Assignment 5

Question 8.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a linear regression model would be appropriate. List some (up to 5) predictors that you might use.

Answer 8.1

I am presently working on a problem where I am trying to predict future production of an oil well using a bunch of predictors. The response variable here is the oil rate and some of the predictors are as follows:

1. Casing Pressure
2. Reservoir Pressure
3. Reservoir Temperature
4. Welbore radius
5. Liquid Ratios

Question 8.2

Using crime data from http://www.statsci.org/data/general/uscrime.txt
(file uscrime.txt, description at http://www.statsci.org/data/general/ uscrime.html), use regression (a useful R function is lm or glm) to predict the observed crime rate in a city with the following data:

M = 14.0 So = 0 Ed = 10.0 Po1 = 12.0 Po2 = 15.5 LF = 0.640 M.F = 94.0 Pop = 150 NW = 1.1 U1 = 0.120 U2 = 3.6 Wealth = 3200 Ineq = 20.1 Prob = 0.04 Time = 39.0

Show your model (factors used and their coefficients), the software output, and the quality of fit.

Answer 8.2

I have tried different models to analyze the data. Since the dataset is small I did not split the data into testing and training data. So, I trained the full dataset.

Also, I investigated the presence of outliers in the dataset and assumed that the data is correct and thus did not further investigate in removing them.

The models are also evaluated to make sure that the assumptions of the regression model holds true such as: 1. Linearity 2. Normality and homoscedasticity 3. No endogeneity with the predictors (meaning that the predictors have no correlation with error) 4. No multi-collinearity between predictors

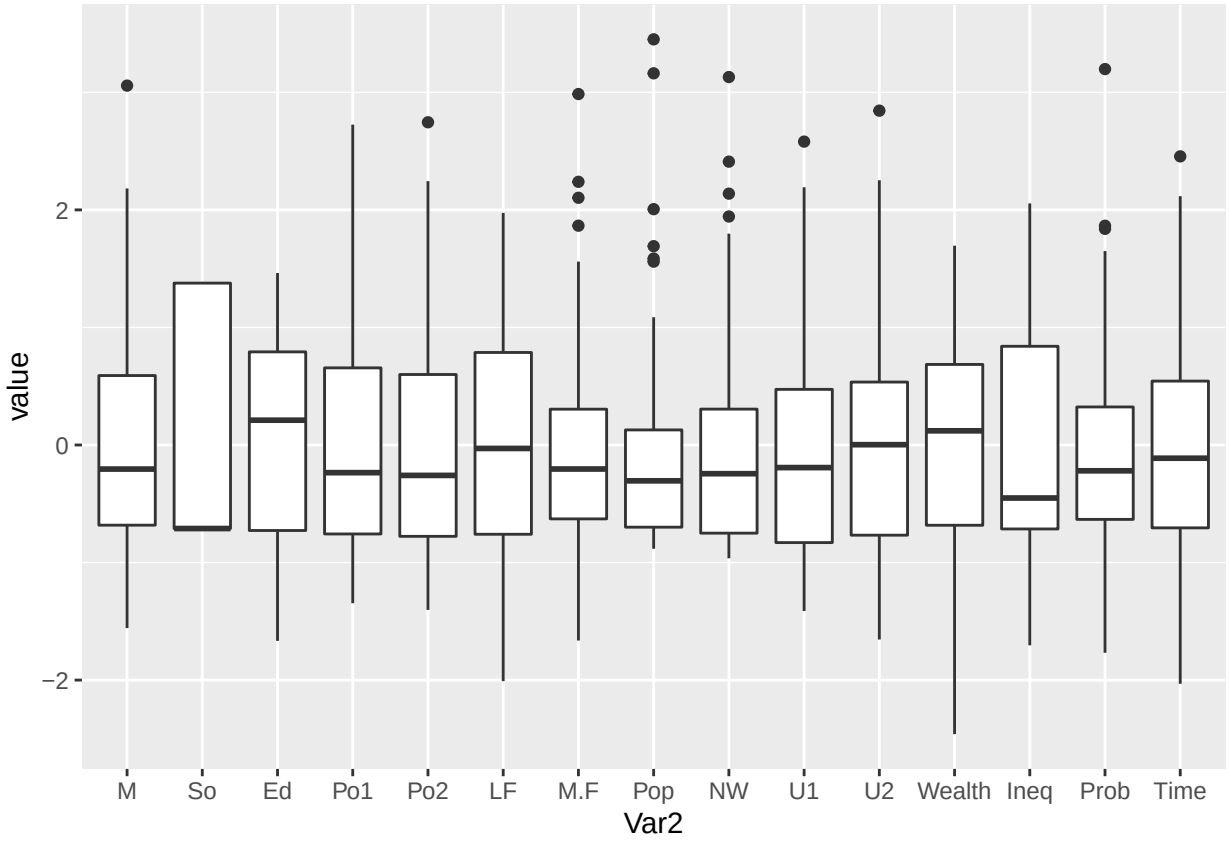
Model 1

```
library(ggplot2)
library(outliers)
library(caret)
library(reshape2)
library(MASS)
```

```
#Getting the training and test data
crime_data<-read.table('uscrime.txt', sep = "", header = TRUE )
test<-read.table('test.txt', sep="", header = TRUE)

max(crime_data$Crime)
```

```
## [1] 1993
```

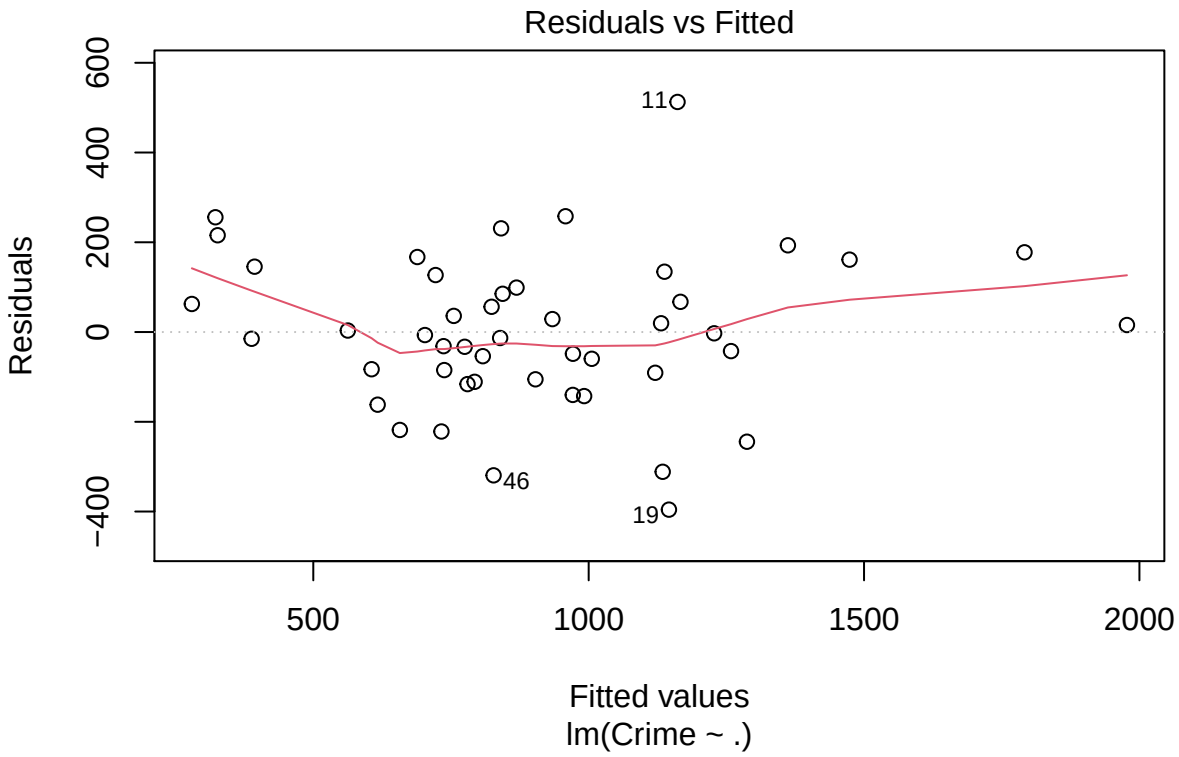
```
#Exploratory data analysis and pre-processing

#We can explore the data for outlier using the strategies we learned earlier.

#Scaling the data to fit the box plot in the plot
crime_scale<-scale(crime_data[,1:15])
out<-melt(crime_scale)

#Box plot to view the outliers
ggplot(out,aes(x=Var2, y=value))+geom_boxplot()
```

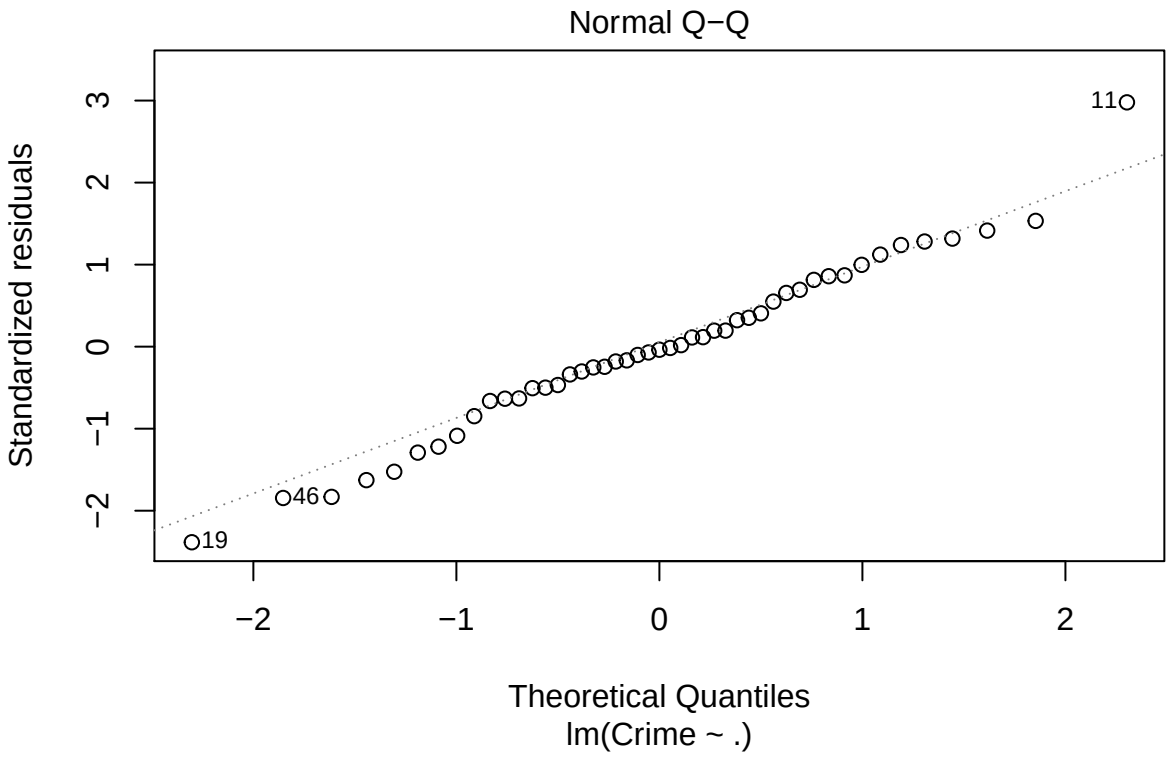


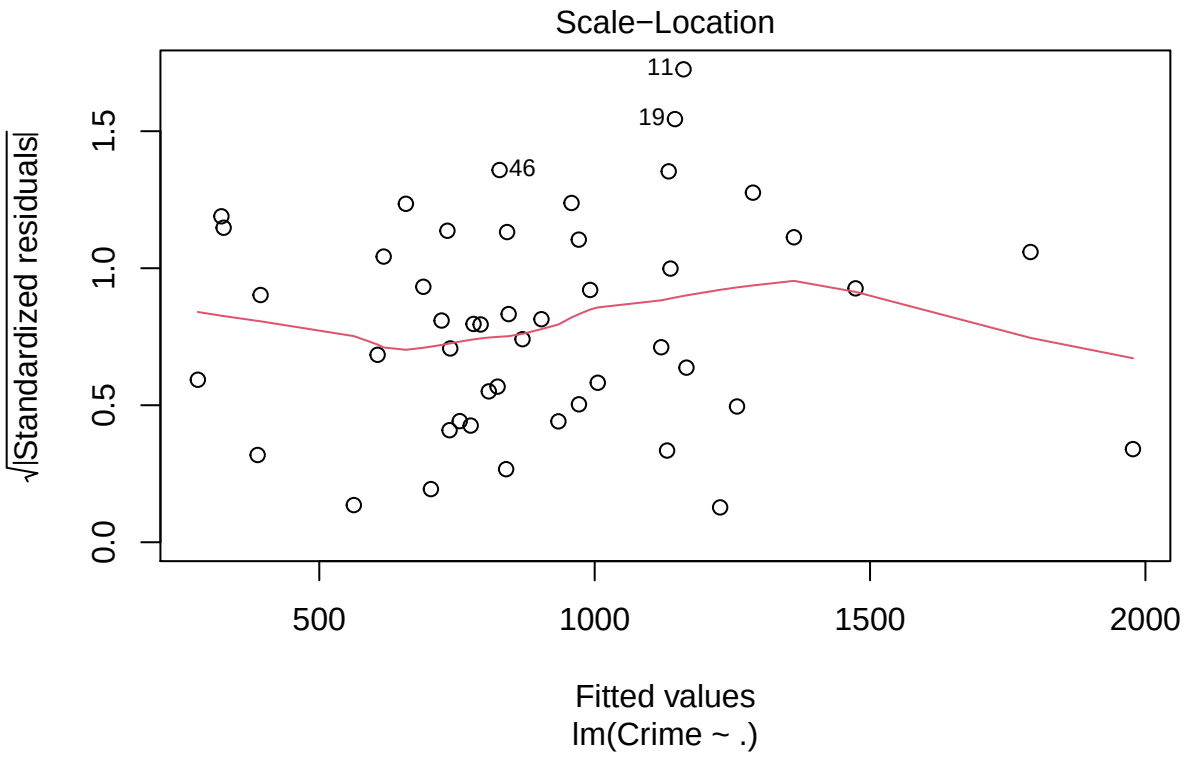
```r
#Some of the predictors seems to have outliers. We cannot just simply remove
#them. Thus, for this exercise we will keep the original dataset without
#further investigating them

#Linear regression model
model1<-lm(Crime~.,data = crime_data)

summary(model1) # Model summary
```

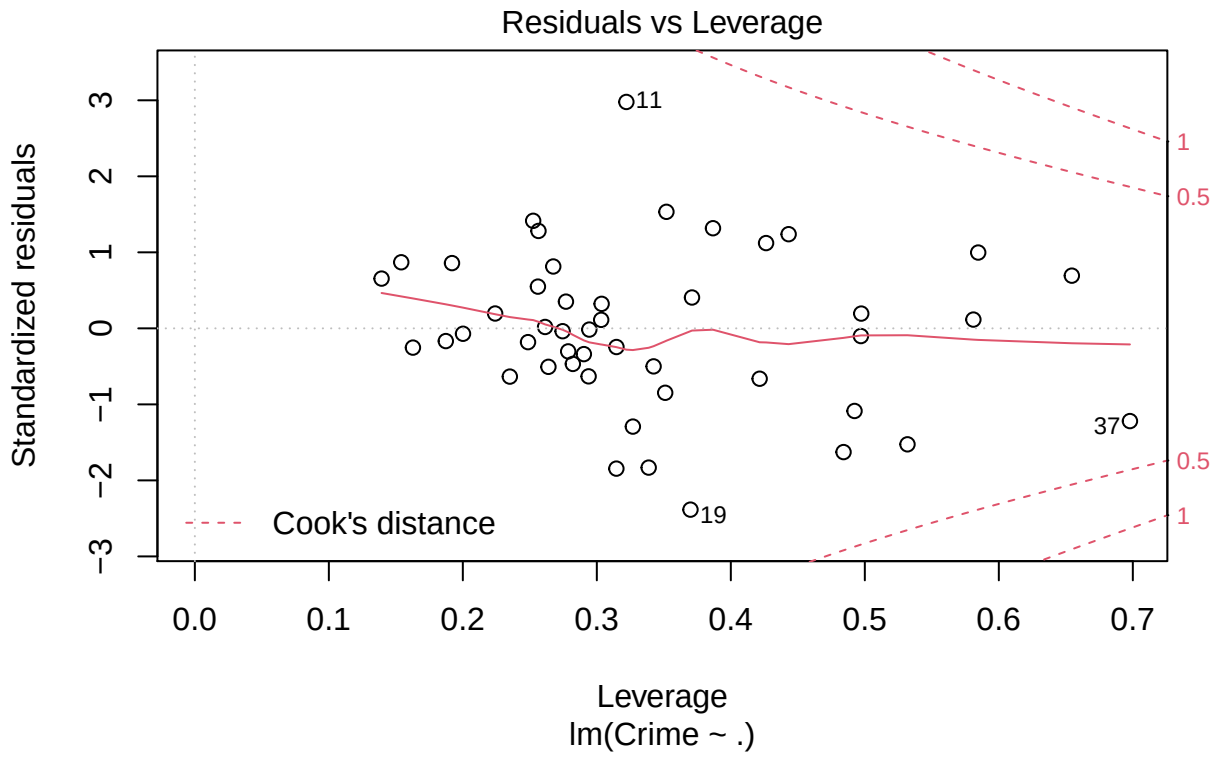
```
##
## Call:
## lm(formula = Crime ~ ., data = crime_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -395.74  -98.09   -6.69  112.99  512.67
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.984e+03  1.628e+03  -3.675 0.000893 ***
## M            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
```

```r
pred1<-predict(model1, test) #Prediction on test data

#Plotting model results
plot(model1)
```


Residuals vs Fitted
Residuals
600
400
200
0
−400
11
46
19
500
1000
1500
2000
Fitted values
lm(Crime ~ .)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Crime ~ .)
11
46
19
−2
−1
0
1
2
3


Scale−Location
11
19
46
√|Standardized residuals|
0.0
0.5
1.0
1.5
500
1000
1500
2000
Fitted values
lm(Crime ~ .)

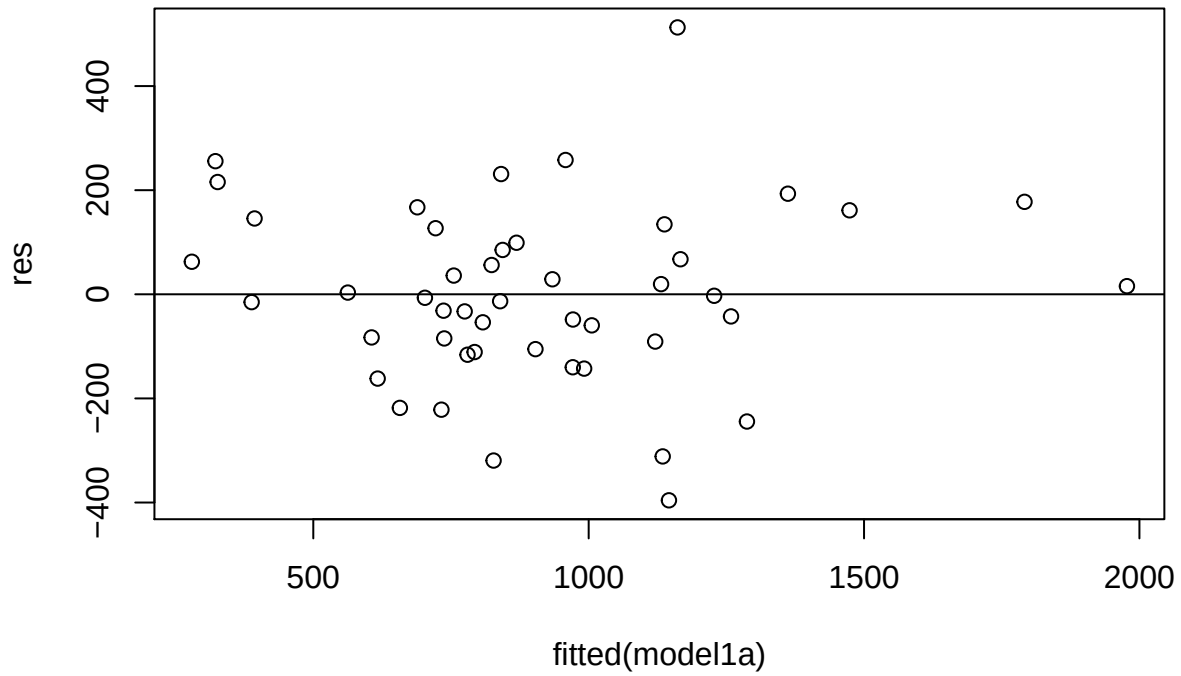



```r
#Linear regression model 1a using cross validation
set.seed(123)
train.control <- trainControl(method = "cv", number = 10)
model1a<-train(Crime~.,data = crime_data, method='lm',  #Linear Model with CV
               trControl=train.control)

summary(model1a) # Model summary
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## 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
```

```r
pred1a<-predict(model1a, test) #Prediction on test data

#Plotting residual vs fitted data to make sure there is no trend in the errors
res<-resid(model1a)
plot(fitted(model1a), res)
abline(0,0) #Adding a horizontal line
```




```r
#The residual plot shows no pattern confirming that errors have no explanatory
#power for the above model and this is what we want that the explanatory power
#of the model resides with the predictors
```

Conclusions: Model 1

Now evaluating the results above. Let's first look into the model summary. The overall efficiency of the model is judged using the adjusted R^2 value which is around 70%, which is good. But the model prediction on the test data gives crime rate as 155 which almost half of the min value of crime rate in the dataset. To investigated this let's look at the model summary plots:

1. Residual vs Fitted: The linearity assumption dose not look good
2. Q-Q Plot: The normality assumption here again is not great specially for earlier values.
3. Scale-Location:To show homoscedasticity the trend line should be straight,but we can see it's somewhat straight

To investigate the next steps the model predictors are assessed for their significance using the p-values. A new model can then be build using only the predictors with strong significance. We will use p-value of 0.1 and any predictor with higher than this will be removed from the new model.

For regression the null hypothesis is that the coefficient and the intercept are not significant. Now for a predictor to be significant we need to reject the null hypothesis by using the p-value.We will use p-value of 0.1 and any predictor with higher than this will be removed from the new model. So using this criteria the following predictors(M,Ed,U2, Ineq, Prob, Po1) are significant using the p-value of Model 1

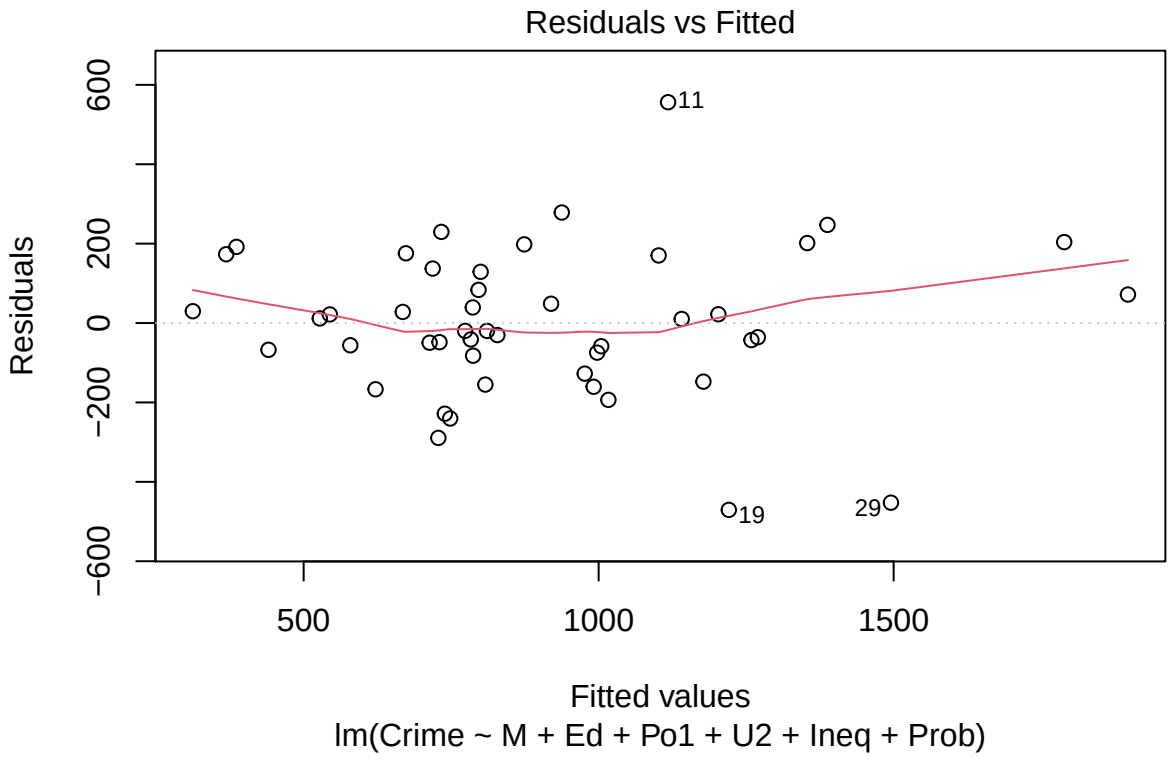
Model 2

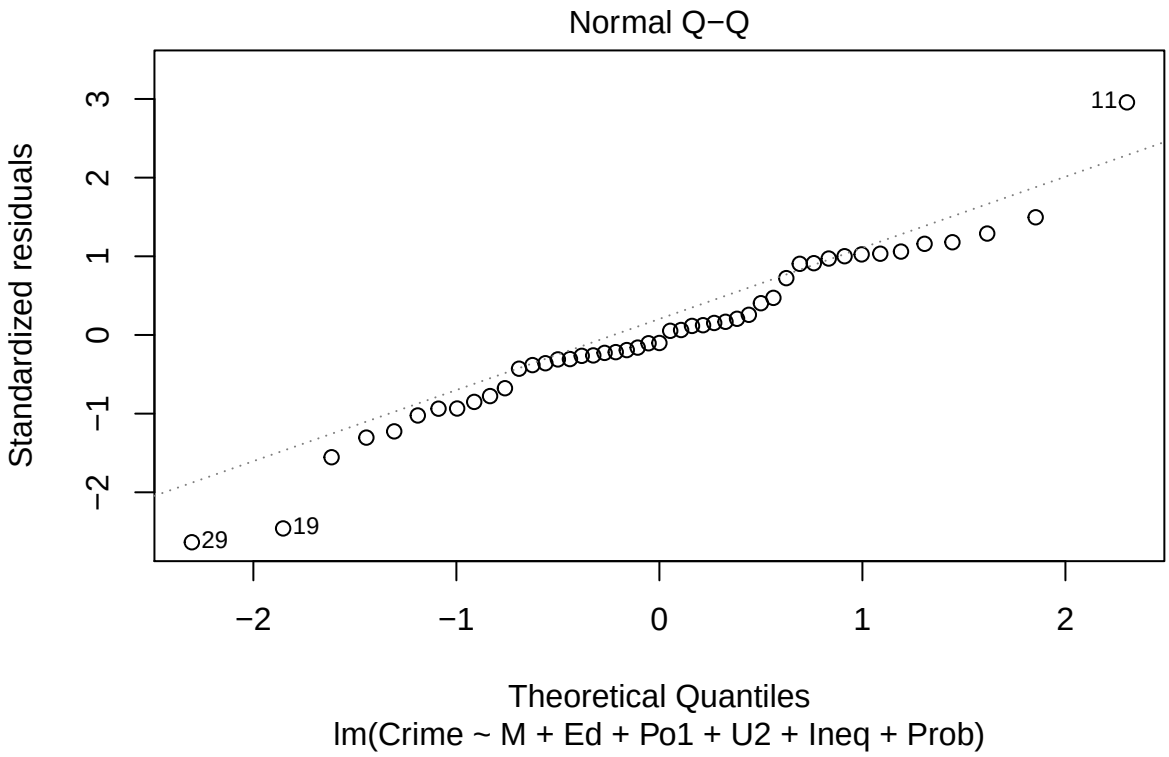
Models with less predictors

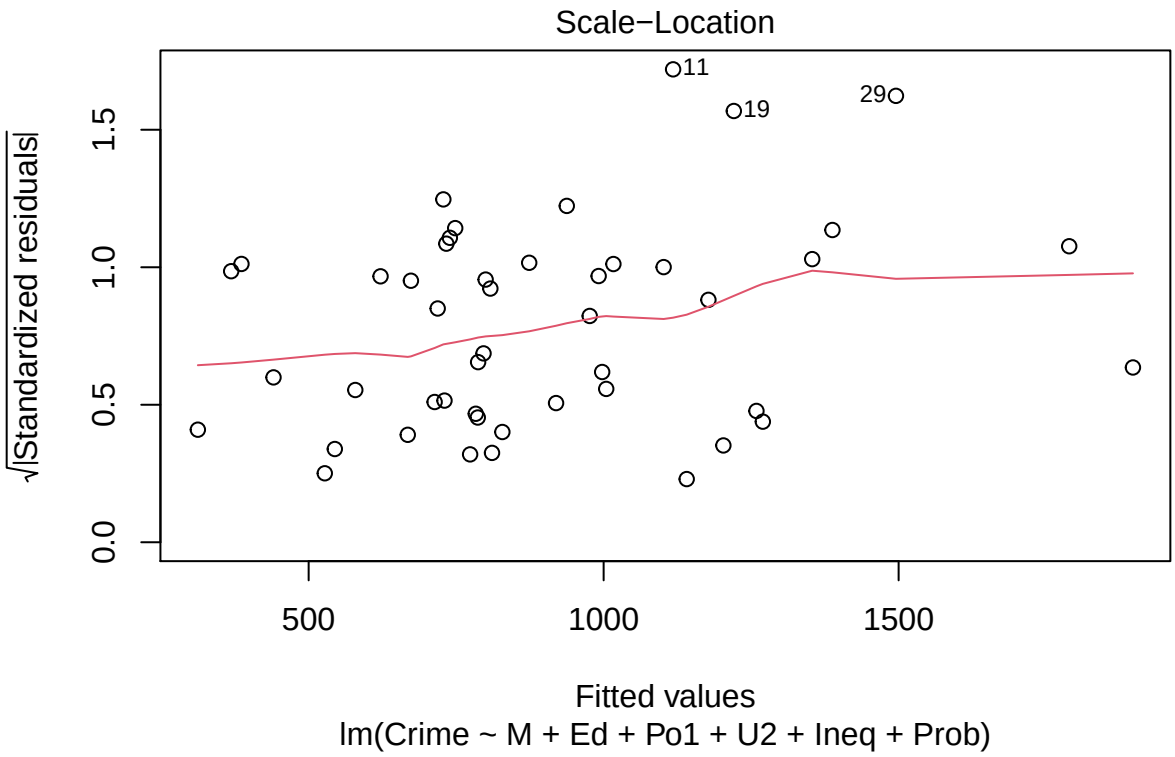
```
#Linear regression model 2
model2<-lm(Crime~M+Ed+Po1+U2+Ineq+Prob,data = crime_data)

summary(model2) # Model summary
```

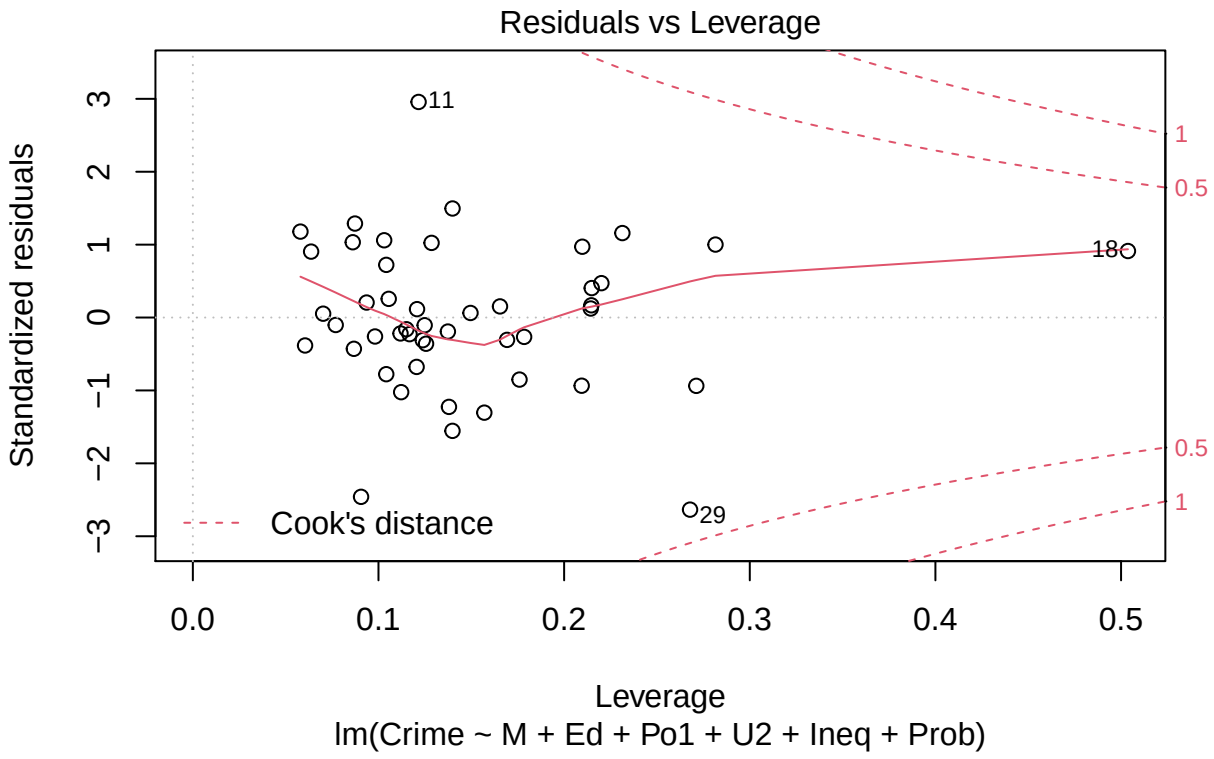
```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = crime_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -470.68  -78.41  -19.68  133.12  556.23
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5040.50     899.84  -5.602 1.72e-06 ***
## M             105.02      33.30   3.154  0.00305 **
## Ed            196.47      44.75   4.390 8.07e-05 ***
## Po1           115.02      13.75   8.363 2.56e-10 ***
## U2             89.37      40.91   2.185  0.03483 *
## Ineq           67.65      13.94   4.855 1.88e-05 ***
## Prob        -3801.84    1528.10  -2.488  0.01711 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 200.7 on 40 degrees of freedom
## Multiple R-squared:  0.7659, Adjusted R-squared:  0.7307
## F-statistic: 21.81 on 6 and 40 DF,  p-value: 3.418e-11
```

```r
pred2<-predict(model2, test) #Prediction on test data

#Plotting model results
plot(model2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Crime ~ M + Ed + Po1 + U2 + Ineq + Prob)
11
19
29


Scale−Location
11
19
29
√|Standardized residuals|
0.0
0.5
1.0
1.5
500
1000
1500
Fitted values
lm(Crime ~ M + Ed + Po1 + U2 + Ineq + Prob)

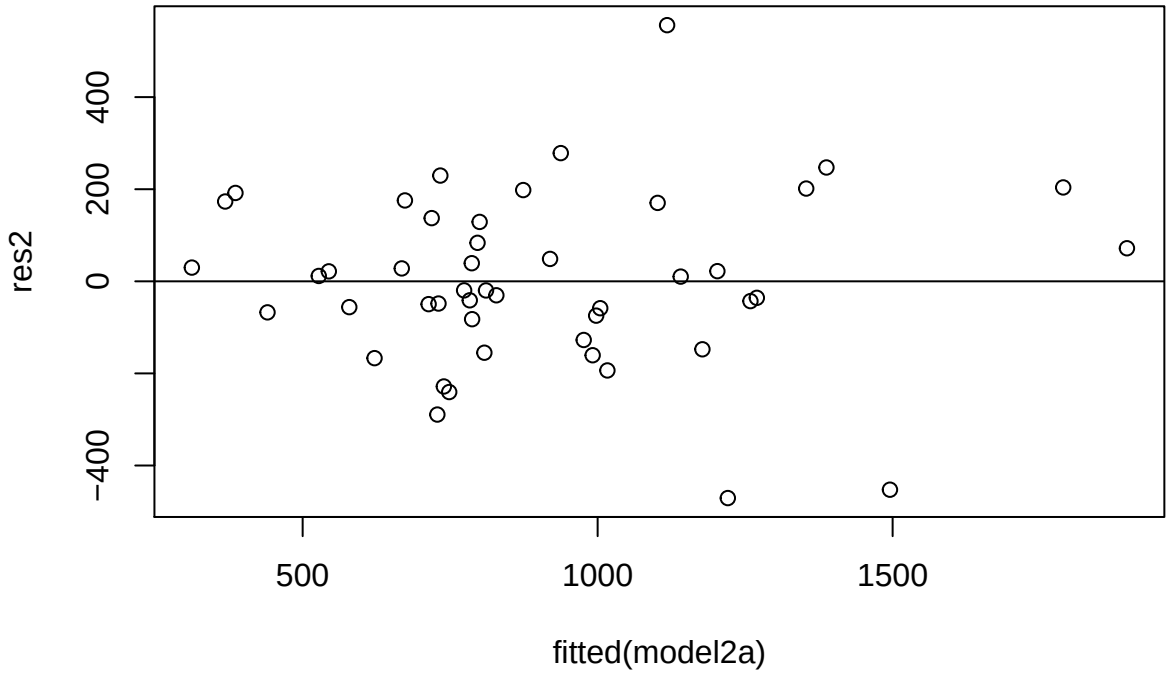



```
#Linear regression model 2 using cross validation
set.seed(123)
train2.control <- trainControl(method = "cv", number = 10)
model2a<-train(Crime~M+Ed+Po1+U2+Ineq+Prob,
               data = crime_data, method='lm',  #Linear Model2 with CV
               trControl=train.control)

summary(model2a) # Model summary
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -470.68  -78.41  -19.68  133.12  556.23
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5040.50     899.84  -5.602 1.72e-06 ***
## M             105.02      33.30   3.154  0.00305 **
## Ed            196.47      44.75   4.390 8.07e-05 ***
## Po1           115.02      13.75   8.363 2.56e-10 ***
## U2             89.37      40.91   2.185  0.03483 *
## Ineq           67.65      13.94   4.855 1.88e-05 ***
## Prob        -3801.84    1528.10  -2.488  0.01711 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 200.7 on 40 degrees of freedom
## Multiple R-squared:  0.7659, Adjusted R-squared:  0.7307
## F-statistic: 21.81 on 6 and 40 DF,  p-value: 3.418e-11
```

```
pred2a<-predict(model2a, test) #Prediction on test data

#Plotting residual vs fitted data to make sure there is no trend in the errors
res2<-resid(model2a)
plot(fitted(model2a), res2)
abline(0,0) #Adding a horizontal line
```




```
#The residual plot shows no pattern confirming that errors have no explanatory
#power for the above model and this is what we want that the explanatory power
#of the model resides with the predictors
```

Conclusion: Model 2

Now evaluating Model 2 shows that all the predictors are significant in terms of the p-value. Also, the model fit shows adjusted R^2 to be 73% better than the previous model.

Investigating the model summary plots:

1. Residual vs Fitted: The linearity assumption looks better than model 1
2. Q-Q Plot: The normality assumption here again is not great specially for earlier and later data.
3. Scale-Location:To show homoscedasticity the trend line should be straight,but we can see it's somewhat straight

Model 2 prediction on test data gives crime rate as 1304 which looks more realistic than predicted by model 1

Since for the above models the assumptions were not great we can try transformation technique

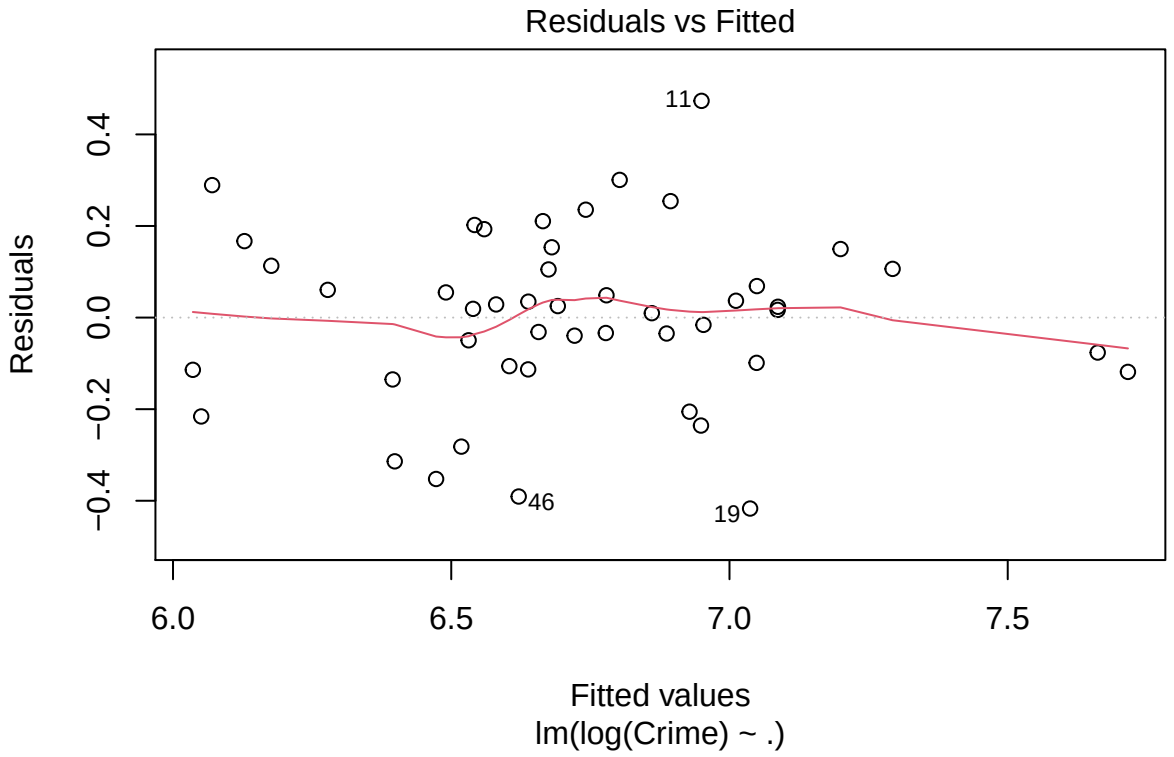
Model 3

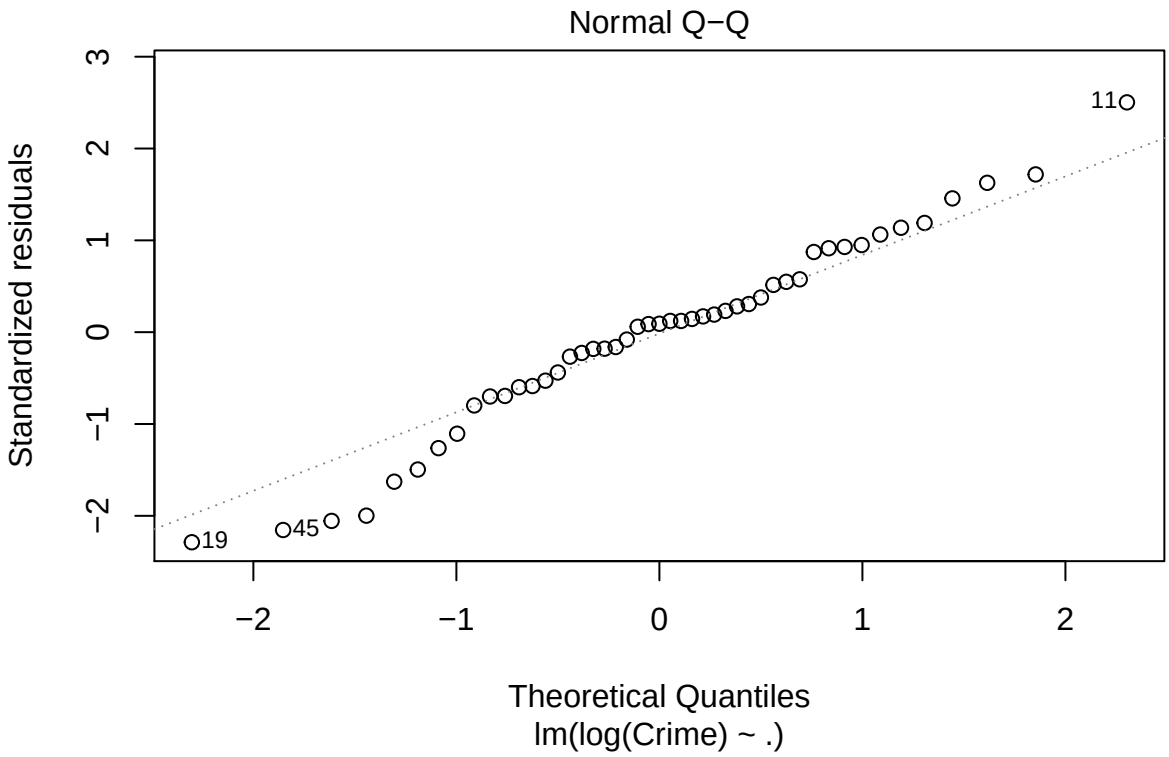
Model transformation using log transformation on response data

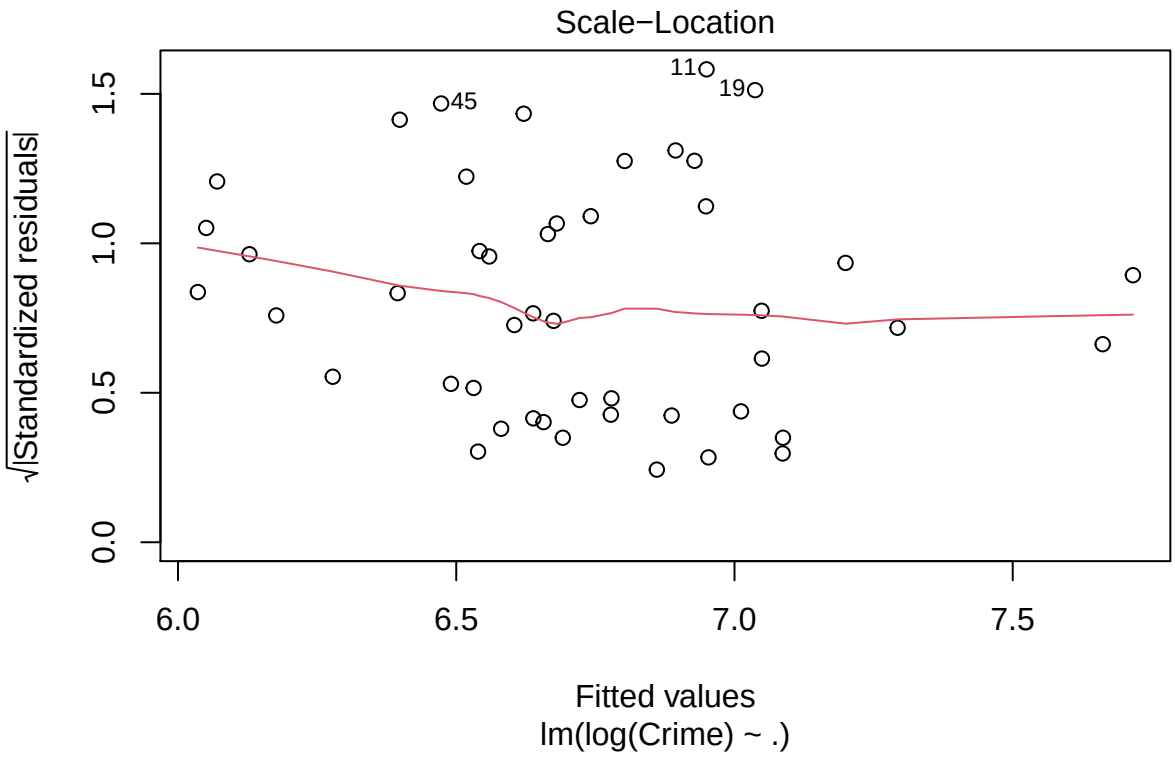
```
#Linear regression model 3a on full dataset using model1
model1_trans<-lm(log(Crime)~.,data = crime_data)

summary(model1_trans)
```

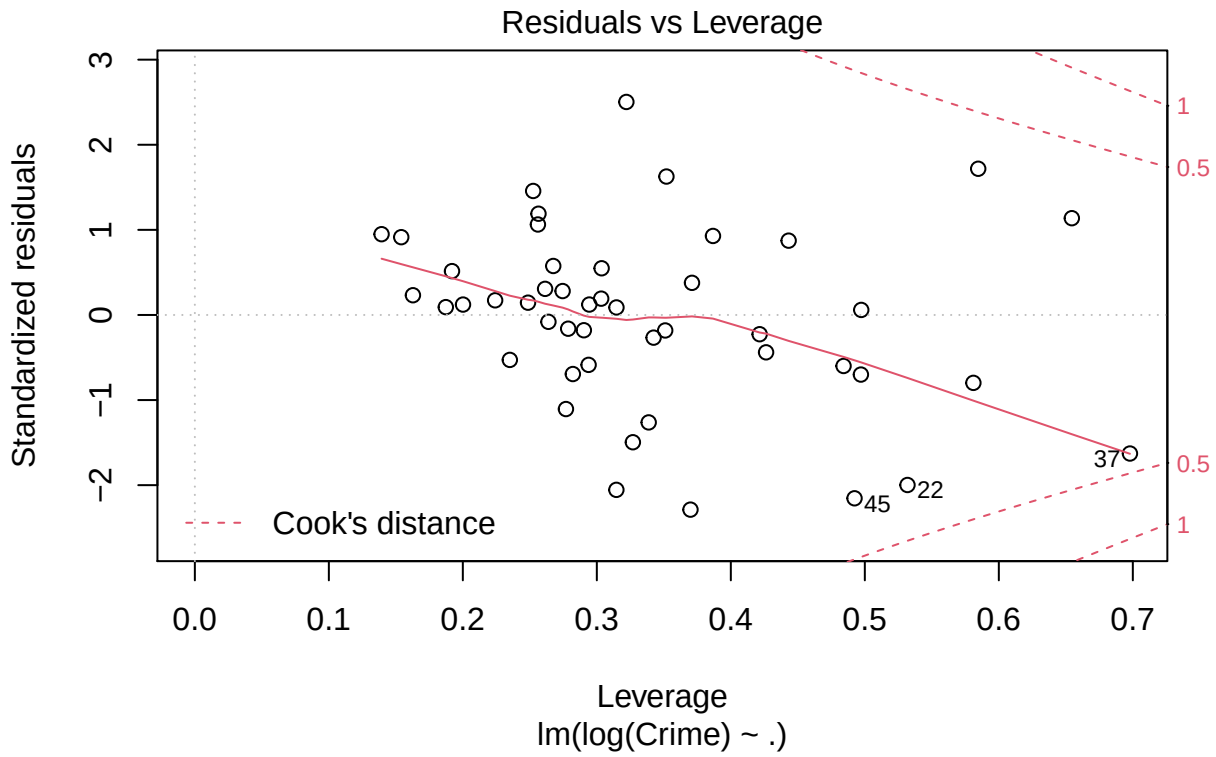
```
##
## Call:
## lm(formula = log(Crime) ~ ., data = crime_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41697 -0.10961  0.01903  0.10971  0.47322
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4345908  1.7884057   0.243  0.80960
## M            0.1160700  0.0458149   2.533  0.01657 *
## So           0.0917576  0.1633799   0.562  0.57841
## Ed           0.2147289  0.0681926   3.149  0.00361 **
## Po1          0.1862712  0.1165418   1.598  0.12012
## Po2         -0.1077276  0.1290273  -0.835  0.41015
## LF           0.1874034  1.6142245   0.116  0.90833
## M.F         -0.0061943  0.0223549  -0.277  0.78355
## Pop         -0.0009638  0.0014163  -0.681  0.50124
## NW           0.0047976  0.0071181   0.674  0.50530
## U1          -4.3033459  4.6242214  -0.931  0.35925
## U2           0.1717980  0.0904308   1.900  0.06680 .
## Wealth       0.0001737  0.0001139   1.526  0.13715
## Ineq         0.0808730  0.0249499   3.241  0.00284 **
## Prob        -6.0950555  2.4957820  -2.442  0.02050 *
## Time        -0.0080381  0.0078697  -1.021  0.31497
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2296 on 31 degrees of freedom
## Multiple R-squared:  0.7897, Adjusted R-squared:  0.688
## F-statistic: 7.761 on 15 and 31 DF,  p-value: 8.862e-07
```

```r
pre_trans<-predict(model1_trans, test)

#Converting prediction to the same scale
final<-exp(pre_trans)

#Plotting model results
plot(model1_trans)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(Crime) ~ .)
11
45
19
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
11
19
45
6.0
6.5
7.0
7.5
Fitted values
lm(log(Crime) ~ .)

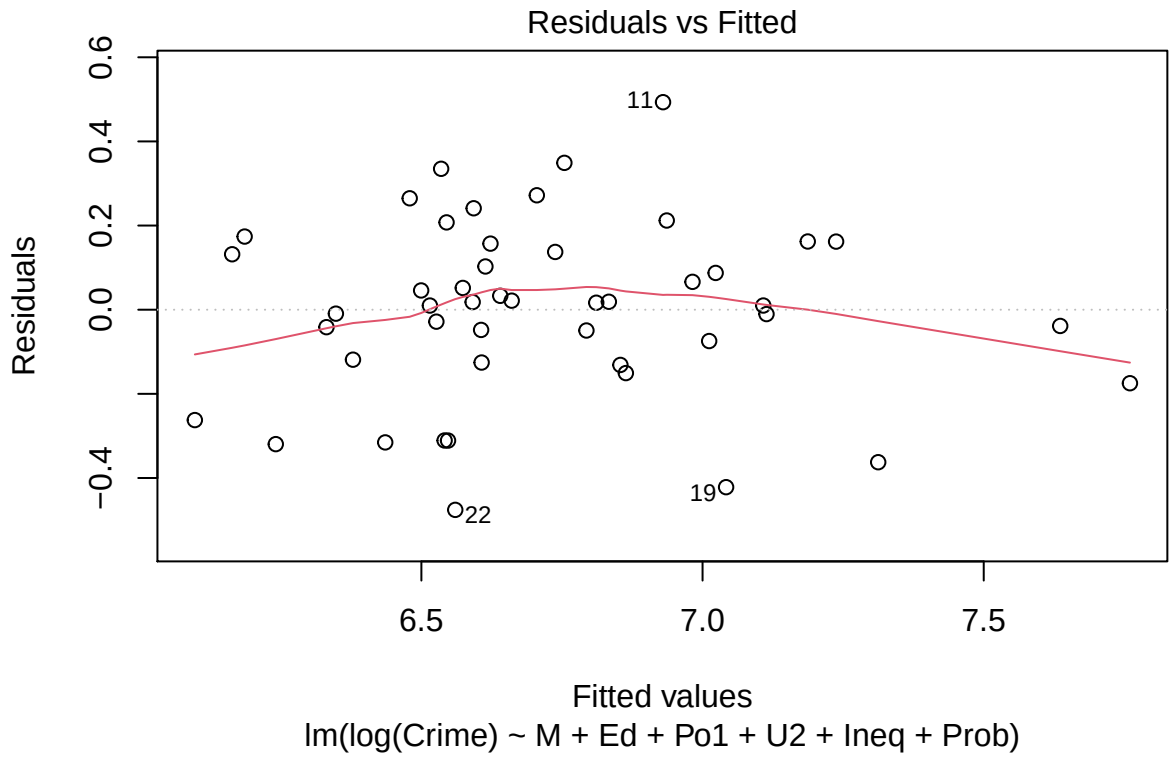


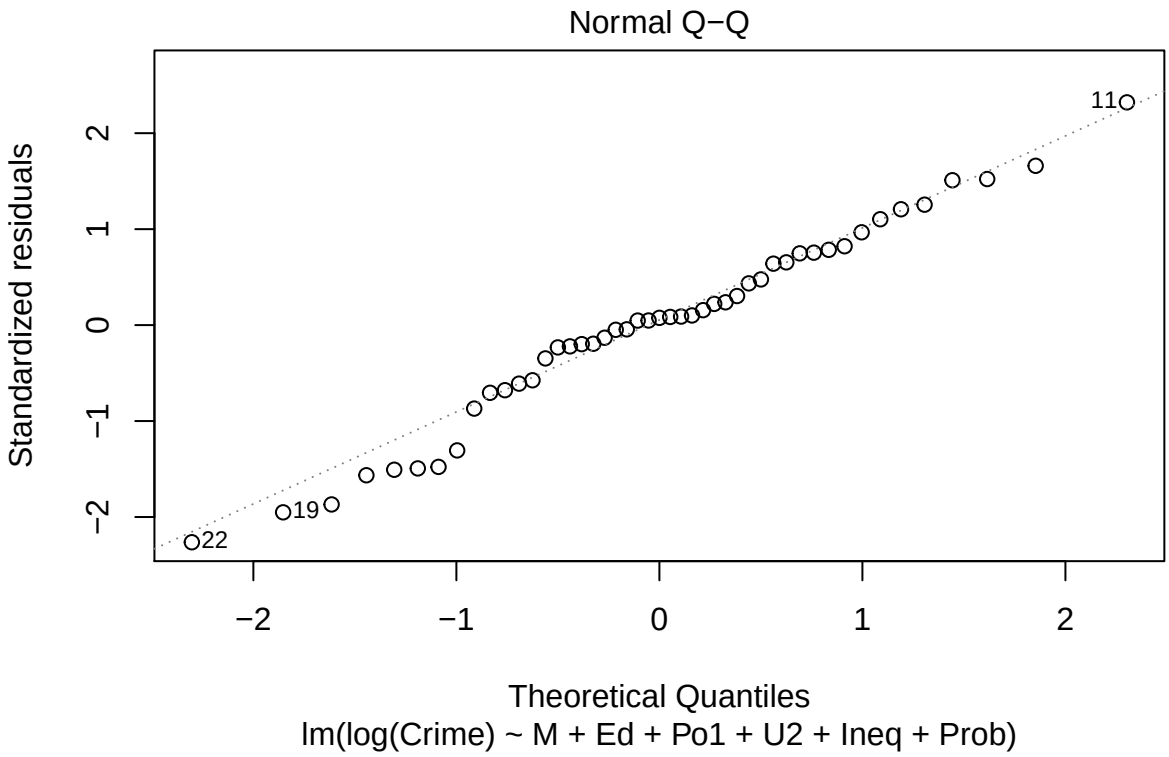

```
#Linear regression model 3b on smaller dataset
modela_trans<-lm(log(Crime)~M+Ed+Po1+U2+Ineq+Prob,data = crime_data)

summary(modela_trans) # Model summary
```

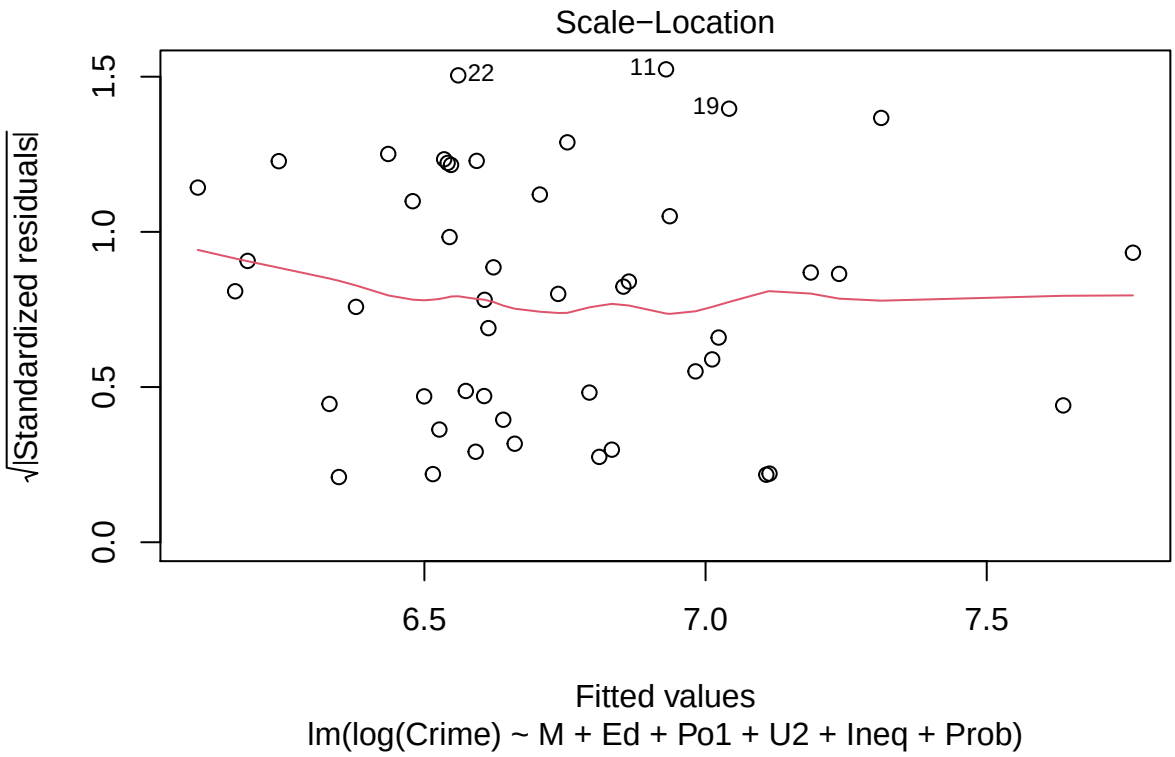
```
##
## Call:
## lm(formula = log(Crime) ~ M + Ed + Po1 + U2 + Ineq + Prob, data = crime_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47574 -0.12217  0.01661  0.14716  0.49322
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.31516    1.01640   0.310 0.758110
## M            0.11927    0.03761   3.171 0.002914 **
## Ed           0.20987    0.05055   4.152 0.000168 ***
## Po1          0.11902    0.01554   7.661 2.29e-09 ***
## U2           0.09523    0.04620   2.061 0.045835 *
## Ineq         0.07206    0.01574   4.578 4.50e-05 ***
## Prob        -4.10406    1.72603  -2.378 0.022287 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2267 on 40 degrees of freedom
```

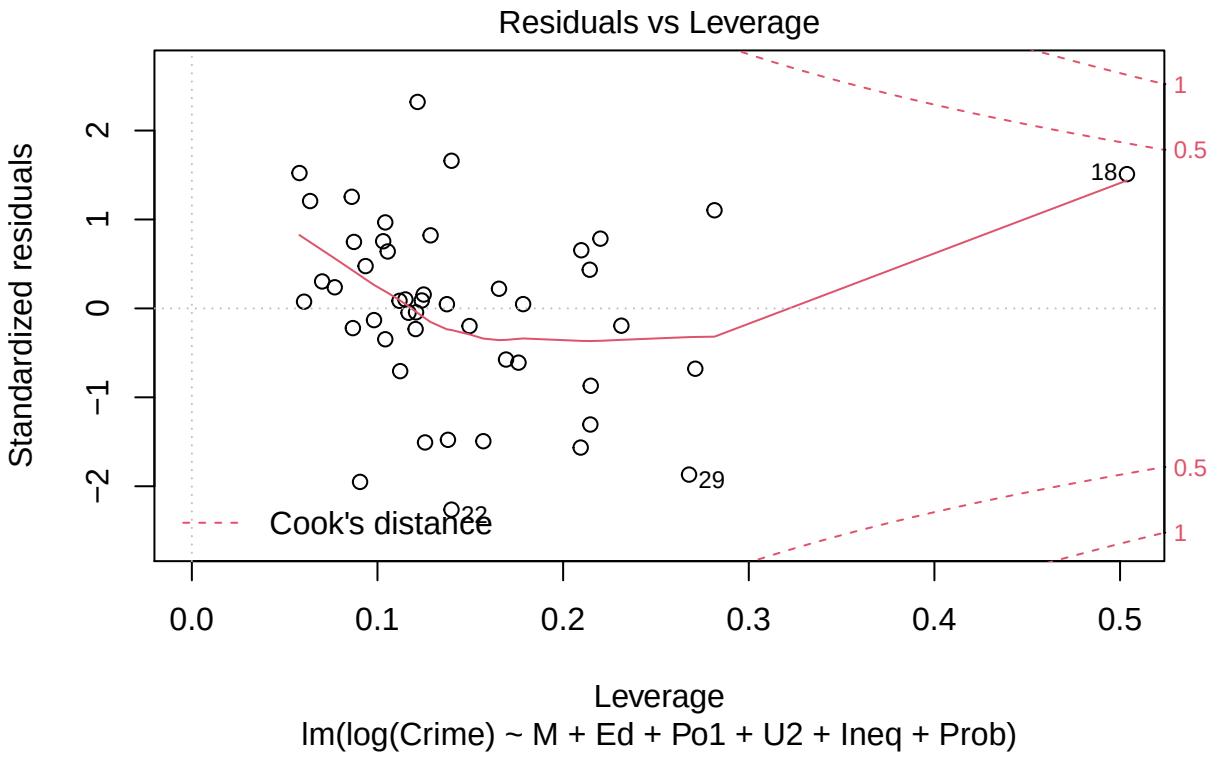
```
## Multiple R-squared:  0.7356, Adjusted R-squared:  0.6959
## F-statistic: 18.54 on 6 and 40 DF,  p-value: 3.618e-10
```

```
pred_trans1<-predict(modela_trans, test) #Prediction on test data

#Converting prediction to the same scale
final1<-exp(pred_trans1)

#Plotting model results
plot(modela_trans)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(Crime) ~ M + Ed + Po1 + U2 + Ineq + Prob)
11
19
22
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
22
11
19
6.5
7.0
7.5
Fitted values
lm(log(Crime) ~ M + Ed + Po1 + U2 + Ineq + Prob)



Conclusion: Model 3

Now evaluating Model 3

Investigating the model summary plots:

1. Residual vs Fitted: The linearity assumption looks best
2. Q-Q Plot: The normality assumption here again is best.
3. Scale-Location: homoscedasticity trend line is better,

Model 3a prediction on test data gives crime rate as 364 better than model 1. Model 3b prediction on test data gives crime rate as 1260 but its difficult to draw a comparison between model 2.