HW5.R

harit

2020-09-22

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

## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(car)

## Loading required package: carData

library(DAAG)

## Loading required package: lattice

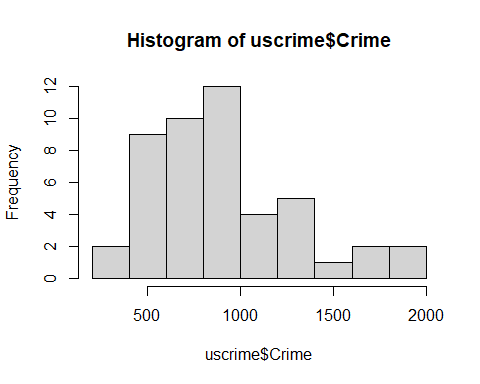
##   
## Attaching package: 'DAAG'

## The following object is masked from 'package:car':  
##   
## vif

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

## M So Ed Po1   
## Min. :11.90 Min. :0.0000 Min. : 8.70 Min. : 4.50   
## 1st Qu.:13.00 1st Qu.:0.0000 1st Qu.: 9.75 1st Qu.: 6.25   
## Median :13.60 Median :0.0000 Median :10.80 Median : 7.80   
## Mean :13.86 Mean :0.3404 Mean :10.56 Mean : 8.50   
## 3rd Qu.:14.60 3rd Qu.:1.0000 3rd Qu.:11.45 3rd Qu.:10.45   
## Max. :17.70 Max. :1.0000 Max. :12.20 Max. :16.60   
## Po2 LF M.F Pop   
## Min. : 4.100 Min. :0.4800 Min. : 93.40 Min. : 3.00   
## 1st Qu.: 5.850 1st Qu.:0.5305 1st Qu.: 96.45 1st Qu.: 10.00   
## Median : 7.300 Median :0.5600 Median : 97.70 Median : 25.00   
## Mean : 8.023 Mean :0.5612 Mean : 98.30 Mean : 36.62   
## 3rd Qu.: 9.700 3rd Qu.:0.5930 3rd Qu.: 99.20 3rd Qu.: 41.50   
## Max. :15.700 Max. :0.6410 Max. :107.10 Max. :168.00   
## NW U1 U2 Wealth   
## Min. : 0.20 Min. :0.07000 Min. :2.000 Min. :2880   
## 1st Qu.: 2.40 1st Qu.:0.08050 1st Qu.:2.750 1st Qu.:4595   
## Median : 7.60 Median :0.09200 Median :3.400 Median :5370   
## Mean :10.11 Mean :0.09547 Mean :3.398 Mean :5254   
## 3rd Qu.:13.25 3rd Qu.:0.10400 3rd Qu.:3.850 3rd Qu.:5915   
## Max. :42.30 Max. :0.14200 Max. :5.800 Max. :6890   
## Ineq Prob Time Crime   
## Min. :12.60 Min. :0.00690 Min. :12.20 Min. : 342.0   
## 1st Qu.:16.55 1st Qu.:0.03270 1st Qu.:21.60 1st Qu.: 658.5   
## Median :17.60 Median :0.04210 Median :25.80 Median : 831.0   
## Mean :19.40 Mean :0.04709 Mean :26.60 Mean : 905.1   
## 3rd Qu.:22.75 3rd Qu.:0.05445 3rd Qu.:30.45 3rd Qu.:1057.5   
## Max. :27.60 Max. :0.11980 Max. :44.00 Max. :1993.0

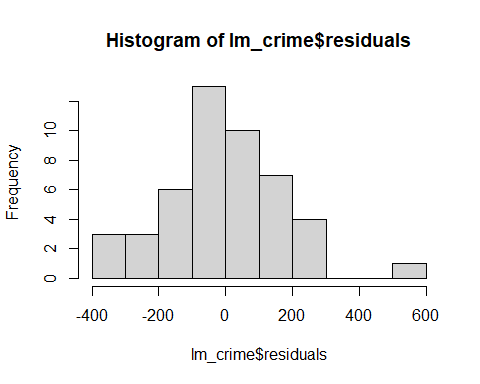
hist(uscrime$Crime)



#For Multiple linear regression model using least squares we can use the lm() function.  
#The crime data set has 15 variables and the 16th variable Crime is used as response.  
  
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

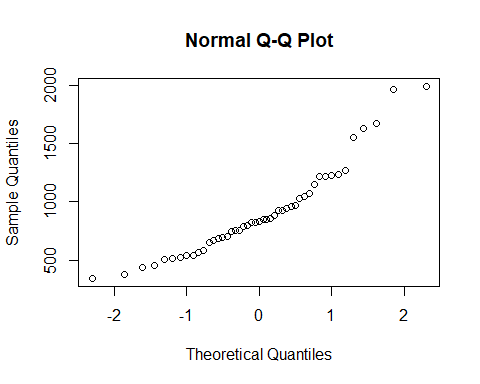
RSquared = summary(lm\_crime)$r.sq  
RSE = summary(lm\_crime)$sigma  
  
hist(lm\_crime$residuals)



#we need to determine if this is normally distributed or not.  
# The residuals are normally distributed with the mean of zero.  
# that is why plotted to see if its true.  
# but shows not so true  
  
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

#the predict value is not in the range of uscrime.   
#it is far below the lower bound. So this shows that our regression model is   
# not the best and may be over fitting rhe data   
  
#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

#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 regression model.  
  
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

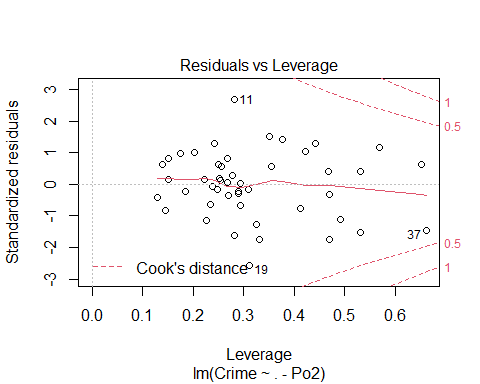
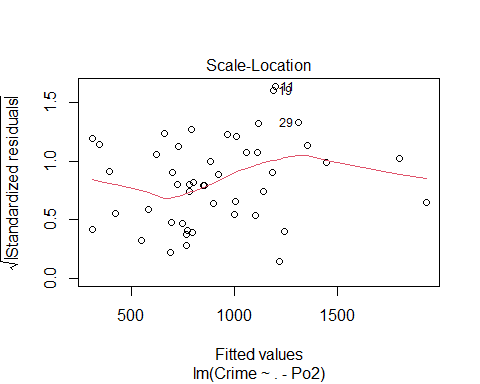
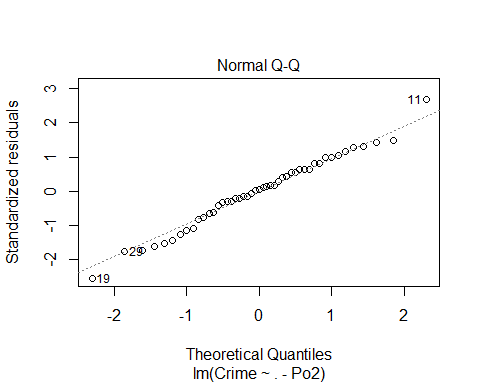
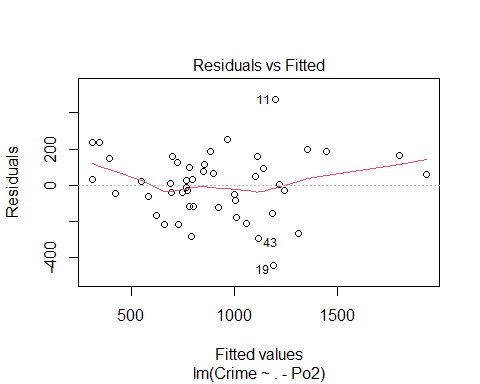
#Predict the crime rate for the data point  
predict\_model\_1 <-predict(lm\_model\_1,testpt)  
predict\_model\_1

## 1   
## 724.8202

range(uscrime$Crime)

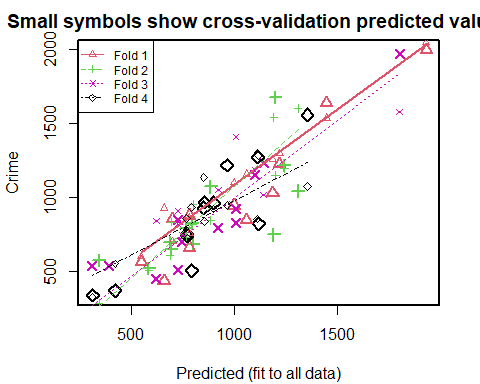
## [1] 342 1993

# The predicted value is within the range  
plot(lm\_model\_1)



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

# Now the predict value is better than the previous . It is in the range.  
  
# This shows that removing the variables that are not necessary actually 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

# 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 and can be removed  
  
# 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

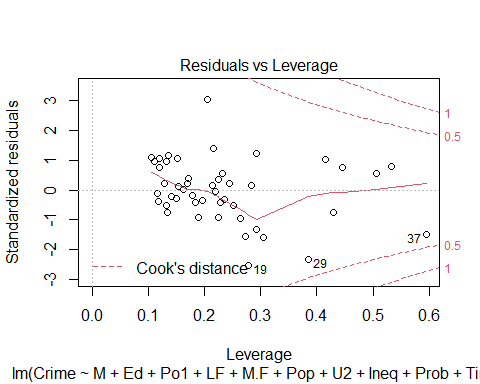
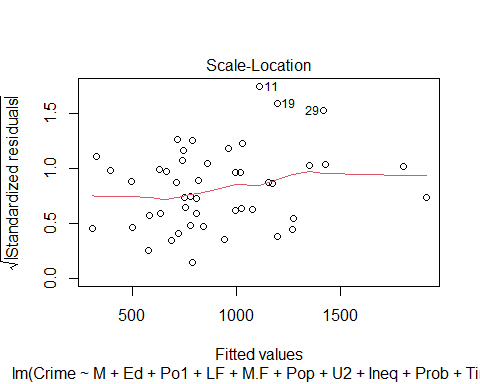
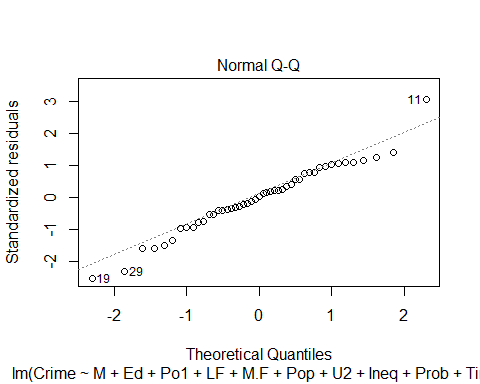
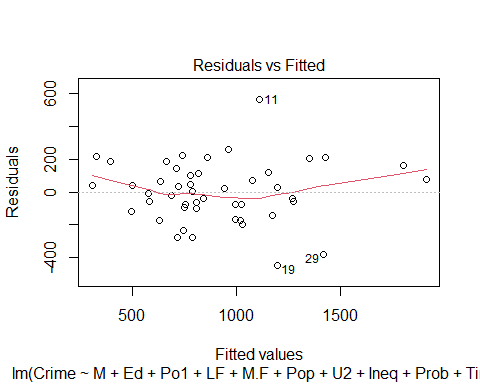
# based on this -We can remove Wealth,Ineq,U1,PO1,SO,U2,NW  
  
#As Wealth,U1, SO,Nw are common in both lists , let us remove those and check the regression model.  
  
  
  
lm\_model\_2 =update(lm\_model\_1,~.-Wealth-U1-So-NW)  
  
  
#Predict the crime rate for the data point  
predict\_model\_2 <-predict(lm\_model\_2,testpt)  
predict\_model\_2

## 1   
## 1254

range(uscrime$Crime)

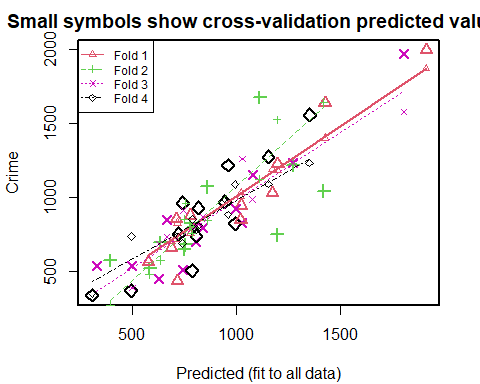
## [1] 342 1993

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

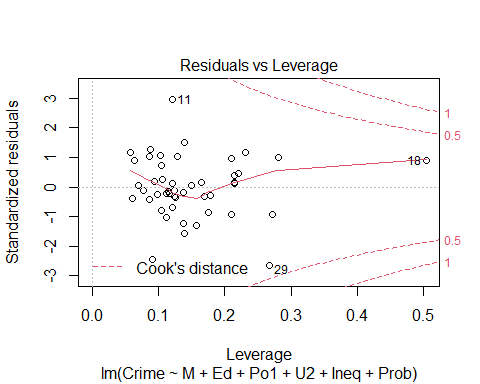
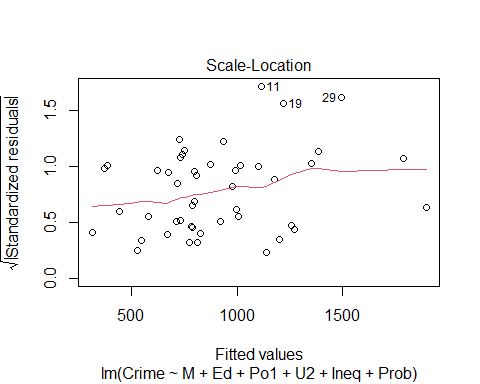
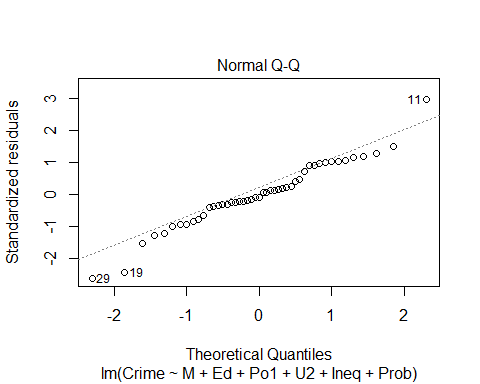
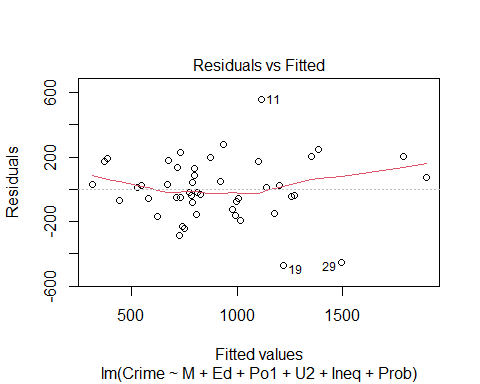
# As we see after removing the factors , the R-squared is little reduced.  
  
#Let is further remove "LF,M.F,Pop,Time"  
  
lm\_model\_3 =update(lm\_model\_2,~.-LF-M.F-Pop-Time)  
  
  
#Predict the crime rate for the data point  
predict\_model\_3 <-predict(lm\_model\_3,testpt)  
predict\_model\_3

## 1   
## 1304

range(uscrime$Crime)

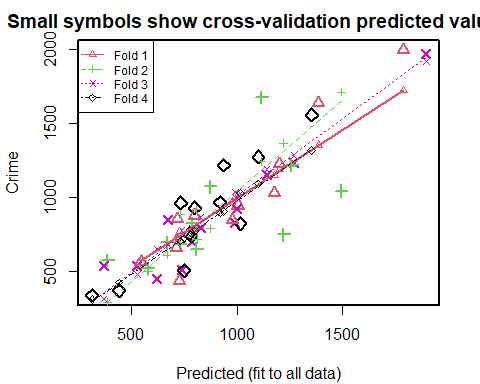
## [1] 342 1993

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

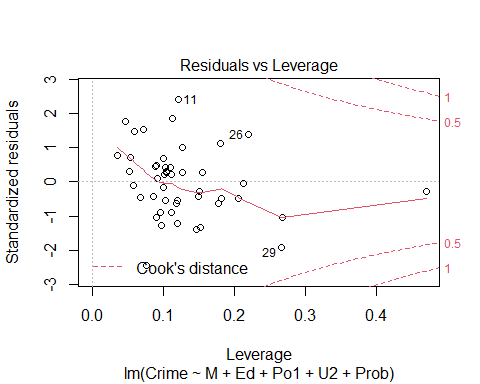
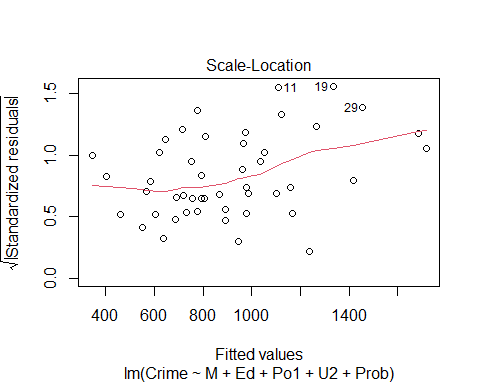
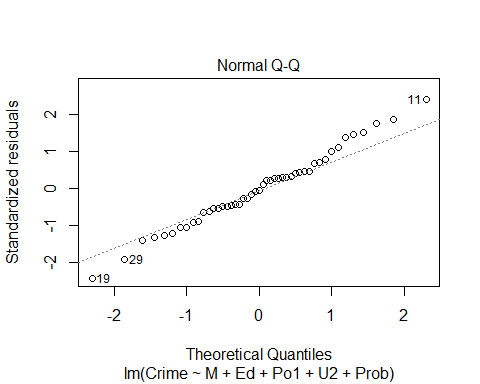
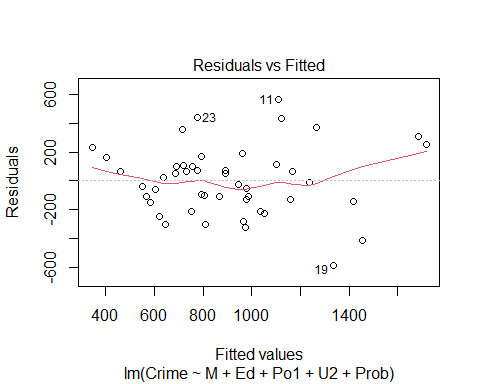
#Let us further remove Ineq and see the results  
lm\_model\_4 =update(lm\_model\_3,~.-Ineq)  
  
  
#Predict the crime rate for the data point  
predict\_model\_4 <-predict(lm\_model\_4,testpt)  
predict\_model\_4

## 1   
## 1250

range(uscrime$Crime)

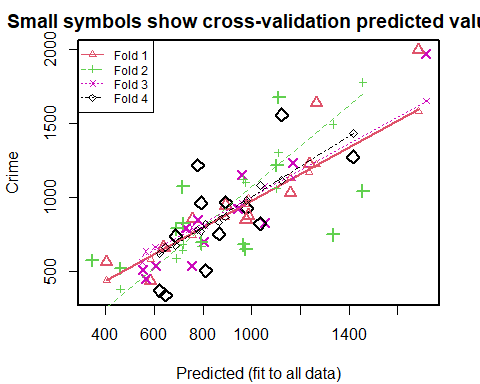
## [1] 342 1993

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