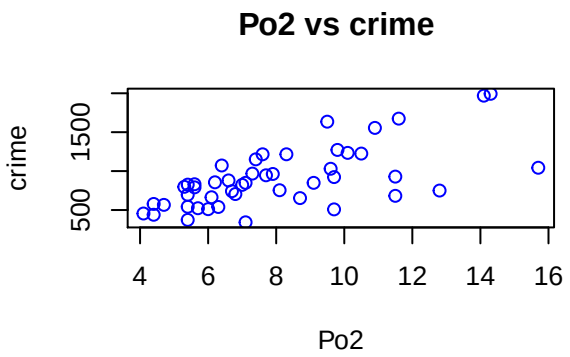
```
par(mfrow=c(2,2))
for (i in colnames(crime_data)[1:15])
{
  m1 = min(crime_data[,i])
  m2 = max(crime_data[,i])
  plot(crime_data[,i], crime_data[,'Crime'], main=paste(i, "vs crime "), xlab=i, ylab="crime", col="blu
}
```

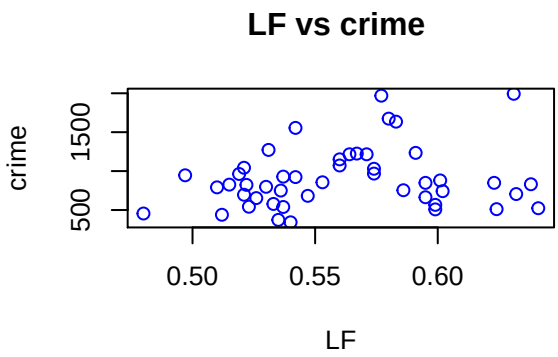

M vs crime
crime
500
1500
12
13
14
15
16
17
M

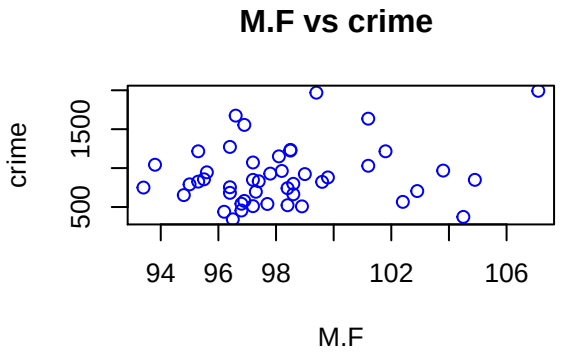
So vs crime
crime
500
1500
0.0
0.2
0.4
0.6
0.8
1.0
So

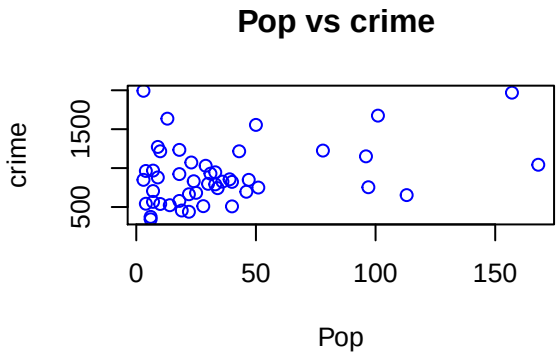

Ed vs crime
crime
500
1500
9.0
10.0
11.0
12.0
Ed

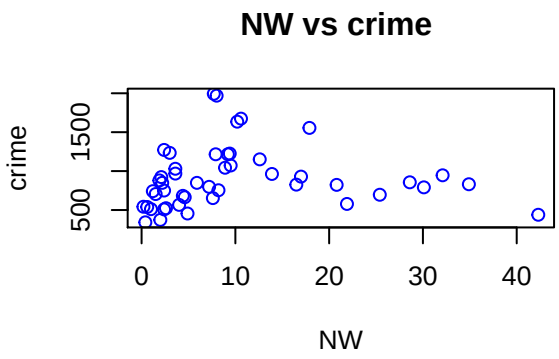

Po1 vs crime
crime
500
1500
6
8
10
12
14
16
Po1

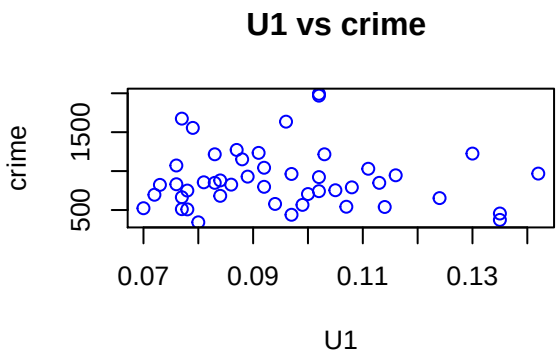

Po2 vs crime
crime
500
1500
4
6
8
10
12
14
16
Po2

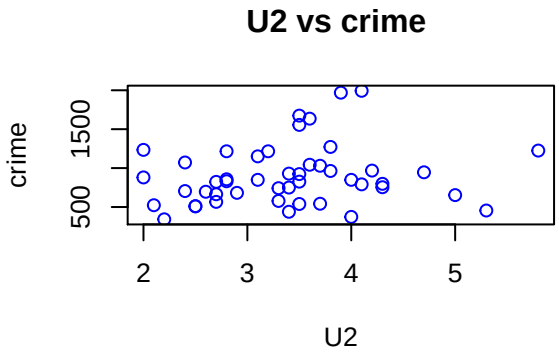

LF vs crime
crime
500
1500
0.50
0.55
0.60
LF

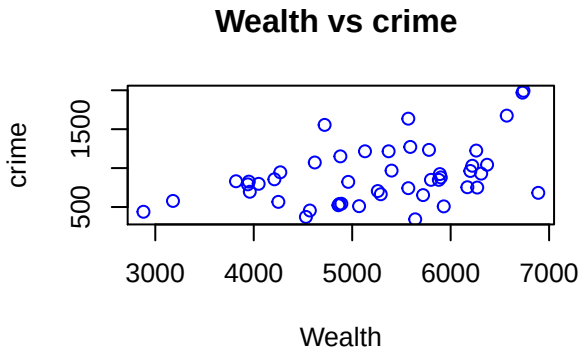

M.F vs crime
crime
500
1500
94
96
98
102
106
M.F

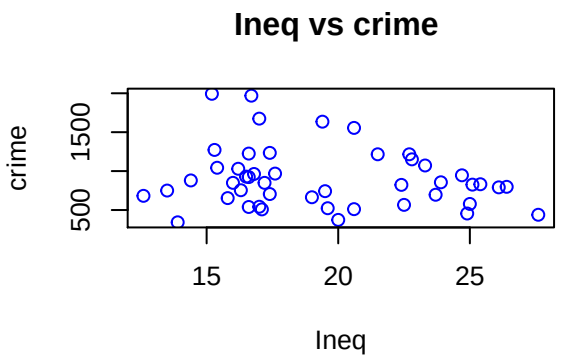

Pop vs crime
crime
500
1500
0
50
100
150
Pop

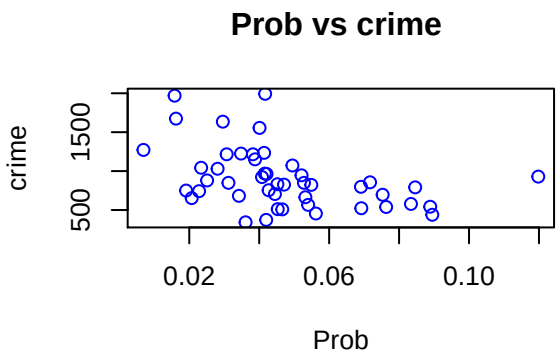

NW vs crime
crime
500
1500
0
10
20
30
40
NW

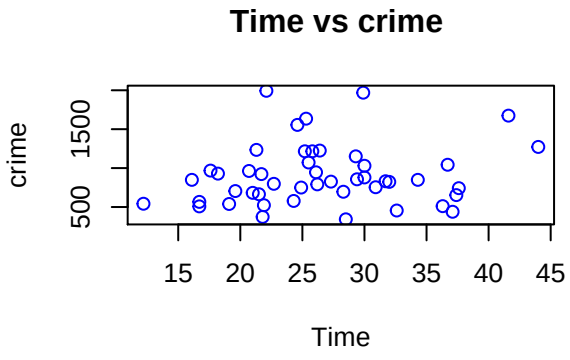

U1 vs crime
crime
500
1500
0.07
0.09
0.11
0.13
U1


U2 vs crime
crime
500
1500
2
3
4
5
U2


Wealth vs crime
crime
500
1500
3000
4000
5000
6000
7000
Wealth


Ineq vs crime
crime
500
1500
15
20
25
Ineq


Prob vs crime
crime
500
1500
0.02
0.06
0.10
Prob


Time vs crime
crime
500
1500
15
20
25
30
35
40
45
Time

Next, I plotted the histogram of the dependent variable to see how the responses are distributed - the idea is to observe if they follow a normal distribution. As shown below the dependent variable doesn't resemble a normal distribution perfectly, but most of the values are indeed centered around the mean, so we can proceed with Linear Regression.

I used the following factors to evaluate the model: 1) SSE/Residual SE 2) AIC and 3) BIC scores 4) Adjusted R-Squared

SSE score is 1,354,946 and the Residual SE is 209.1. The AIC and BIC scores are 650.02 and 681.48. Adjusted R2 0.8078 - this means the model can explain 80% of the variability in the data, which is good.

In the Residuals plot the red line representing the mean of the residuals is not a straight line which tells me that the data is not perfectly suited for linear regression.

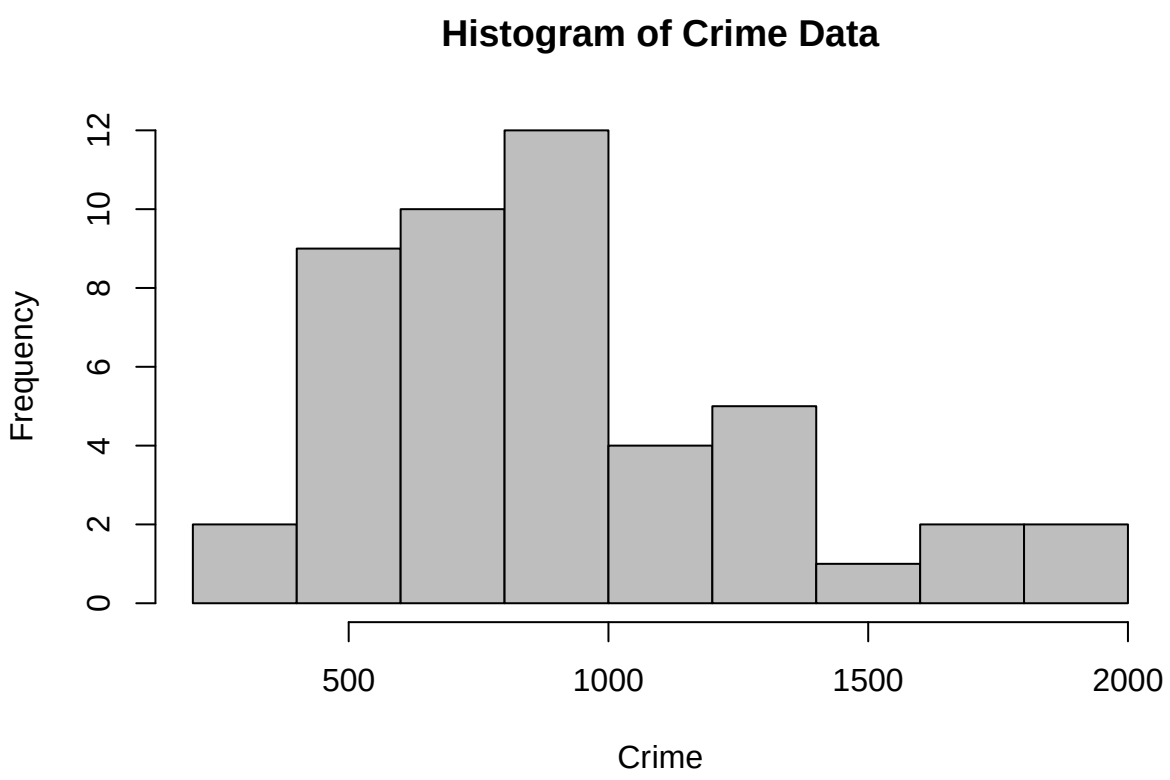
In the Q-Q plot the residuals from the model are almost perfectly in line with the theoretical results from a perfect model, thus passing the normality test.

Next, I looked at the P values of each attribute to see which attributes do make the most impact and which ones don't. I looked at each attribute and retained the attributes that had a P-value of less than 0.05. I then fit a model for this revised data to find out the attributes mentioned above performed far worse. For example, the predicted value was 897.2, whereas the AIC and BIC scores were 690.06 and 701.16 worse than the previous model. In addition, the Residual SE was 347.5 worse than for the previous model. Also, the adjusted R squared is only 0.1927.

Therefore, the conclusion would be to use the first model.

Lastly, I just wanted to see if transforming the dependent variable assuming an Exponential and Quadratic relationship with the independent variables would be a better fit. However, as shown below the predictions obtained don't seem to be accurate. Hence I will use the model fit using the provided data/attributes.

```
hist(crime_data$Crime, main="Histogram of Crime Data", col="Grey",xlab="Crime")
```




```
mod = lm(Crime ~., data=crime_data)
vec = as.vector(t(coefficients(mod)))

inp = c(1,14,0,10.0,12.0,15.5,0.640,94.0,150,1.1,0.120,3.6,3200,20.1,0.04,39)

pred = sum(inp * vec)
cat("The predicted crime rate using the formula is", pred, "\n")
```

```
## The predicted crime rate using the formula is 155.4349
```

```
## Predict using the predict function
cat("The prediction using the predict function", predict(mod, test_data))
```

```
## The prediction using the predict function 155.4349
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = Crime ~ ., data = crime_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -395.74  -98.09   -6.69  112.99  512.67
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.984e+03  1.628e+03  -3.675 0.000893 ***
## M            8.783e+01  4.171e+01   2.106 0.043443 *
## So          -3.803e+00  1.488e+02  -0.026 0.979765
## Ed           1.883e+02  6.209e+01   3.033 0.004861 **
## Po1          1.928e+02  1.061e+02   1.817 0.078892 .
## Po2         -1.094e+02  1.175e+02  -0.931 0.358830
## LF          -6.638e+02  1.470e+03  -0.452 0.654654
## M.F          1.741e+01  2.035e+01   0.855 0.398995
## Pop         -7.330e-01  1.290e+00  -0.568 0.573845
## NW           4.204e+00  6.481e+00   0.649 0.521279
## U1          -5.827e+03  4.210e+03  -1.384 0.176238
## U2           1.678e+02  8.234e+01   2.038 0.050161 .
## Wealth       9.617e-02  1.037e-01   0.928 0.360754
## Ineq         7.067e+01  2.272e+01   3.111 0.003983 **
## Prob        -4.855e+03  2.272e+03  -2.137 0.040627 *
## Time        -3.479e+00  7.165e+00  -0.486 0.630708
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209.1 on 31 degrees of freedom
## Multiple R-squared:  0.8031, Adjusted R-squared:  0.7078
## F-statistic: 8.429 on 15 and 31 DF,  p-value: 3.539e-07
```

```
cat("The SSE for this model is", sum(mod$residuals**2), "\n")
```

```
## The SSE for this model is 1354946
```

```
cat("The Residual standard error is", sqrt(sum(mod$residuals**2)/mod$df.residual), "\n")
```

```
## The Residual standard error is 209.0644
```

```
cat("The confidence interval of the model is \n")
```

```
## The confidence interval of the model is
```

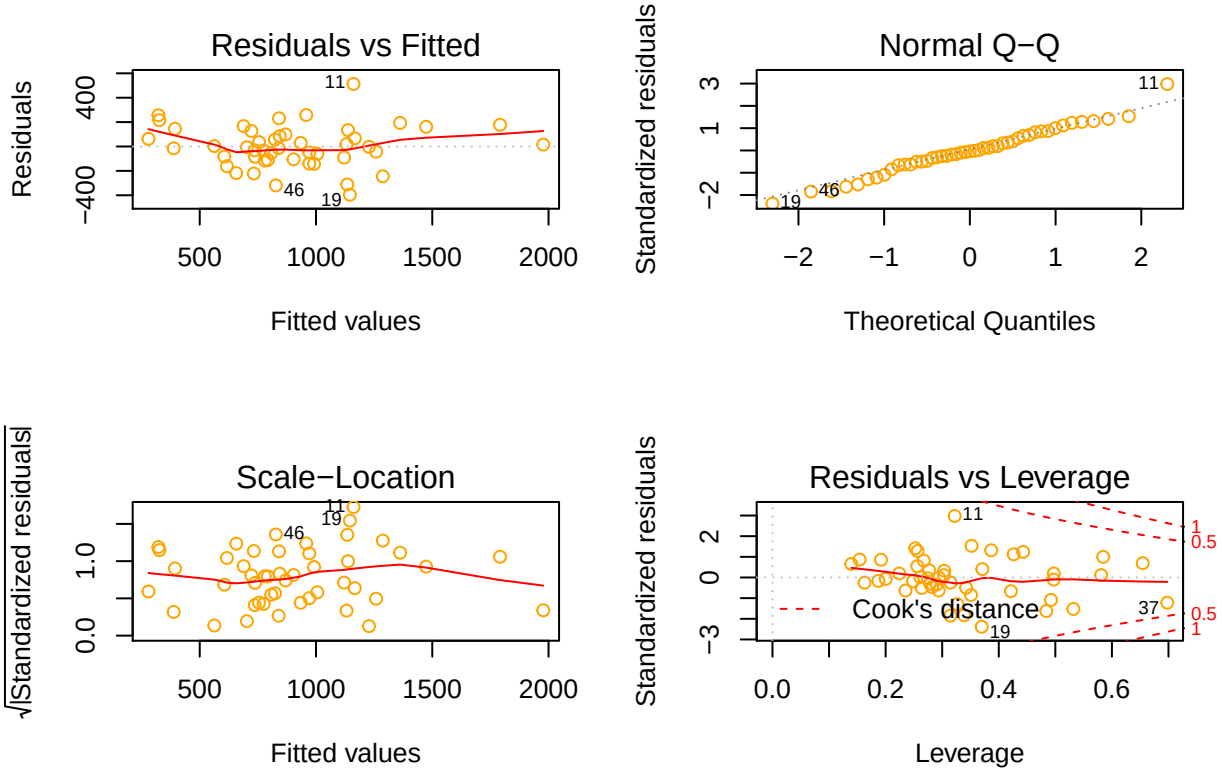
```
confint(mod)
```

```
##                     2.5 %        97.5 %
## (Intercept) -9.305265e+03 -2663.3103872
## M            2.754182e+00   172.9061646
## So          -3.071916e+02   299.5846578
## Ed           6.169424e+01   314.9543927
## Po1         -2.360777e+01   409.2164486
## Po2         -3.490189e+02   130.1750880
## LF          -3.661358e+03  2333.7055475
## M.F         -2.410508e+01    58.9187914
## Pop         -3.363074e+00     1.8970574
## NW          -9.013406e+00    17.4223277
```

```
## U1          -1.441404e+04  2759.8383788
## U2          -1.256155e-01   335.7249600
## Wealth      -1.152621e-01     0.3075945
## Ineq         2.434145e+01   117.0027500
## Prob        -9.489804e+03  -220.7272203
## Time        -1.809269e+01    11.1346572
```

```
cat("\n")
```

```
par(mfrow=c(2,2))
plot(mod, col="orange")
```




```
## Akaic
cat("The AKAIC score is ", AIC(mod), "\n")
```

```
## The AKAIC score is  650.0291
```

```
## BIC
cat("The BIC score is ", BIC(mod), "\n")
```

```
## The BIC score is  681.4816
```

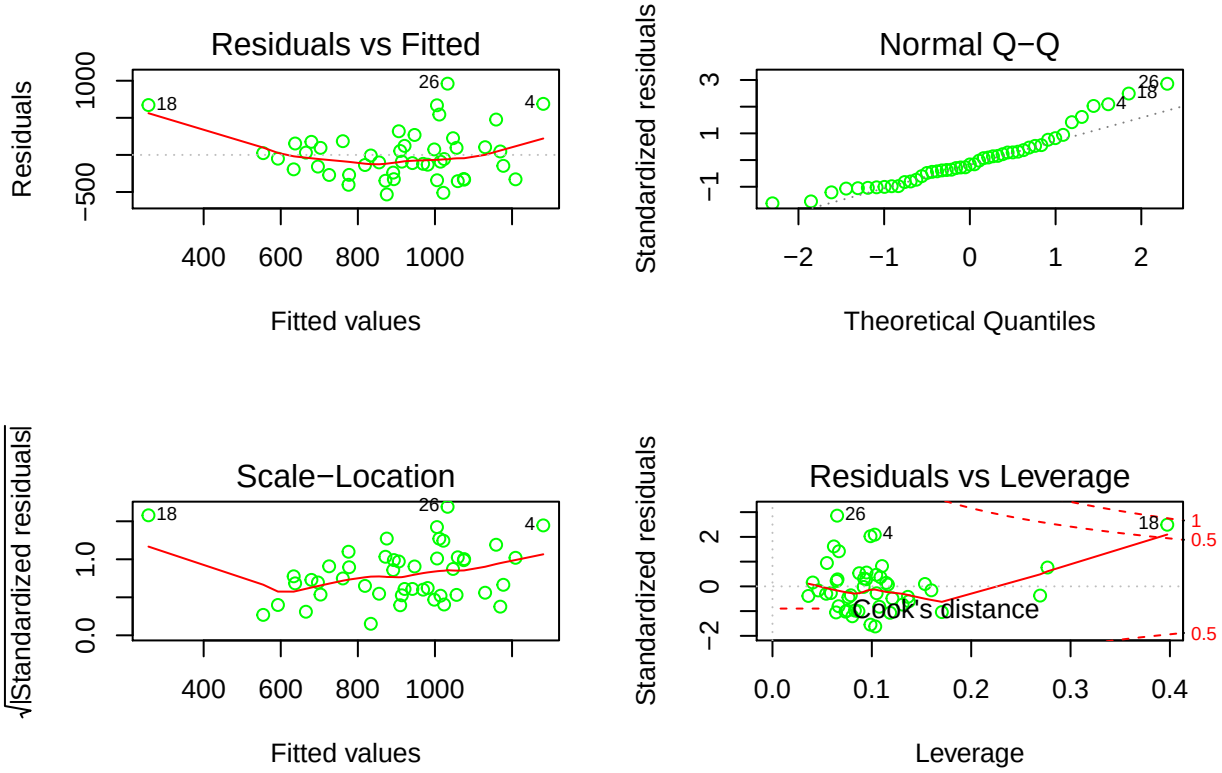
```
#########################################################################################################################
#########################################################################################################################
p = summary(mod)$coefficients[, "Pr(>|t|)"]

p = p[p <0.05]
pnames = names(p)
pnames = c(pnames, "Crime")

pnames = pnames[c(2:length(pnames))]

new_data = crime_data[,pnames]

## Fit for this new data
mod2 = lm(Crime ~., data=new_data)
vec = as.vector(t(coefficients(mod2)))
summary(mod2)
```

```
##
## Call:
## lm(formula = Crime ~ ., data = new_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -532.97 -254.03  -55.72  137.80  960.21
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1339.35    1247.01  -1.074  0.28893
## M              35.97      53.39   0.674  0.50417
## Ed            148.61      71.92   2.066  0.04499 *
## Ineq           26.87      22.77   1.180  0.24458
## Prob        -7331.92    2560.27  -2.864  0.00651 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 347.5 on 42 degrees of freedom
## Multiple R-squared:  0.2629, Adjusted R-squared:  0.1927
## F-statistic: 3.745 on 4 and 42 DF,  p-value: 0.01077
```

```
inp = c(1,14,10.0,20.1,0.04)
pred = sum(inp * vec)
cat("The predicted crime rate is", pred, "\n")
```

```
## The predicted crime rate is 897.2307
```

```
par(mfrow=c(2,2))
plot(mod2, col="green")
```




```
## Akaic
cat("The AIC score is ", AIC(mod2), "\n")
```

```
## The AIC score is  690.0666
```

```
## BIC
cat("The BIC score is ", BIC(mod2), "\n")
```

```
## The BIC score is  701.1675
```

```
##############################################################################################################
##Sensitivity Analysis
mod3 = lm(sqrt(Crime) ~., data=crime_data)
vec = as.vector(t(coefficients(mod3)))

inp = c(1,14,0,10.0,12.0,15.5,0.640,94.0,150,1.1,0.120,3.6,3200,20.1,0.04,39)

pred = sum(inp * vec)
cat("The predicted crime rate - Quadratic model is", pred, "\n")
```

```
## The predicted crime rate - Quadratic model is 17.43855
```

```r
mod3 = lm(log(Crime) ~., data=crime_data)
vec = as.vector(t(coefficients(mod3)))

inp = c(1,14,0,10.0,12.0,15.5,0.640,94.0,150,1.1,0.120,3.6,3200,20.1,0.04,39)

pred = sum(inp * vec)
cat("The predicted crime rate - Exponential model is", pred, "\n")
```

```
## The predicted crime rate - Exponential model is 5.897035
```