Homework 5 by Haritha Pulletikurti

2020-09-23

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:

We can use the regression model to predict the profit/loss rate of a business. The predictors for this scenario could be Daily Sales, Advertisement Budget, Qualified work staff budget, Supply budget, Number of Competitors, Number of Customers, etc.

Answer

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

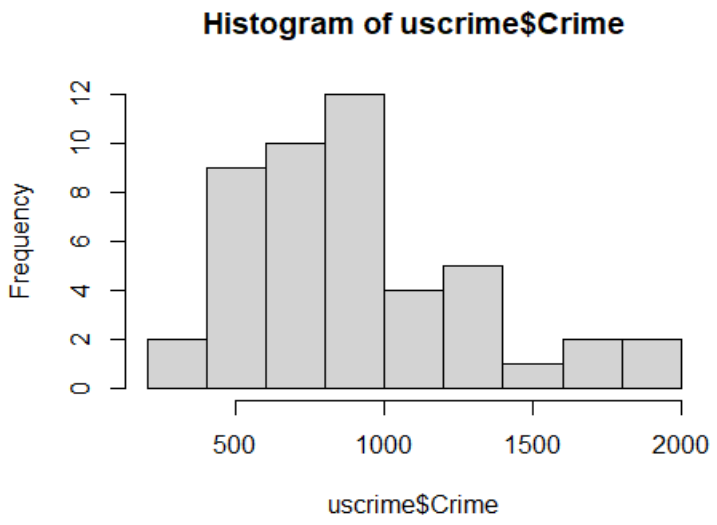
Show your model (factors used and their coefficients), the software output, and the quality of fit. Note that because there are only 47 data points and 15 predictors, you'll probably notice some overfitting. We'll see ways of dealing with this sort of problem later in the course.

Answer:

Model 0: Initial Model with 15 predictors and One Response Variable -Crime

Step1: Read and Analyze the data.

```
#Start Fresh
rm(list=ls())
#Load the GGally library for ggpairs()
library(GGally)

library(car) #using vif

library(DAAG)

#Read the data
uscrime <- read.table("uscrime.txt",stringsAsFactors = FALSE,header = TRUE)
hist(uscrime$Crime)
```



Analysis: The above histogram shows that the crime data is not distributed by the center mean.

To perform the Multiple Linear Regression using least squares, we can use the lm() function . Here the crime data set has 15 predictors or factors and the 16th variable "Crime" is the Response.

I will now look at the Regression Model with all the 15 predictors and will analyze the Test data point that is provided in the question to predict the Crime Rate. I will see if the prediction is good enough or needs further enhancement.

<u>Step 2:</u> Perform Multiple Regression on Data with all 15 predictors and analyze the predicted Y-HAT. Call this Model – 0

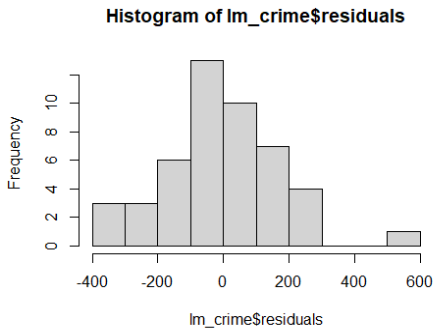
```
lm_crime <- lm(Crime~.,data = uscrime)
summary(lm_crime)

##
## Call:
## lm(formula = Crime ~ ., data = uscrime)
##
## 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

RSquared = summary(lm_crime)$r.sq
RSE = summary(lm_crime)$sigma

hist(lm_crime$residuals)
```

Analysis for Model-0[Regression model with All predictors Included]:

The lm() Multiple Regression function returned 0.8031 R squared Error and if the Threshold for Pvalue is 0.1 then there are about 6 predictors that add value. Please see the highlighted lines.

- The **F Statistic** in regression is used along the p-value tells us whether the results are significant enough to reject the null hypothesis. If F is bigger than some Predictor is significant enough to contribute in the regression equation that affects the Response variable.

- **<u>Null Hypothesis</u>:** No predictors are significant enough to change the Response Variable.Here P = 3.539e-07 and F > 8 telling us that we can **reject the Null Hypothesis**.
- The **Degrees of freedom** are used to compute the F value = Ratio of Chi Square Distribution and Degrees of freedom.
- **Residuals** = difference between observed Crime value "y" and the Predicted Crime value "yhat"

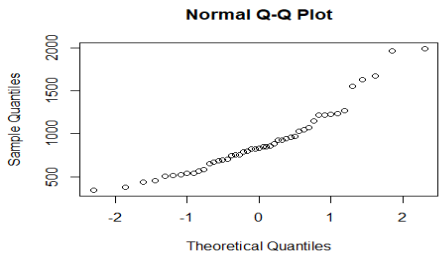


The linear regression model assumes that the relationship between n the Predictors(X) and the Response variable(Y) is linear. The above histogram shows that the values are normally distributed but not by the mean = 0 at the center. This shows the relationship between X and Y could be either polynomialic or logarithmic.

```
testpt <- 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)


#Predict the crime rate for the data point
predict_model <-predict(lm_crime,testpt)
predict_model
```

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

```
#Is this a good prediction?
qqnorm(uscrime$Crime)
```



```
range(uscrime$Crime)
## [1]  342 1993
```

Analysis after Predicting Y-HAT on Model-0:

The Above Predicted Crime Value (Y-hat) for the Test Data, Y-Hat = 155.4349.
The Range of Response Variable (Y) is lower limit = 342 and Upper Limit = 1993.
The Y-Hat value is out of range and far below the lower bound of Y.

This means that our Regression model is probably overfitting the data and needs to be enhanced.
Let us now look at the Variance Inflation Factors obtained by vif() function.

```
#compute variance inflation factors using vif
vif(lm_crime)
```

```
##        M       So       Ed      Po1      Po2       LF      M.F      Pop
##   2.8924   5.3428   5.0774 104.6600 113.5600   3.7127   3.7859   2.5367
##       NW       U1       U2   Wealth     Ineq     Prob     Time
##   4.6741   6.0639   5.0889  10.5300   8.6445   2.8095   2.7138
```

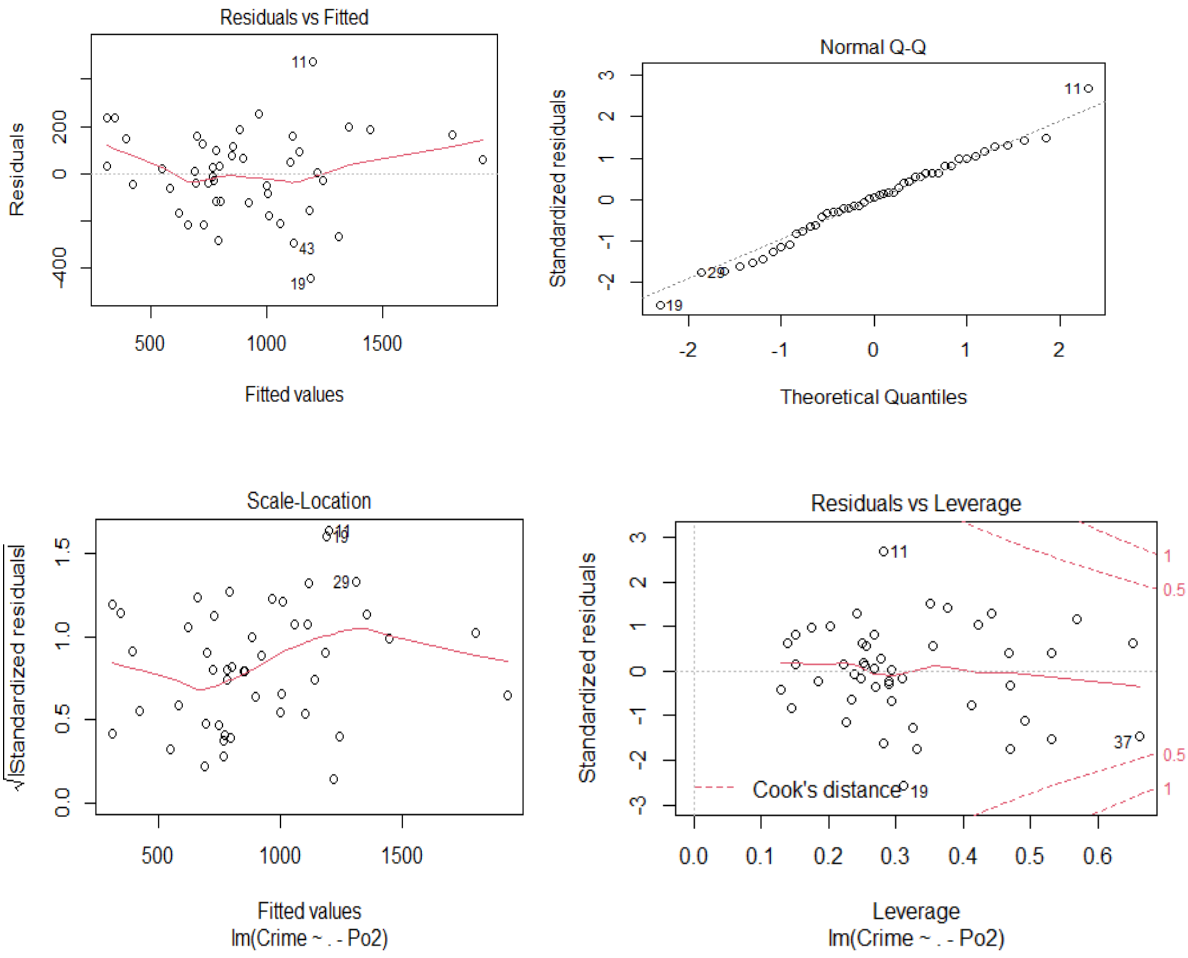
Analysis – Which Predictors can be removed?

Based on this result, the p-value is very high for the following factors
Po2,Po1,Wealth,Ineq,So,Ed,M.f,L.f
let us remove one by one and check the quality of the regression model.

Model 1: Initial Model with only 14 predictors, removing "Po2" predictor.

Step 1: Remove the Highest "Po2" variable

```
lm_model_1 =lm(Crime~.-Po2,data = uscrime)
summary(lm_model_1)$r.sq
## [1] 0.797576
summary(lm_model_1)$sigma
## [1] 208.6313
#Predict the crime rate for the data point
predict_model_1 <-predict(lm_model_1,testpt)
predict_model_1
##        1
## 724.8202
range(uscrime$Crime)
## [1]  342 1993
# The predicted value is within the range
plot(lm_model_1)
```



Analysis for Model-1:

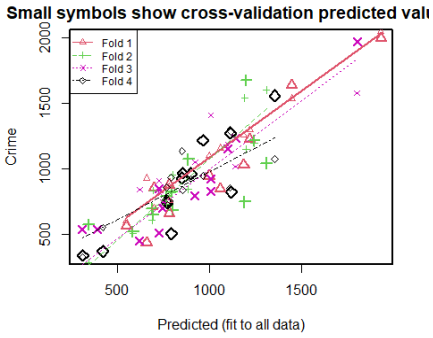
Sigma (Residual Standard Error (RSA)) = 208.63
R-Squared = 0.797576
Predicted YHat = 724.8202 (within the Range of Y: 342 – 1993)

Cross Validation Linear Regression using 4 folds:
Let us now perform Cross Validation Linear Regression with number of folds = 4.
We are doing this to get the Total Sum of Squared errors – Mean value from all the four folds.

Then we compute the R- Squared error of the Model by
SSE = <- Mean value * NRows(Data)
SST = sum((uscrime$Crime-mean(uscrime$Crime))^2)
RSQ of the Model = 1-(sse/sst)

Implementation:

```
set.seed(42)
options(warn=-1)
lm_model_1_cv<-cv.lm(uscrime,lm_model_1,m=4)

## Analysis of Variance Table
##
## Response: Crime
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## M          1   55084   55084    1.27   0.269
## So         1   15370   15370    0.35   0.557
## Ed         1  905668  905668   20.81 7.1e-05 ***
## Po1        1 3076033 3076033   70.67 1.3e-09 ***
## LF         1  120696  120696    2.77   0.106
## M.F        1  138150  138150    3.17   0.084 .
## Pop        1   52880   52880    1.21   0.279
## NW         1    7274    7274    0.17   0.685
## U1         1   15514   15514    0.36   0.555
## U2         1  280663  280663    6.45   0.016 *
## Wealth     1   42944   42944    0.99   0.328
## Ineq       1  566547  566547   13.02   0.001 **
## Prob       1  210003  210003    4.82   0.035 *
## Time       1    1236    1236    0.03   0.867
## Residuals 32 1392865   43527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
##
## fold 1
## Observations in test set: 11
##                 2   9   14   16     20   22     26    38  41   44   47
## Predicted   1449.1 700  780  998 1221.5  658 1933.7 546.7 781 1188 1061
## cvpred      1535.2 706  867 1100 1298.8  931 2044.2 603.1 757 1256 1158
## Crime       1635.0 856  664  946 1225.0  439 1993.0 566.0 880 1030  849
## CV residual   99.8 150 -203 -154  -73.8 -492  -51.2 -37.1 123 -226 -309
##
## Sum of squares = 510914    Mean square = 46447    n = 11
##
## fold 2
## Observations in test set: 12
##                1   3    6   11   19    25   28   29    30   33   35    39
## Predicted   764.9 342  799 1201 1193 583.8 1244 1310 687.1  886  693 796.4
## cvpred      734.2 287  955 1149 1539 509.5 1197 1602 610.1  848  836 821.7
## Crime       791.0 578  682 1674  750 523.0 1216 1043 696.0 1072  653 826.0
## CV residual  56.8 291 -273  525 -789  13.5   19 -559  85.9  224 -183   4.3
##
## Sum of squares = 1464789    Mean square = 122066    n = 12
##
## fold 3
## Observations in test set: 12
##                4    5    10  12   13   15    17     34   37     40  42   45
## Predicted   1804 1142 746.7 725  727  922 392.1 1006.8 1008 1100.5 308  620
```

```
## cvpred      1578 1017 748.1 826  910 1052  99.1  824.1 1409 1183.8 -92  845
## Crime       1969 1234 705.0 849  511  798 539.0  923.0  831 1151.0 542  455
## CV residual  391  217 -43.1  23 -399 -254 439.9   98.9 -578  -32.8 634 -390
##
## Sum of squares = 1518656    Mean square = 126555    n = 12
##
## fold 4
## Observations in test set: 12
##                 7    8   18    21   23  24    27   31   32   36   43   46
## Predicted   900.4 1356  852 772.1  964 855 310.5  420  768 1113 1119  793
## cvpred      922.9 1075 1138 727.5  949 841 330.2  554  862  856 1286  934
## Crime       963.0 1555  929 742.0 1216 968 342.0  373  754 1272  823  508
## CV residual  40.1  480 -209  14.5  267 127  11.8 -181 -108  416 -463 -426
##
## Sum of squares = 976401    Mean square = 81367    n = 12
##
## Overall (Sum over all 12 folds)
##    ms
## 95123
```

```
# Let us calculate the Rsquared error
sse_model1<-95123*nrow(uscrime)
sst_model1<-sum((uscrime$Crime-mean(uscrime$Crime))^2)
rsq_model1<-1-sse_model1/sst_model1
rsq_model1
```

```
## [1] 0.35
```

Analysis for Model1 (Removed "Po2" Predictor):

Sigma (Residual Standard Error (RSA)) = 208.63
R-Squared of cv_lm()= 0.797576
R Squared value of Model 1 (Computed Manually) = 0.35
Predicted YHat = 724.8202 (within the Range of Y: 342 – 1993)
Conclusion: This shows that removing the variables that are not necessary reduces over fitting of the data.

```
# Let us analyse the results of the summary of lm_model_1
summary(lm_model_1)
```

```
##
## Call:
## lm(formula = Crime ~ . - Po2, data = uscrime)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -442.6 -116.5    8.9  118.3  473.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.38e+03   1.57e+03   -4.07  0.00029 ***
## M            8.99e+01   4.16e+01    2.16  0.03823 *
## So           5.67e+00   1.48e+02    0.04  0.96970
## Ed           1.77e+02   6.08e+01    2.92  0.00644 **
## Po1          9.65e+01   2.39e+01    4.04  0.00032 ***
## LF          -2.80e+02   1.41e+03   -0.20  0.84354
## M.F          1.82e+01   2.03e+01    0.90  0.37603
## Pop         -7.84e-01   1.29e+00   -0.61  0.54652
## NW           2.45e+00   6.19e+00    0.40  0.69524
## U1          -5.42e+03   4.18e+03   -1.30  0.20416
## U2           1.69e+02   8.21e+01    2.06  0.04744 *
## Wealth       9.07e-02   1.03e-01    0.88  0.38629
## Ineq         7.27e+01   2.26e+01    3.22  0.00292 **
## Prob        -4.29e+03   2.18e+03   -1.96  0.05848 .
## Time        -1.13e+00   6.69e+00   -0.17  0.86725
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209 on 32 degrees of freedom
## Multiple R-squared:  0.798,  Adjusted R-squared:  0.709
## F-statistic: 9.01 on 14 and 32 DF,  p-value: 1.67e-07
```

Analysis for Model 1 on which predictors can be removed to enhance it further:

1. The Summary shows that if we Threshold(P_value) = 0.1, then factors above the threshold value if removed might give us a better regression model
   SO,LF,M.F,POP,NW,U1,Wealth,Time are above the threshold P-Value = 0.1

2. Also let us consider the variance inflation factors

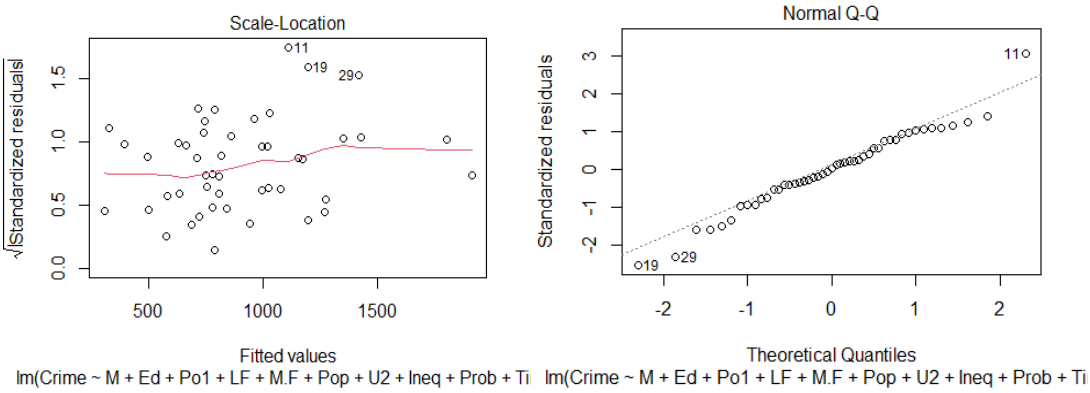
```
vif(lm_model_1)
##       M      So      Ed     Po1      LF     M.F     Pop      NW      U1      U2  Wealth
##    2.88    5.32    4.89    5.34    3.42    3.78    2.53    4.28    6.00    5.09   10.50
##    Ineq    Prob    Time
##    8.56    2.61    2.38
```

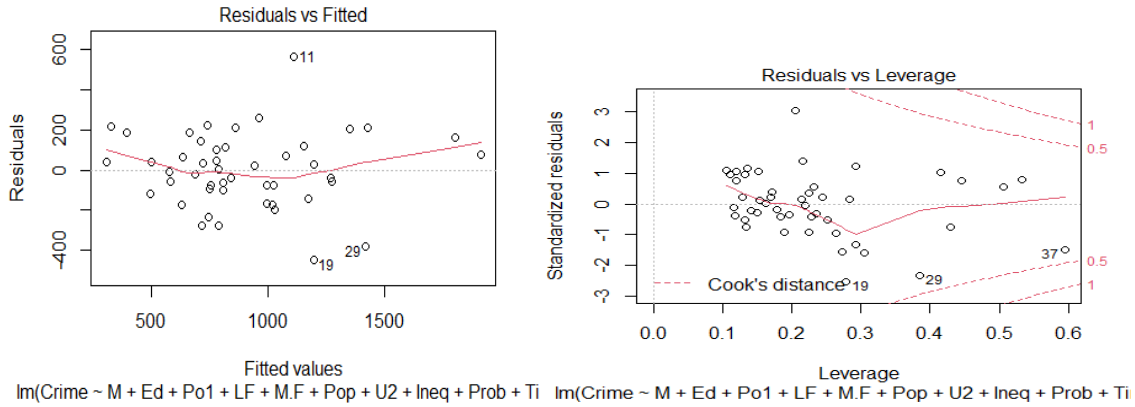
Based on 1 and 2 this -We can remove Wealth, Ineq, U1, PO1, SO, U2, NW as they are common in both the lists.

Model 2: Initial Model with only 10 predictors, removing "P02 ,Wealth,U1, SO,NW" predictor.

Step 1 : As Wealth,U1, SO,NW are common in both lists , let us remove those and check the regression model.

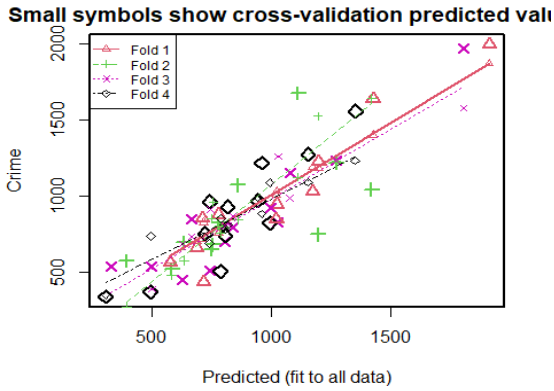
```
lm_model_2 =update(lm_model_1,~.-Wealth-U1-So-NW)

#Predict the crime rate for the data point
predict_model_2 <-predict(lm_model_2,testpt)
predict_model_2
##    1
## 1254
range(uscrime$Crime)
## [1]  342 1993
plot(lm_model_2)
```






```
lm_model_2_cv<-cv.lm(uscrime,lm_model_2,m=4)
```

```
## Analysis of Variance Table
##
## Response: Crime
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## M          1   55084   55084    1.28 0.26510
## Ed         1  725967  725967   16.89 0.00022 ***
## Po1        1 3173852 3173852   73.84   3e-10 ***
## LF         1   62131   62131    1.45 0.23711
## M.F        1  130888  130888    3.05 0.08952 .
## Pop        1   50474   50474    1.17 0.28574
## U2         1  175814  175814    4.09 0.05061 .
## Ineq       1  698861  698861   16.26 0.00027 ***
## Prob       1  260103  260103    6.05 0.01883 *
## Time       1     332     332    0.01 0.93044
## Residuals 36 1547420   42984
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
##
## fold 1
## Observations in test set: 11
##                2   9    14     16     20   22   26  38  41   44   47
## Predicted   1426 710 687.2 1022.1 1199.9  716 1914 578 779 1174 1022
## cvpred      1402 720 705.4 1035.8 1184.7  829 1868 581 760 1190 1007
## Crime       1635 856 664.0  946.0 1225.0  439 1993 566 880 1030  849
## CV residual  233 136 -41.4  -89.8   40.3 -390  125 -15 120 -160 -158
##
## Sum of squares = 316840    Mean square = 28804    n = 11
##
## fold 2
## Observations in test set: 12
##                1   3    6   11   19    25     28   29  30   33   35    39
## Predicted   787.0 393  757 1112 1195 582.2 1272.2 1421 633  859  749 781.1
## cvpred      759.5 277  960 1125 1523 482.6 1231.6 1641 577  845  855 806.2
## Crime       791.0 578  682 1674  750 523.0 1216.0 1043 696 1072  653 826.0
## CV residual  31.5 301 -278  549 -773  40.4  -15.6 -598 119  227 -202  19.8
##
## Sum of squares = 1535050    Mean square = 127921    n = 12
##
## fold 3
## Observations in test set: 12
##                4      5   10  12   13    15  17      34   37   40  42   45
## Predicted   1806 1271.0  807 664  745 840.0 499 997.090 1028 1078 328  628
## cvpred      1575 1236.9  807 735  914 870.8 393 922.361 1261  988 174  689
## Crime       1969 1234.0  705 849  511 798.0 539 923.000  831 1151 542  455
## CV residual  394   -2.9 -102 114 -403 -72.8 146   0.639 -430  163 368 -234
##
## Sum of squares = 768628    Mean square = 64052    n = 12
##
## fold 4
## Observations in test set: 12
##                7    8  18    21   23    24    27   31    32   36   43   46
## Predicted   741 1353 817 807.2  959 944.6 305.0  494 723.3 1154  993  787
## cvpred      691 1233 810 785.6  884 995.8 353.4  739 734.7 1092 1087  855
## Crime       963 1555 929 742.0 1216 968.0 342.0  373 754.0 1272  823  508
## CV residual 272  322 119 -43.6  332 -27.8 -11.4 -366  19.3  180 -264 -347
##
## Sum of squares = 661860    Mean square = 55155    n = 12
##
## Overall (Sum over all 12 folds)
##    ms
## 69838
```

```
# Let us calculate the Rsquared error
sse_model2<-69838*nrow(uscrime)
sst_model2<-sum((uscrime$Crime-mean(uscrime$Crime))^2)
rsq_model2<-sse_model2/sst_model2
rsq_model2
```

```
## [1] 0.477
```

```
# Let us analyse the results of the summary of lm_model_2

summary(lm_model_2)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + LF + M.F + Pop + U2 + Ineq +
##     Prob + Time, data = uscrime)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -445.2  -98.8    4.0  114.6  562.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5272.819   1417.693   -3.72  0.00068 ***
## M              95.731     37.412    2.56  0.01485 *
## Ed            168.379     58.120    2.90  0.00637 **
## Po1           124.136     16.916    7.34  1.2e-08 ***
## LF            375.021   1165.213    0.32  0.74943
## M.F             3.865     16.687    0.23  0.81816
## Pop            -1.049      1.252   -0.84  0.40793
## U2             91.767     50.272    1.83  0.07624 .
## Ineq           68.360     15.400    4.44  8.2e-05 ***
## Prob        -3984.293   1972.674   -2.02  0.05090 .
## Time            0.568      6.460    0.09  0.93044
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 207 on 36 degrees of freedom
## Multiple R-squared:  0.775,  Adjusted R-squared:  0.713
## F-statistic: 12.4 on 10 and 36 DF,  p-value: 6.08e-09
```

```
# Also let us consider the variance inflation factors
vif(lm_model_2)
```

```
##    M   Ed  Po1   LF  M.F  Pop   U2 Ineq Prob Time
## 2.37 4.52 2.70 2.37 2.59 2.43 1.93 4.04 2.15 2.24
```

As we see after removing the factors, the R-squared is little reduced.

Analysis:

Sigma (Residual Standard Error (RSA)) = 207

R-Squared of lm()= 0.775

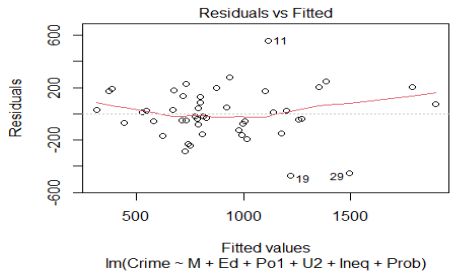
R-Squared from CV_LM()= 0.477

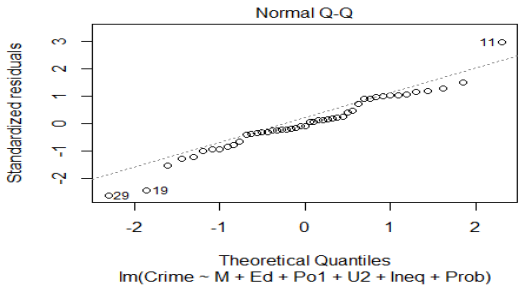
Predicted YHat = 1254 (within the Range of Y: 342 – 1993)

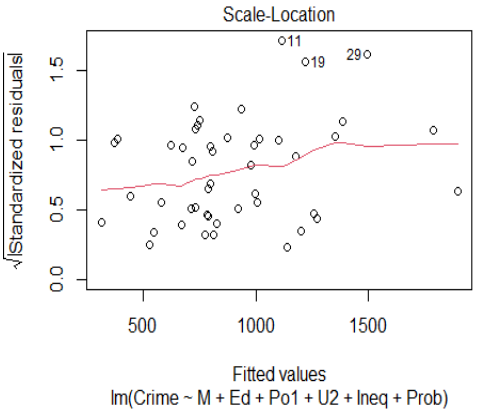
Common Factors from cv_lm(Model2) summary(Threshold P_Value > 0.1) and High Variance Inflation factors are "LF,M.F, Pop, Time"

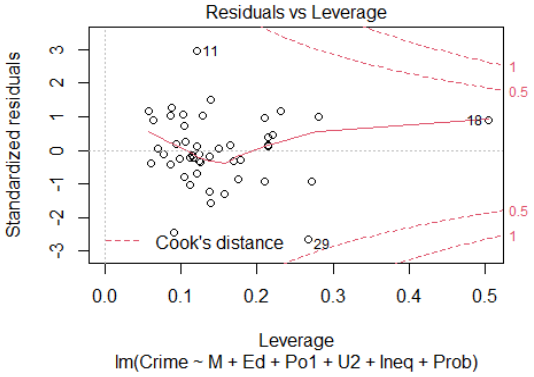
Model 3: Initial Model with only 10 predictors, removing "P02,Wealth,U1, SO,NW, LF,M.F, Pop, Time" predictor.

```
lm_model_3 =update(lm_model_2,~.-LF-M.F-Pop-Time)
#Predict the crime rate for the data point
predict_model_3 <-predict(lm_model_3,testpt)
predict_model_3##    1
## 1304
range(uscrime$Crime)
## [1]  342 1993
plot(lm_model_3)
```

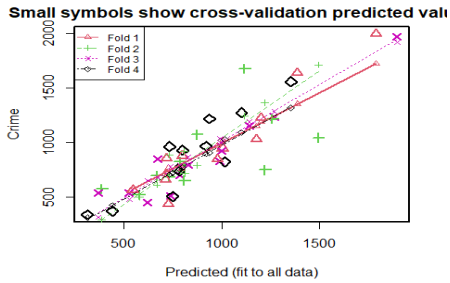









```
# Let us analyse the results of the summary of lm_model_1
lm_model_3_cv<-cv.lm(uscrime,lm_model_3,m=4)

## Analysis of Variance Table
##
## Response: Crime
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## M          1   55084   55084    1.37 0.24914
## Ed         1  725967  725967   18.02 0.00013 ***
## Po1        1 3173852 3173852   78.80 5.3e-11 ***
## U2         1  217386  217386    5.40 0.02534 *
## Ineq       1  848273  848273   21.06 4.3e-05 ***
## Prob       1  249308  249308    6.19 0.01711 *
## Residuals 40 1611057   40276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
##
## fold 1
## Observations in test set: 11
##                2   9    14     16     20   22   26    38  41   44   47
## Predicted   1388 719 713.6 1004.4 1203.0  728 1789 544.4 796 1178  976
## cvpred      1355 731 731.1 1023.2 1187.6  771 1720 588.4 763 1150  970
## Crime       1635 856 664.0  946.0 1225.0  439 1993 566.0 880 1030  849
## CV residual  280 125 -67.1  -77.2   37.4 -332  273 -22.4 117 -120 -121
##
## Sum of squares = 334042    Mean square = 30367    n = 11
##
## fold 2
## Observations in test set: 12
##                 1   3    6   11   19    25     28   29    30   33   35    39
## Predicted   810.8 386  730 1118 1221 579.1 1259.0 1495 668.0  874  808 786.7
## cvpred      716.9 296  888 1241 1363 504.3 1208.7 1711 614.2  792  919 736.6
## Crime       791.0 578  682 1674  750 523.0 1216.0 1043 696.0 1072  653 826.0
## CV residual  74.1 282 -206  433 -613  18.7    7.3 -668  81.8  280 -266  89.4
##
## Sum of squares = 1300449    Mean square = 108371    n = 12
##
## fold 3
## Observations in test set: 12
##                  4      5    10  12   13  15    17    34   37     40  42   45
## Predicted   1897.2 1269.8 787.3 673  739 828 527.4 997.5  992 1140.8 369  622
## cvpred      1916.6 1282.8 791.8 680  778 867 483.3 998.2 1037 1190.7 317  656
## Crime       1969.0 1234.0 705.0 849  511 798 539.0 923.0  831 1151.0 542  455
## CV residual   52.4  -48.8 -86.8 169 -267 -69  55.7 -75.2 -206  -39.7 225 -201
##
## Sum of squares = 261503    Mean square = 21792    n = 12
##
## fold 4
## Observations in test set: 12
##                7    8  18  21   23    24    27  31  32   36   43   46
## Predicted   733 1354 800 783  938 919.4 312.2 440 774 1102 1017  748
## cvpred      708 1319 771 759  909 896.3 316.2 426 740 1093 1027  723
## Crime       963 1555 929 742 1216 968.0 342.0 373 754 1272  823  508
## CV residual 255  236 158 -17  307  71.7  25.8 -53  14  179 -204 -215
##
## Sum of squares = 369549    Mean square = 30796    n = 12
##
## Overall (Sum over all 12 folds)
##    ms
## 48203

# Let us calculate the Rsquared error
sse_model3<-48203*nrow(uscrime)
sst_model3<-sum((uscrime$Crime-mean(uscrime$Crime))^2)
rsq_model3<-1-sse_model3/sst_model3
rsq_model3
```

```
## [1] 0.671

summary(lm_model_3)

##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = uscrime)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -470.7  -78.4  -19.7  133.1  556.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5040.5      899.8   -5.60  1.7e-06 ***
## M              105.0       33.3    3.15   0.0031 **
## Ed             196.5       44.8    4.39  8.1e-05 ***
## Po1            115.0       13.8    8.36  2.6e-10 ***
## U2              89.4       40.9    2.18   0.0348 *
## Ineq            67.7       13.9    4.85  1.9e-05 ***
## Prob         -3801.8     1528.1   -2.49   0.0171 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 201 on 40 degrees of freedom
## Multiple R-squared:  0.766,  Adjusted R-squared:  0.731
## F-statistic: 21.8 on 6 and 40 DF,  p-value: 3.42e-11

# Also let us consider the variance inflation factors
vif(lm_model_3)

##    M   Ed  Po1   U2 Ineq Prob
## 2.00 2.86 1.91 1.36 3.53 1.38
```

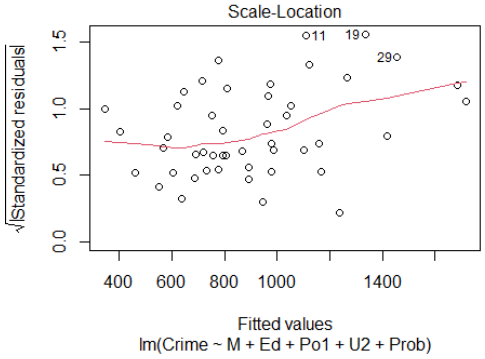
Analysis of Model 3:

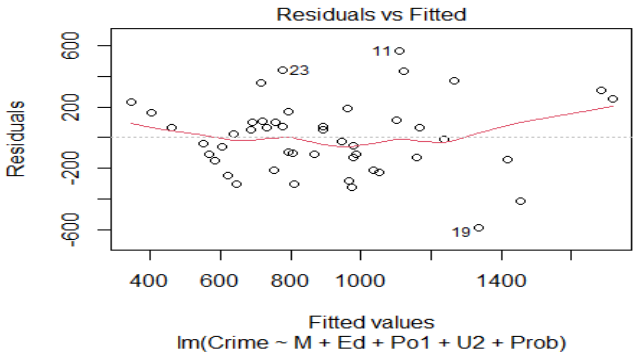
Sigma (Residual Standard Error (RSA)) = 201
Multiple R-Squared = 0.766
R-Squared from CV_LM()= 0.671
Predicted YHat = 1304 (within the Range of Y: 342 – 1993)

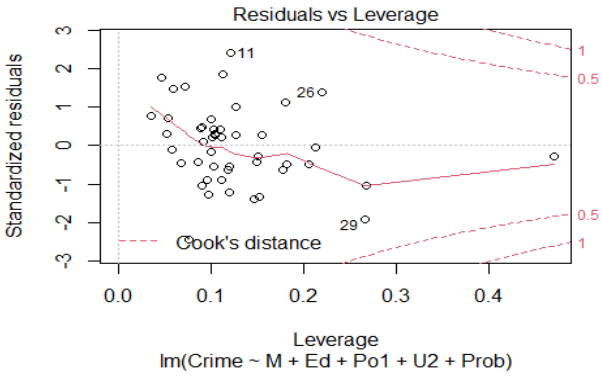
Common Factors from CV_LM(Model2) summary(Threshold P_Value > 0.1) and High Variance Inflation factors are "Ineq"

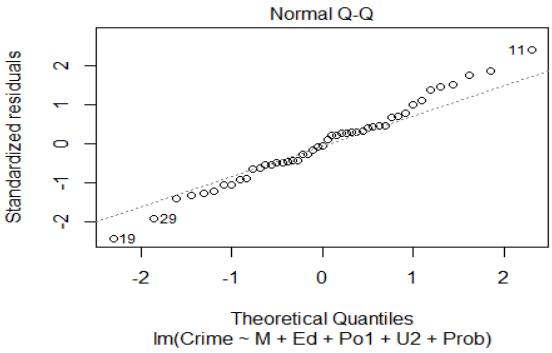
Model 4: Initial Model with only 10 predictors, removing "P02,Wealth,U1, SO,NW, LF,M.F, Pop, Time,Ineq" predictor.

```
Model 4 : Let is further remove " Ineq "
lm_model_4 =update(lm_model_3,~.-Ineq)
#Predict the crime rate for the data point
predict_model_4 <-predict(lm_model_4,testpt)
predict_model_4
##    1
## 1250
range(uscrime$Crime)
## [1]  342 1993
plot(lm_model_4)
```

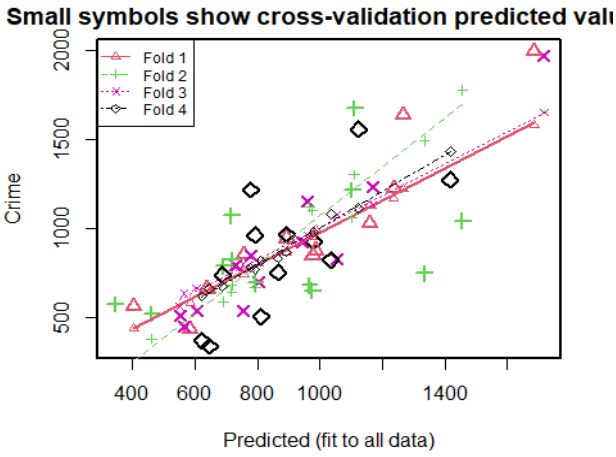









```
# Let us analyse the results of the summary of lm_model_1
lm_model_4_cv<-cv.lm(uscrime,lm_model_4,m=4)

## Analysis of Variance Table
##
## Response: Crime
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## M          1   55084   55084    0.88  0.3531
## Ed         1  725967  725967   11.63  0.0015 **
## Po1        1 3173852 3173852   50.83 1.1e-08 ***
## U2         1  217386  217386    3.48  0.0692 .
## Prob       1  148360  148360    2.38  0.1309
## Residuals 41 2560278   62446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
##
## fold 1
## Observations in test set: 11
##                 2   9     14    16     20   22   26  38   41   44   47
## Predicted    1267 756 639.04 893.7 1235.3  585 1685 404  987 1159  976
## cvpred       1225 749 660.49 864.2 1172.9  588 1580 443  996 1132  953
## Crime        1635 856 664.00 946.0 1225.0  439 1993 566  880 1030  849
## CV residual   410 107   3.51  81.8   52.1 -149  413 123 -116 -102 -104
##
## Sum of squares = 431031    Mean square = 39185    n = 11
##
## fold 2
## Observations in test set: 12
##                1   3    6   11   19  25   28   29    30   33   35  39
## Predicted    690 344  966 1110 1334 461 1103 1456 793.1  717  976 718
## cvpred       591 243 1124 1302 1492 385 1059 1775 670.1  647 1102 684
## Crime        791 578  682 1674  750 523 1216 1043 696.0 1072  653 826
## CV residual 200 335 -442  372 -742 138  157 -732  25.9  425 -449 142
##
## Sum of squares = 2020015    Mean square = 168335    n = 12
##
## fold 3
## Observations in test set: 12
##                4    5   10  12    13    15   17  34   37   40   42   45
## Predicted    1717 1169  805 778 551.5 729.8  754 945 1054  960  604  566
## cvpred       1651 1132  814 782 563.4 786.6  798 938 1080  962  667  639
## Crime        1969 1234  705 849 511.0 798.0  539 923  831 1151  542  455
## CV residual   318  102 -109  67 -52.4  11.4 -259 -15 -249  189 -125 -184
##
## Sum of squares = 345481    Mean square = 28790    n = 12
##
## fold 4
## Observations in test set: 12
##                7    8    18    21   23  24   27   31  32   36   43   46
## Predicted    792 1123 980.0 688.2  778 893  646  623 865 1417 1035  812
## cvpred       772 1118 982.8 677.5  783 869  665  622 836 1433 1082  819
## Crime        963 1555 929.0 742.0 1216 968  342  373 754 1272  823  508
## CV residual 191  437 -53.8  64.5  433  99 -323 -249 -82 -161 -259 -311
##
## Sum of squares = 795058    Mean square = 66255    n = 12
##
## Overall (Sum over all 12 folds)
```

```
##    ms
## 76417

# Let us calculate the Rsquared error
sse_model4<-76417*nrow(uscrime)
sst_model4<-sum((uscrime$Crime-mean(uscrime$Crime))^2)
rsq_model4<-sse_model4/sst_model4
rsq_model4

## [1] 0.522

summary(lm_model_4)

##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + U2 + Prob, data = uscrime)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -584.0 -136.9  -10.3  110.9  563.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3009.8      992.0   -3.03  0.00418 **
## M              154.3       39.5    3.91  0.00034 ***
## Ed              76.1       46.4    1.64  0.10844
## Po1             93.2       16.2    5.76  9.6e-07 ***
## U2              93.7       50.9    1.84  0.07315 .
## Prob         -2911.6     1889.0   -1.54  0.13091
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 250 on 41 degrees of freedom
## Multiple R-squared:  0.628,  Adjusted R-squared:  0.583
## F-statistic: 13.8 on 5 and 41 DF,  p-value: 6.24e-08
```

Analysis of Model 4:

Sigma (Residual Standard Error (RSA)) = 250
Multiple R-Squared of lm()= 0.628
R-Squared of Model 4 = 0.671
Predicted YHat = 1250 (within the Range of Y: 342 – 1993)

Conclusion:

- To perform the Multiple Linear Regression using least squares, I used the lm() function. Here the crime data set has 15 predictors or factors and the 16th variable "Crime" is the Response.
- I looked at the Regression Model (Model-0) with all the 15 predictors (X)and predicted the Response Y-Hat using the Test data point (X-Hat) that is provided in the question.

  Analysis of Model 0:

  - **Null Hypothesis:** No predictors are significant enough to change the Response Variable.

- **Alternate Hypothesis**: Some or all Predictors add value to change the Response Variable.

1. The lm() Multiple Regression function returned 0.8031 R squared Error and I have chosen the Threshold for Pvalue as 0.1 and concluded that there are about 6 predictors that add value to the response variable.
2. The F Statistic in regression is used along the p-value tells us whether the results are significant enough to reject the null hypothesis. If F is bigger then some Predictor is significant enough to contribute in the regression equation that affects the Response variable.
3. Here P = 3.539e-07 and F > 8 telling us that we can **reject the Null Hypothesis**.
4. The Predicted Crime Value Y-Hat = 155.4349.
   The Range of Response Variable (Y) is lower limit = 342 and Upper Limit = 1993.
   The Y-Hat value is out of range and far below the lower bound of Y.
5. This means that Regression model is probably overfitting the data and needs to be enhanced.
6. Based on this result, the p-value threshold = 0.1 and the following factors "Po2, Po1, Wealth, Ineq, So, Ed, M.f, L.f" exceeded the threshold
7. I used Variance Inflation Factors and created another list with the predictors showing high variance. I have chosen the common factors from both the lists and Then I created 4 models by removing few Predictors in each and analyzed the following.

- Model 1 : I removed "Po2"
- Model 2 : I removed " Po2 ,Wealth,U1,So,NW"
- Model 3: I removed "Po2,Wealth,U1,So,NW,LF,M.F,Pop,Time "
- Model 4: I removed "Po2,Wealth,U1,So,NW,LF,M.F,Pop,Time,Ineq "

Please see the below table and the results:

Model	lm() Removed Factors	P-Value	RSE	F	Multiple R-Squared	Y-Hat
0	None – All 15 are present	3.539e-07	209.1	8.829(15,31)DF	0.8031	155.4349
1	Crime~.-Po2	1.67e-07	208.6313	9.01(14,32)DF	0.797576	724.8202
2	Crime~.-Po2-Wealth-U1-So-NW	6.08e-09	207	12.4(10,36)DF	0.775	1254

3	Crime~.-Po2-Wealth-U1-So-NW LF-M.F-Pop-Time	3.42e-11	201	21.8(6,40)DF	0.766	1304
4	Crime~.-Po2-Wealth-U1-So-NW LF-M.F-Pop-Time -Ineq	6.24e-08	250	13.8(5,41)DF	0.628	1250

As we see from the above, Model-0 Y-Hat is not in the range of Y. Based on the P and F values being greater, null hypothesis is rejected.

As the number of Predictors being removed increased the Multiple R-Squared Error is reduced.

As discussed in the office hours, if there are 10 data points one predictor is a good estimate. As we have 47 data points about 5 predictors would be better to have in the Multiple Regression model.

I would say Model -3 performed the best among all the models with the Better R-Squared. The Predicted Crime Value for the Test data point is 1304 with 201 RSE and a decent R-Squared. I am rejecting Model 4 as the RSE significantly increased to 250 while model 3 has it at 201.

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5040.5      899.8   -5.60  1.7e-06 ***
## M              105.0       33.3    3.15   0.0031 **
## Ed             196.5       44.8    4.39  8.1e-05 ***
## Po1            115.0       13.8    8.36  2.6e-10 ***
## U2              89.4       40.9    2.18   0.0348 *
## Ineq            67.7       13.9    4.85  1.9e-05 ***
## Prob         -3801.8     1528.1   -2.49   0.0171 *
```