Homework 5

Anonymous

2024-02-13

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.

A situation that a regression model could be appropriate is when data scientists for professional sport teams, such as the NBA, use them to measure the effects of different training regimens on players performances.

A few predictors could be if players are engaging in yoga, weightlifting, HIIT, track & field, and aquatics. Yoga trains for flexibility and composure. Weightlifting improves strength when taking a charge or shooting the ball from a long distance. HIIT workouts focus on high intensity so, the player can last longer on the court in a shorter amount of time. Track & field improves endurance and power while doing long and high jump. Aquatics can influence a stronger heart and resistance training. Although, not all players benefit from each of the workouts depending on which position they play on the court. The data scientist will have to measure each workout, or multiple, for each player based on his position on the court.

Question 8.2

Using Crime data, 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.

Model 1

```r
library(ggplot2)
library(outliers)
library(caret)
```

```
## Loading required package: lattice
```

```r
library(reshape2)
library(MASS)

rm(list = ls())
crime_data <- read.table(file= "C:\\Users\\sheya\\OneDrive\\Desktop\\uscrime.txt",
                         header = TRUE)
```

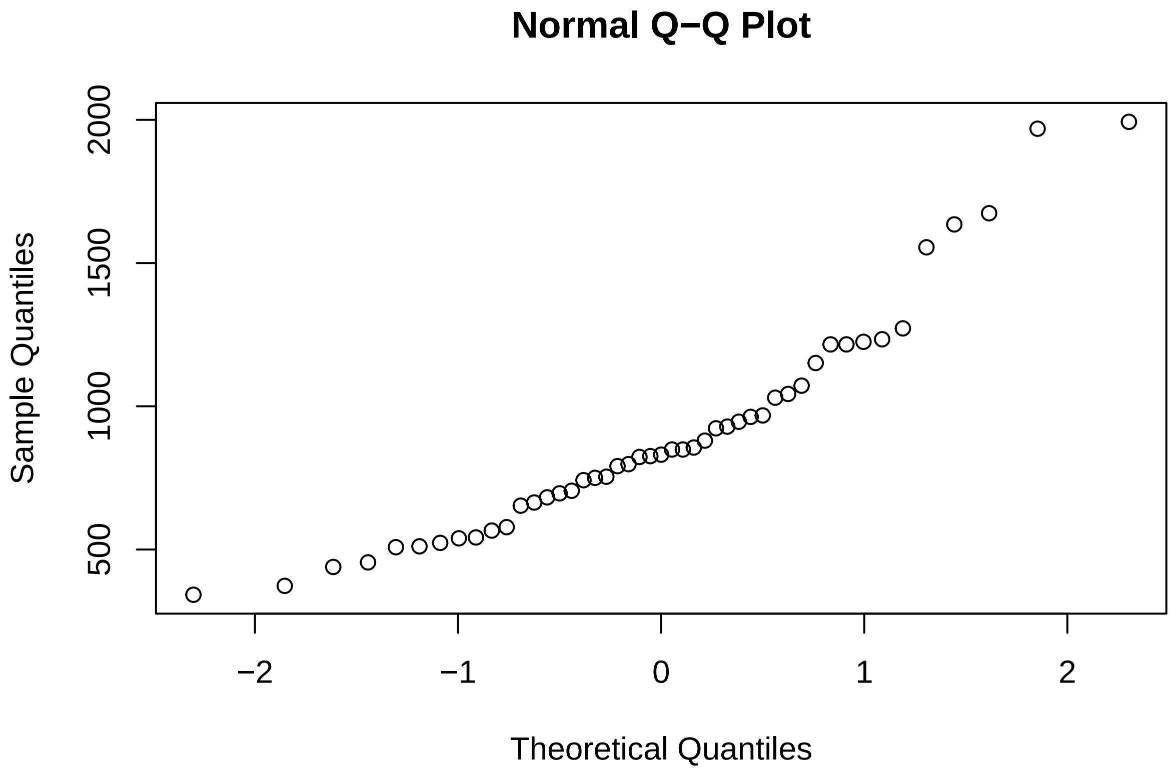
```r
#Linear regression model
lm_crime <- lm(Crime~., data = crime_data)
summary(lm_crime) #Model summary
```

```
##
## 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
```

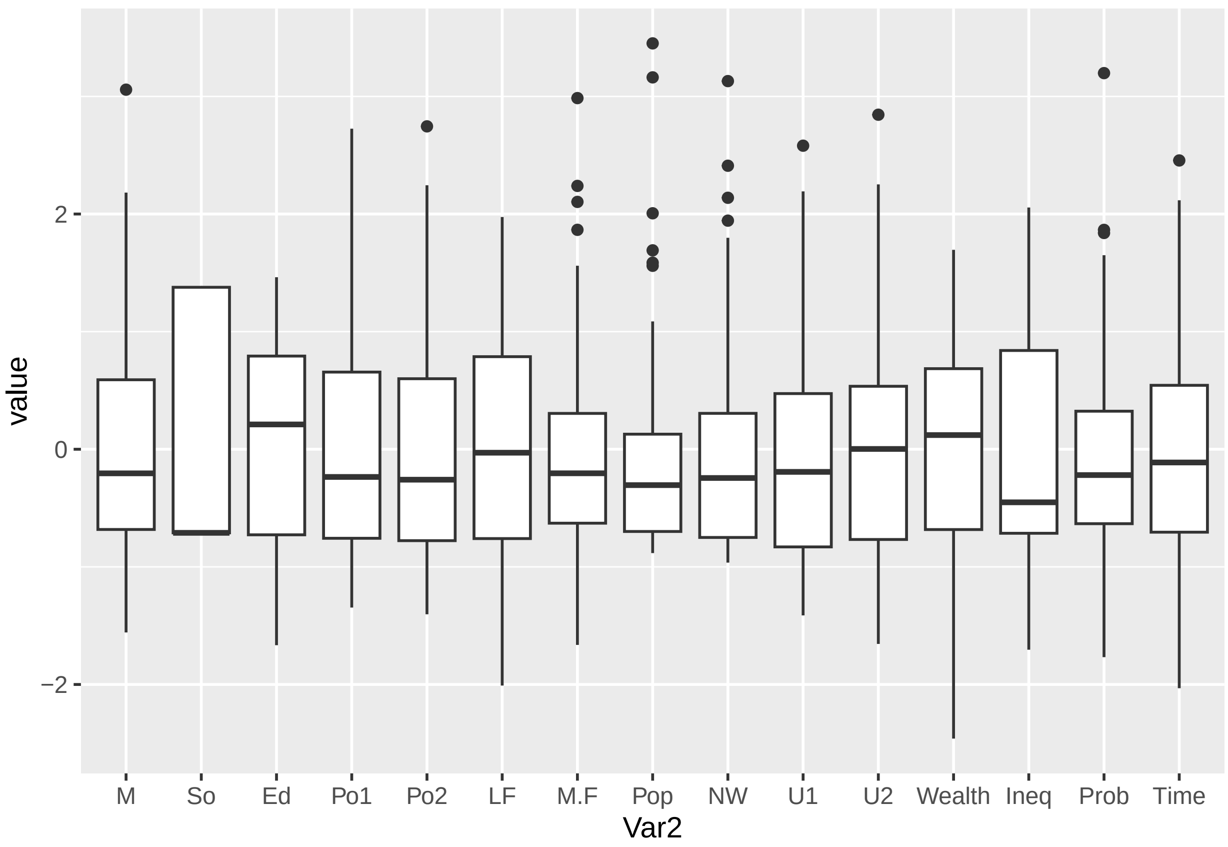
```r
#Test data
test <- data.frame(M = 14.0, So = 0, Ed = 10.0, Po1 = 12.0, Po2 = 15.5, LF = 0.640,
                   M.F = 94.0, Pop = 150, NW = 1.1, U1 = 0.120, U2 = 3.6, Wealth = 3200,
                   Ineq = 20.1, Prob = 0.04, Time = 39.0)

#Prediction on the data set
pred <- predict(lm_crime, test)
pred
```

```
##        1
## 155.4349
```

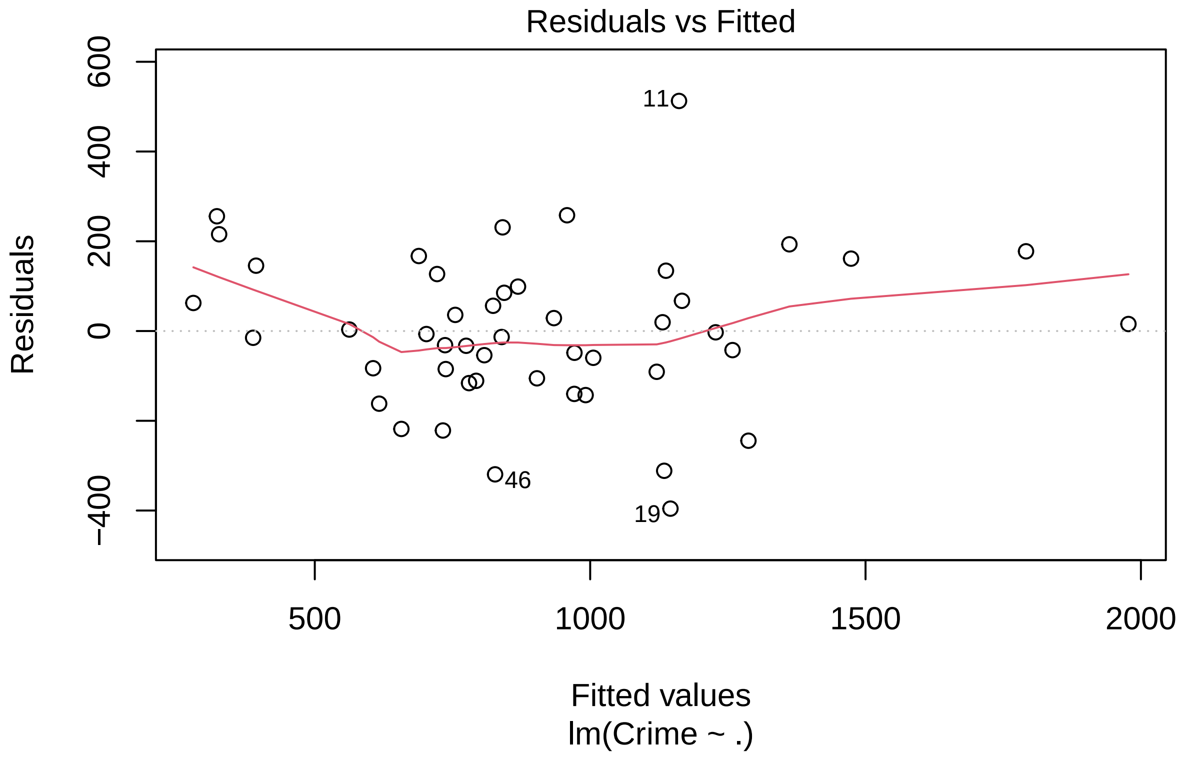
```r
#Check normality of the data set
qqnorm(crime_data$Crime)
```

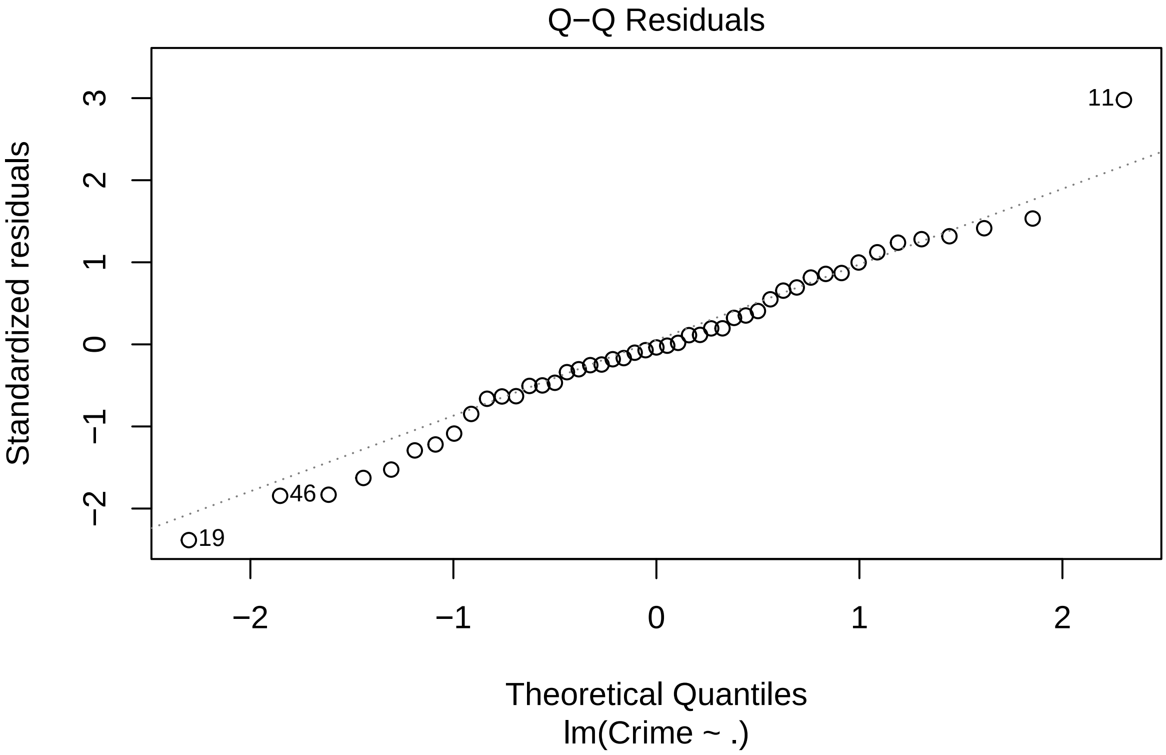


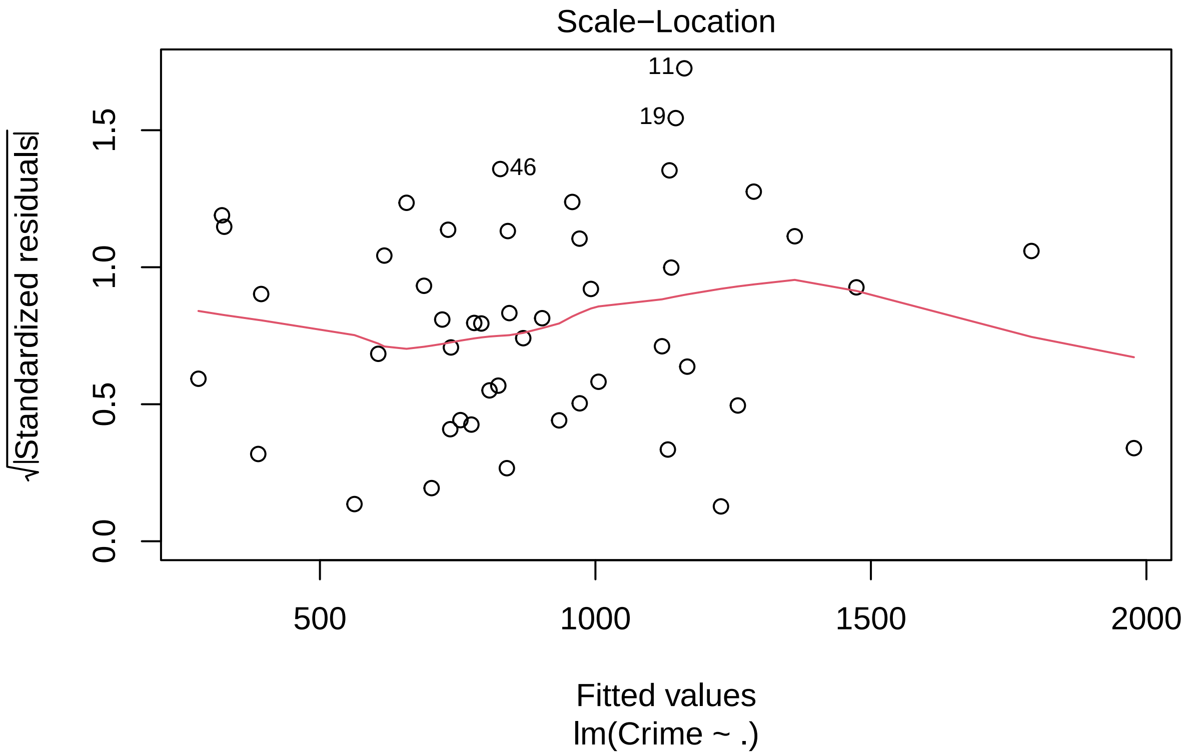
```
#It appears some of the predictors might have outliers considering
#the normality Q-Q graph is not straight.
#However, I cannot remove them. Therefore, I will keep the original
#data set without investigating the outliers.

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

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

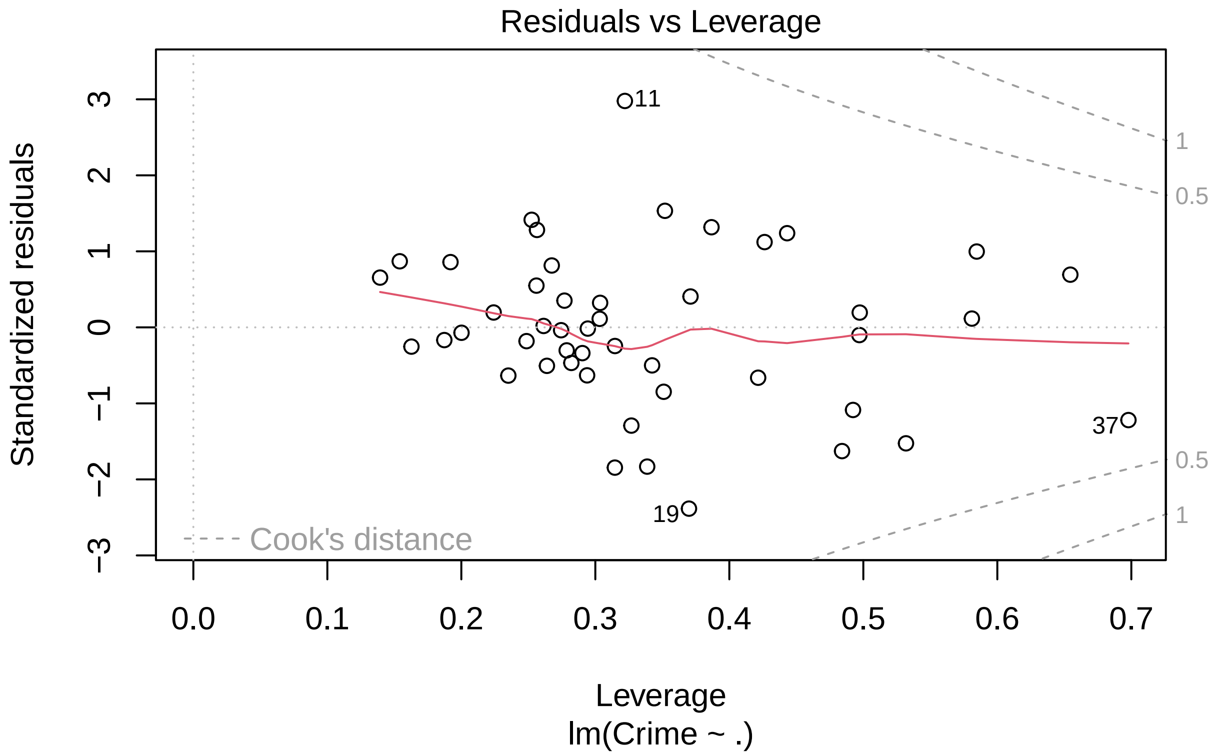


```r
plot(lm_crime)
```


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


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


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



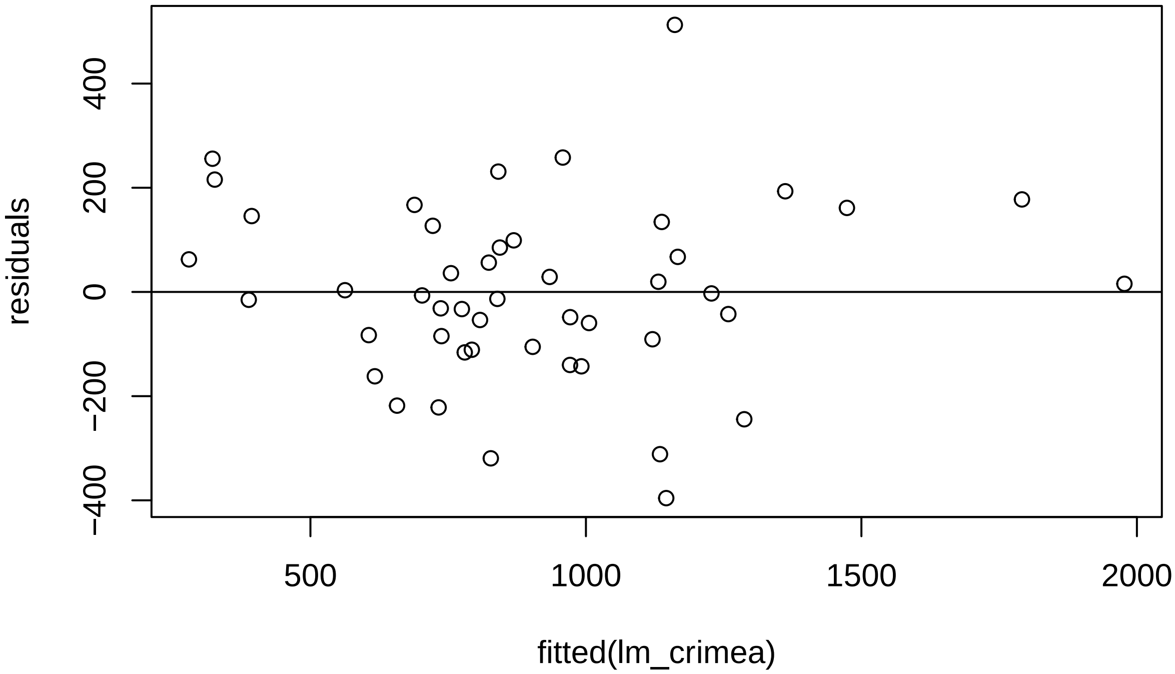
```r
set.seed(123)

train <- trainControl(method = "cv", number = 10)

lm_crimea <- train(Crime~., data = crime_data, method = 'lm', trControl = train)

summary(lm_crimea) #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
```

```
#Prediction on the test data
predModela <- predict(lm_crimea, test)
predModela
```

```
##        1
## 155.4349
```

```
#Plotting residuals vs fitted data to ensure no trends in errors.
residuals <- resid(lm_crimea)
plot(fitted(lm_crimea), residuals)
abline(0,0) #Adding a horizontal line
```



```
#The residual plot graph shows no pattern confirming that errors have no explanatory
#power for the above model.
```

In conclusion from the model above, the R^2 on training data is 0.803 since the model included some factors with high p-values it lead to some overfitting. This then led me to use cross-validation to estimate the quality of the model. However, the model's prediction on the test data set gives crime rate at 155 which is a problem considering the lowest crime in the data set is 342. The problem is that the model includes 15 factors that not all of them are significant. However, removing those factors isn't always efficient, especially when they have high p-values.

Model 2

```
lm_crime2 <- lm(Crime~M + Ed + Po1 + U2 + Ineq + Prob, data = crime_data)

summary(lm_crime2)
```

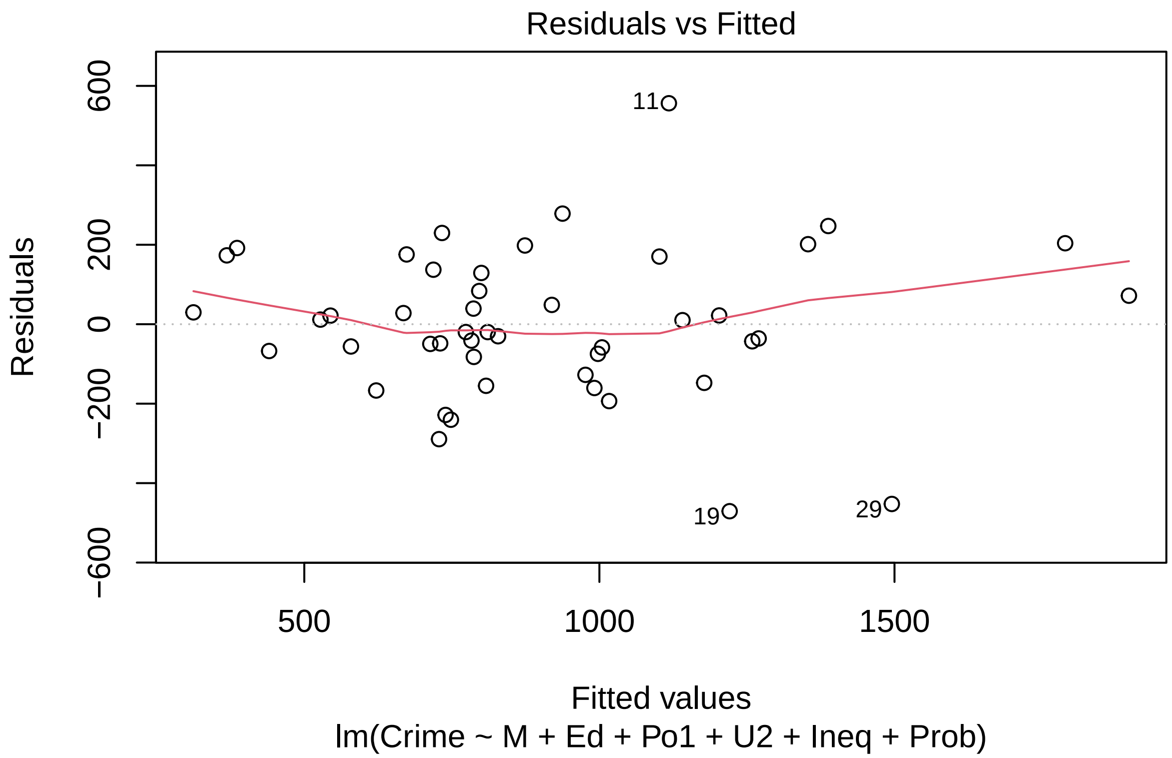
```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = crime_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -470.68  -78.41  -19.68  133.12  556.23
```

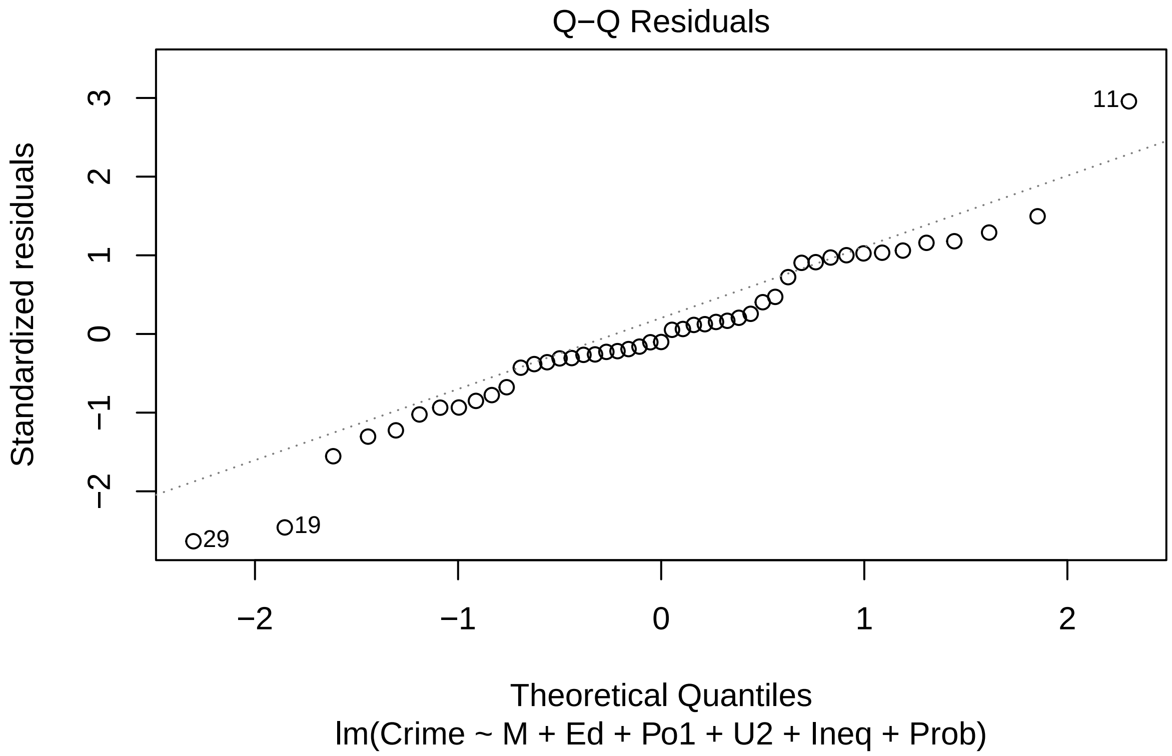
```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5040.50     899.84  -5.602 1.72e-06 ***
## M             105.02      33.30   3.154  0.00305 **
## Ed            196.47      44.75   4.390 8.07e-05 ***
## Po1           115.02      13.75   8.363 2.56e-10 ***
## 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
```

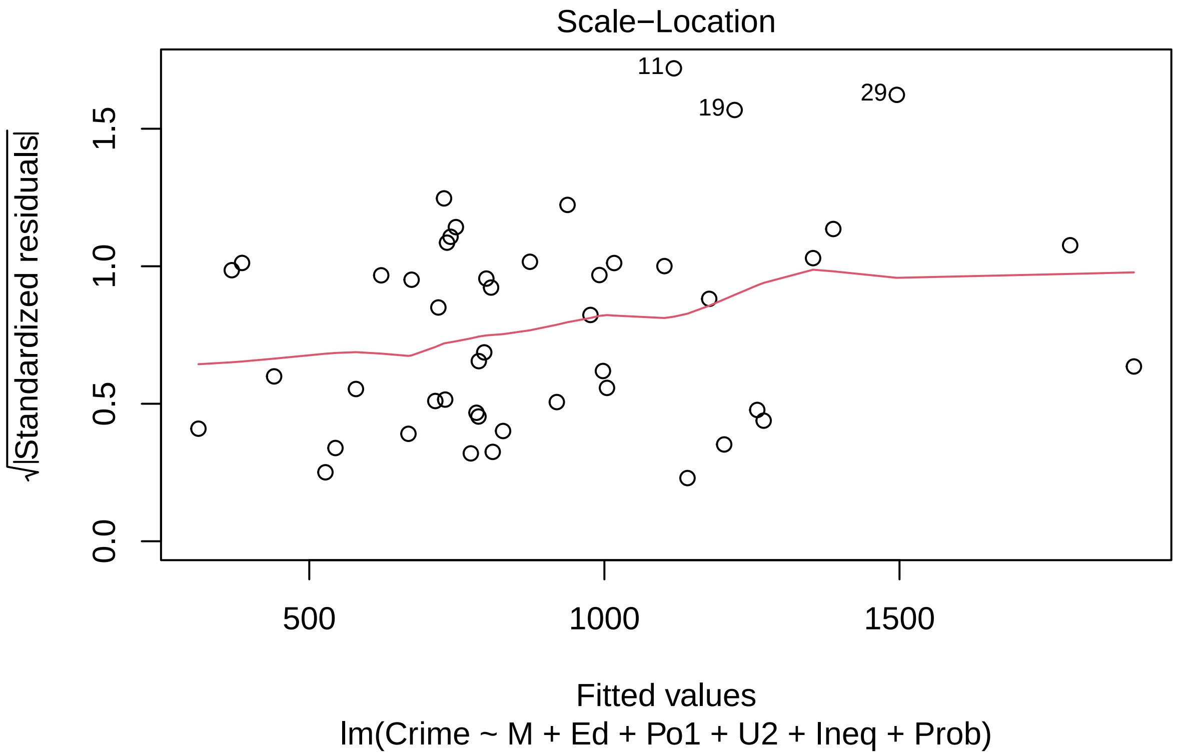
```r
pred2 <- predict(lm_crime2, test)
pred2
```

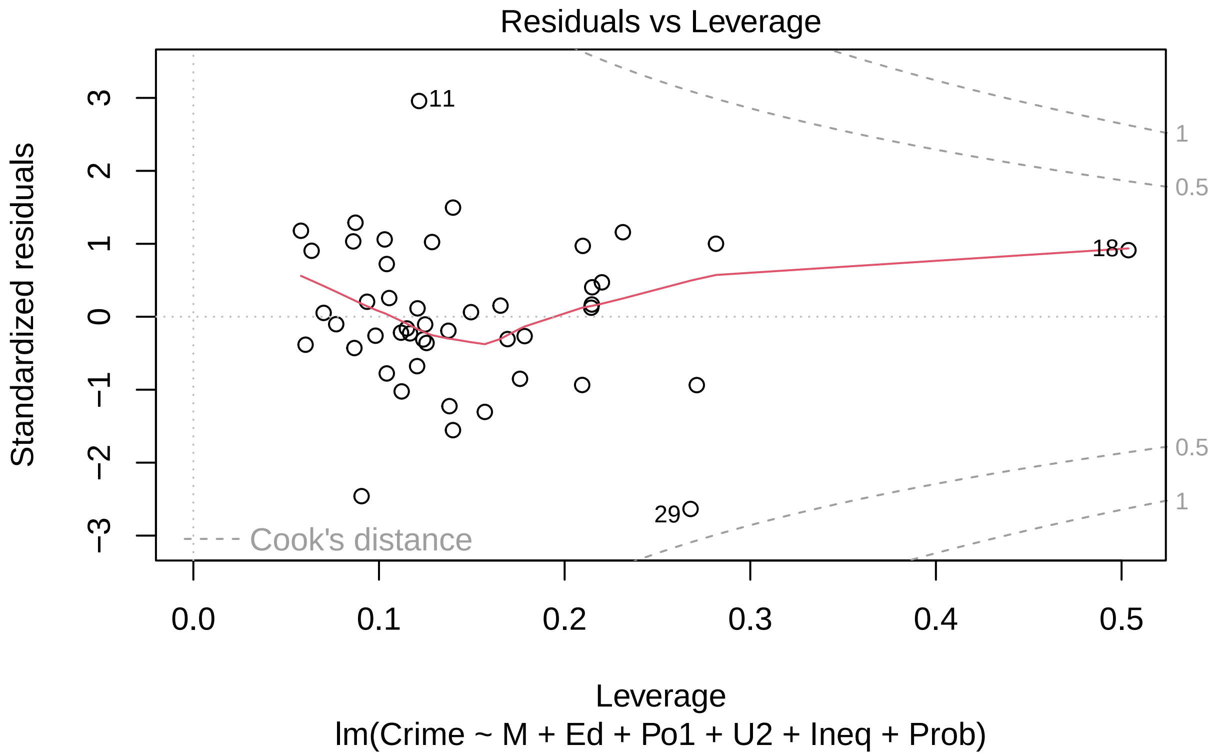
```
##        1
## 1304.245
```

```r
plot(lm_crime2)
```




Q−Q Residuals
Standardized residuals
3
2
1
0
−1
−2
11
29
19
−2
−1
0
1
2
Theoretical Quantiles
lm(Crime ~ M + Ed + Po1 + U2 + Ineq + Prob)


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



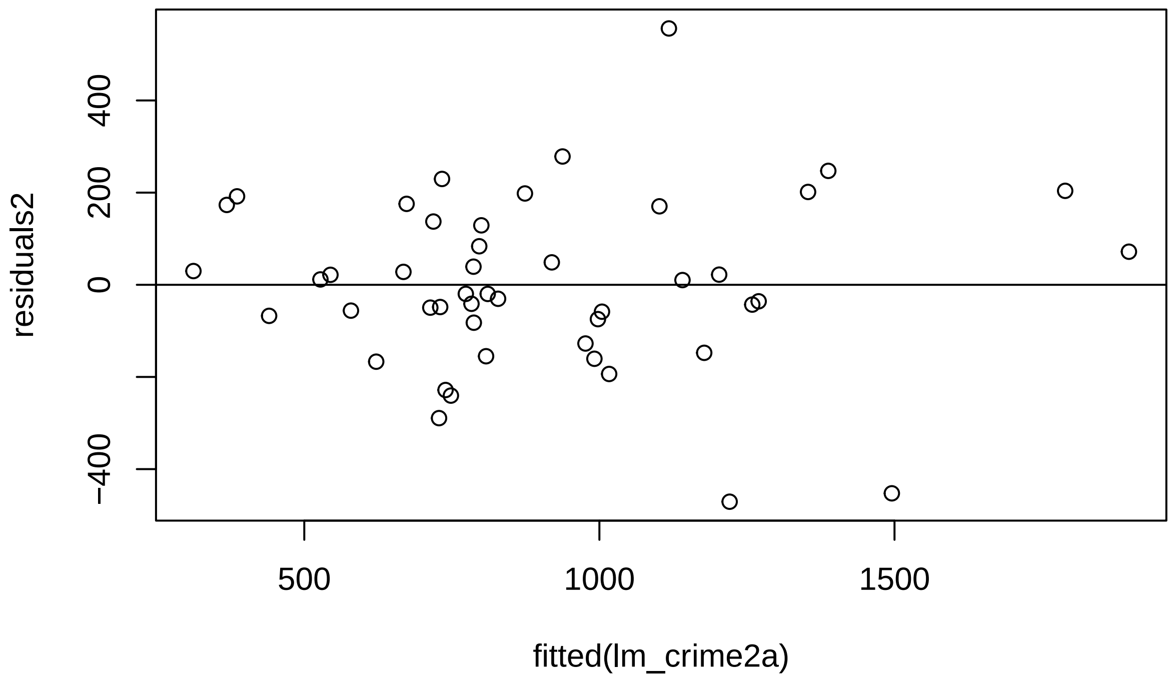
```r
#Linear regression using cross validation on the second model
set.seed(123)
train2 <- trainControl(method = "cv", number = 10)
lm_crime2a <- train(Crime~ M + Ed + Po1 + U2 +Ineq + Prob,
                    data = crime_data, method = 'lm',
                    trControl = train)
summary(lm_crime2a)
```

```
##
## 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(lm_crime2a, test)
pred2a
```

```
##        1
## 1304.245
```

```
#Plotting residual vs fitted data to ensure no trend in errors
residuals2 <- resid(lm_crime2a)
plot(fitted(lm_crime2a), residuals2)
abline(0,0)
```



```
#The residual plot graph shows no patterns confirming that errors have no power
#and the power of the model resides with the predictors.
```

In the second model above, I chose to reduce the number of factors to achieve a reasonable crime prediction. I re-fit the model using only factors with initial p-values 0.10 or lower since they showed significance in the summary of the model. This then gave a crime prediction of 1304 which a more-reasonable prediction for this model. Therefore, the model shows that the predictors are significant in terms of their p-values.