Week3_Linear Regression

HT

6/5/2019

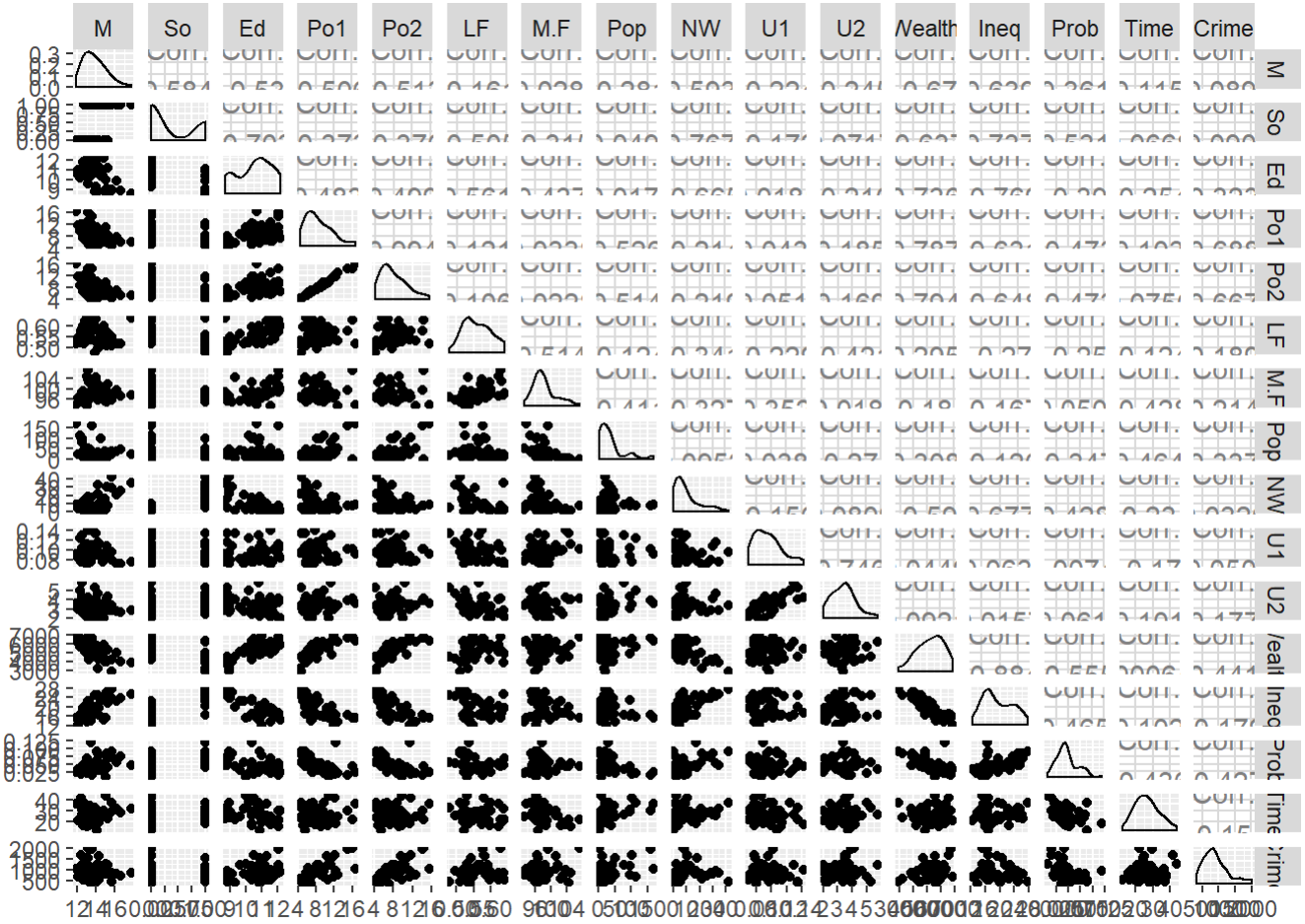
```
library(GGally)
```

```
## Loading required package: ggplot2
```

```
# Reading the file and exploring header
crime_df <- read.table("uscrime.txt", header = TRUE)

head(crime_df)
```

```
##      M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq
## 1 15.1  1  9.1  5.8  5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1
## 2 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4
## 3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0
## 4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7
## 5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4
## 6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6
##       Prob    Time Crime
## 1 0.084602 26.2011   791
## 2 0.029599 25.2999  1635
## 3 0.083401 24.3006   578
## 4 0.015801 29.9012  1969
## 5 0.041399 21.2998  1234
## 6 0.034201 20.9995   682
```

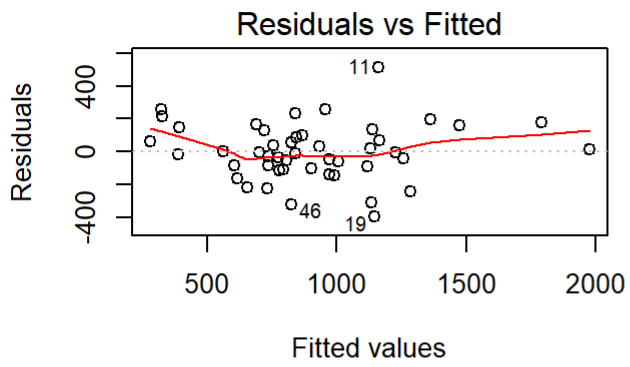
```
ggpairs(crime_df)
```

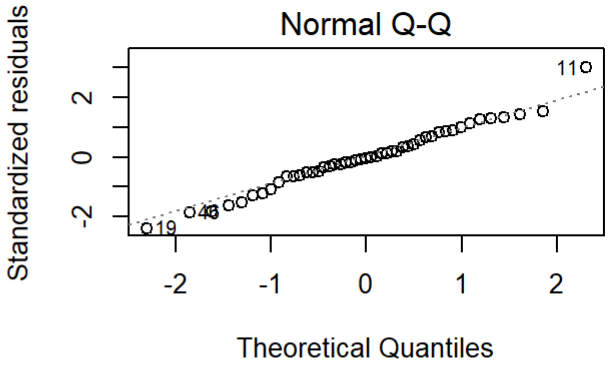


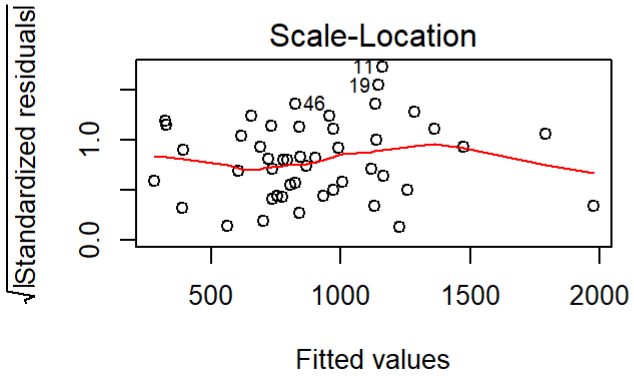

```
# Creating linear model with all columns (except crime) to predict Crime
crime.model <- lm(Crime~M+So+Ed+Po1+Po2+LF+M.F+Pop+NW+U1+U2+Wealth+Ineq+Prob+Time, data = crime_df)
summary(crime.model)
```

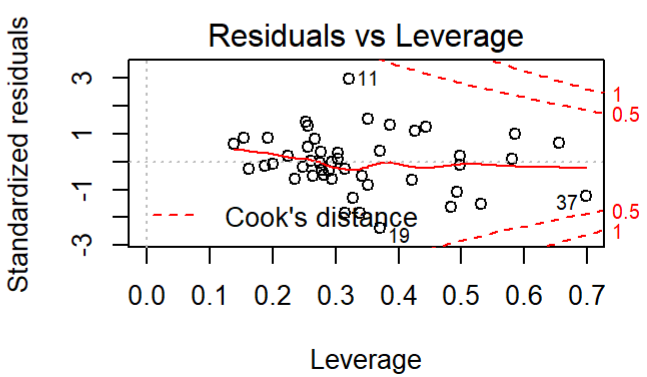
```
##
## Call:
## lm(formula = Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop +
##     NW + U1 + U2 + Wealth + Ineq + Prob + Time, data = crime_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -395.74  -98.09   -6.69  112.99  512.67
##
## Coefficients:
##               Estimate Std. 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
```

```
par(mfrow = c(2,2))
plot(crime.model)
```





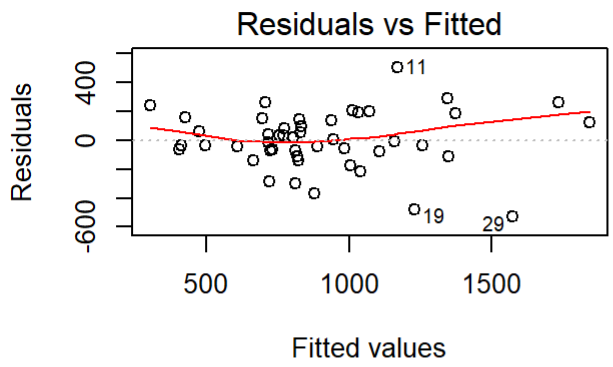


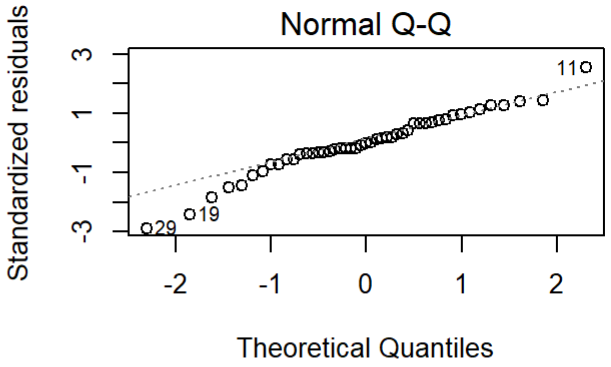


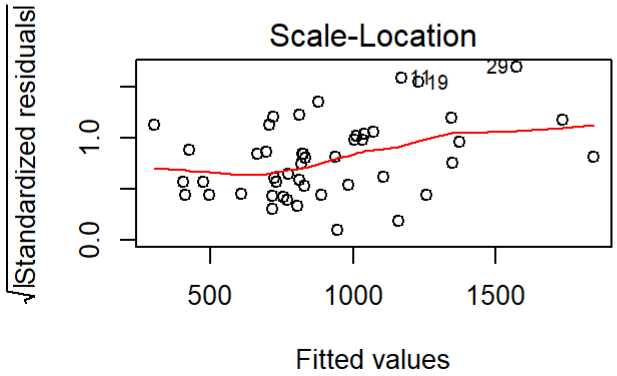

```
#Based on model summary value, create a new model with P-value ess than .05

crime.model.refined <- lm(Crime~M+Ed+Po1+Ineq+Prob,crime_df)
summary(crime.model.refined)
```

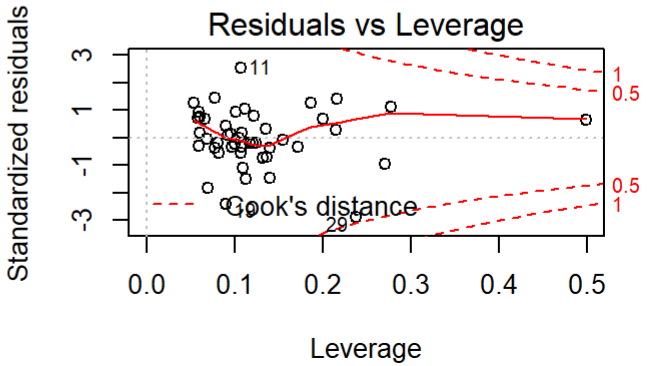
```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + Ineq + Prob, data = crime_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -528.2  -74.0   -7.0  139.8  503.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4064.57     816.28  -4.979 1.20e-05 ***
## M              79.69      32.62   2.443 0.018964 *
## Ed            160.15      43.42   3.688 0.000656 ***
## Po1           121.23      14.06   8.621 9.47e-11 ***
## Ineq           68.31      14.56   4.692 3.00e-05 ***
## Prob        -3867.27    1596.55  -2.422 0.019930 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209.7 on 41 degrees of freedom
## Multiple R-squared:  0.7379, Adjusted R-squared:  0.706
## F-statistic: 23.09 on 5 and 41 DF,  p-value: 5.926e-11
```

```
plot(crime.model.refined)
```

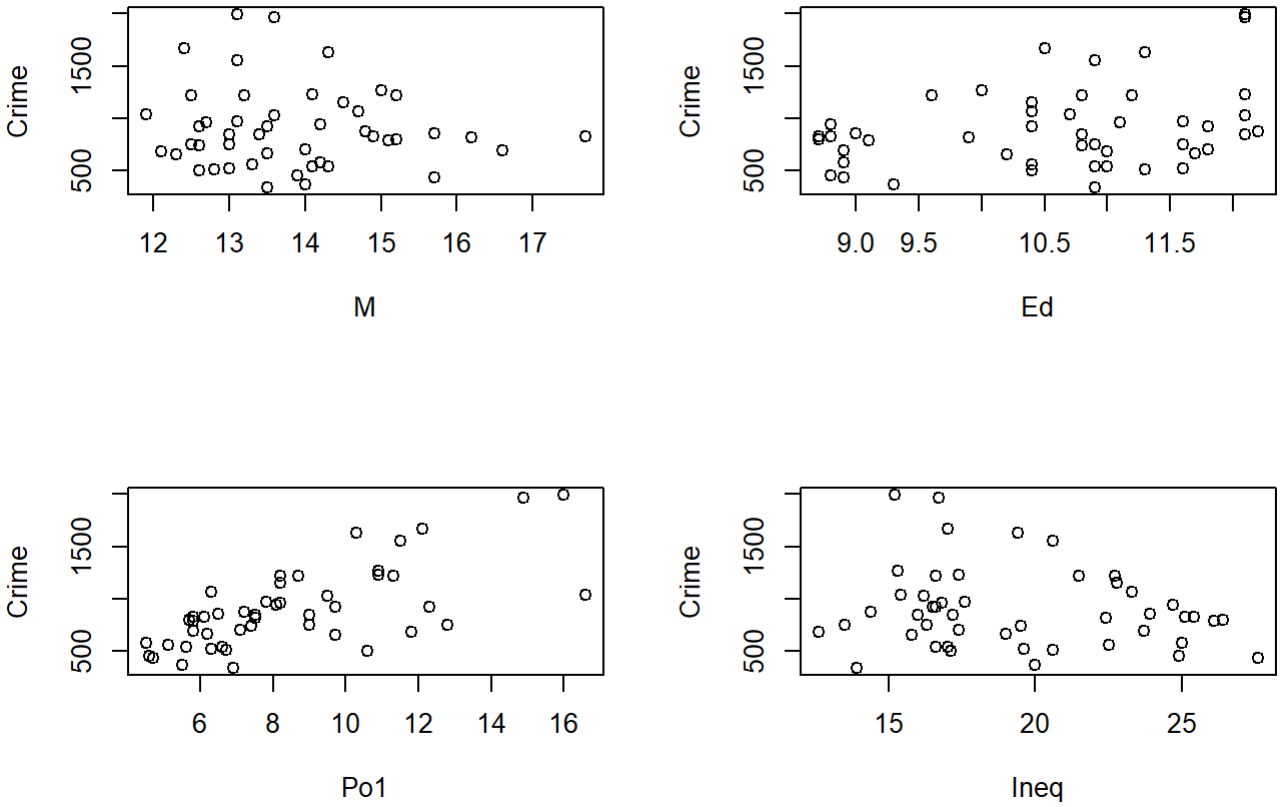









```
# Visualize the distribution, to get an estimated idea of predicted crime value later
plot(Crime~M+Ed+Po1+Ineq+Prob,crime_df)
```




```
# Set dataframe for test purpose
crime.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,We
alth= 3200,Ineq= 20.1,Prob= 0.04,Time = 39.0)

# Predicting Crime using Model 1 (All predictors)
crime.predict.model1 <- predict(crime.model, crime.test)
summary(crime.predict.model1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.4   155.4   155.4   155.4   155.4   155.4
```

```
# Predicting Crime using Model 2 (Refined, Using 5 predictors)
crime.predict.model2 <- predict(crime.model.refined, crime.test)
summary(crime.predict.model2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1326    1326    1326    1326    1326    1326
```

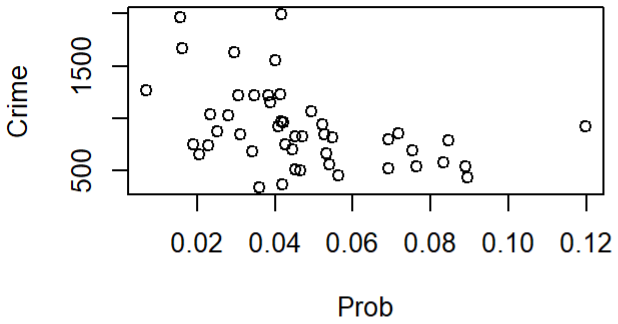
```
library(rsq) # to calculated R-squared values and check the variability in data. I will be using the rsq.lr, R squared value for linear model.

# For Model 1
rsq.lr(crime.model)
```

```
## [1] 0.8030868
```

```
# For Model 2
rsq.lr(crime.model.refined)
```

```
## [1] 0.7379292
```

```
# Although both models shows good variability and a small random/other effects, model 2 looks to be better at predicting the crime rate. As value predicted by Model 1 is very small, and when we look at distribution of data the prediction of Model 2 >1000 looks more reasonable.
```


Crime
500
1500
0.02
0.04
0.06
0.08
0.10
0.12
Prob