HW6
ISYE 6501

Question 9.1

Apply Principal Component Analysis and then create a regression model using the first few principal components. Specify your new model in terms of the original variables (not the principal components), and compare its quality to that of your solution to Question 8.2.

The process for applying PCA to the us crime dataset is as follows:

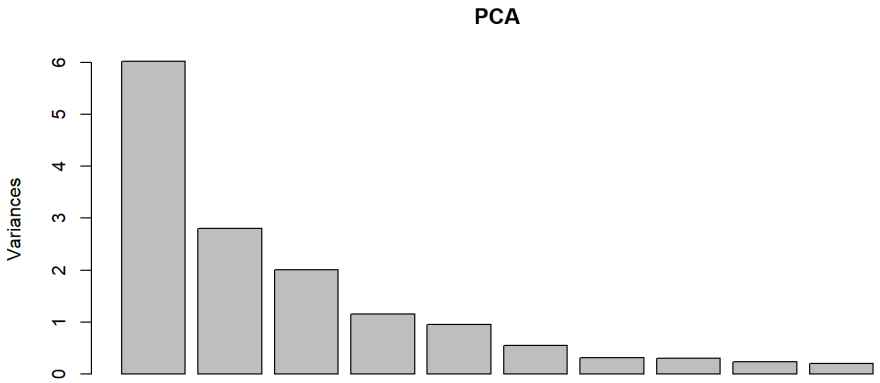
1. Scale the data set, crime_df, done in next step
2. Use prcomp() for the features of the dataset
3. Create a model using lm() for the PCA values
    a. Select first five PCs
4. Reverse the PCA scaling of the data using:
    a. $x_{unscaled} = stdev * x_{scaled} + mean$
    b. Stdev: PCA$center
    c. Mean: PCA$scale
5. Descale the coefficients other than the intercept
6. Use the given values to predict the outcome for the given model
7. Compare with last week's outcome
    a. $1 - \frac{sum(residuals^2)}{(x - avg)^2}$

The model was not successfully completed despite the understanding of the process. By doing Principal Component Analysis, the correlation was removed from the dataset before choosing the best 5 predictors and creating the linear model. The data should then be reverted to its original scaling using the equation provided. The coefficients of the scaled model also should be reverted. The predictor values provided are then input to the model to estimate a prediction value before comparing the models and choosing the best.

1, 2. Scaling the data

> Scaling the input data was solved by setting the logical argument in prcomp() to True. The function returned the Principal Components and their properties

3. Use lm() to create a regression model on the top 5 sorted PCs. The elbow in the component was around PC5.



The p-values for the smaller selection of PCs gave much more statistically significant results.

4. Reverse scaling of data

Using the above equation the data was attempted to be unscaled back to its original value

5. Descale coefficients

```
Coefficients:
(Intercept)          PC1          PC2          PC3          PC4          PC5
     905.09        65.22       -70.08        25.19        69.45      -229.04
```

6. Predict

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

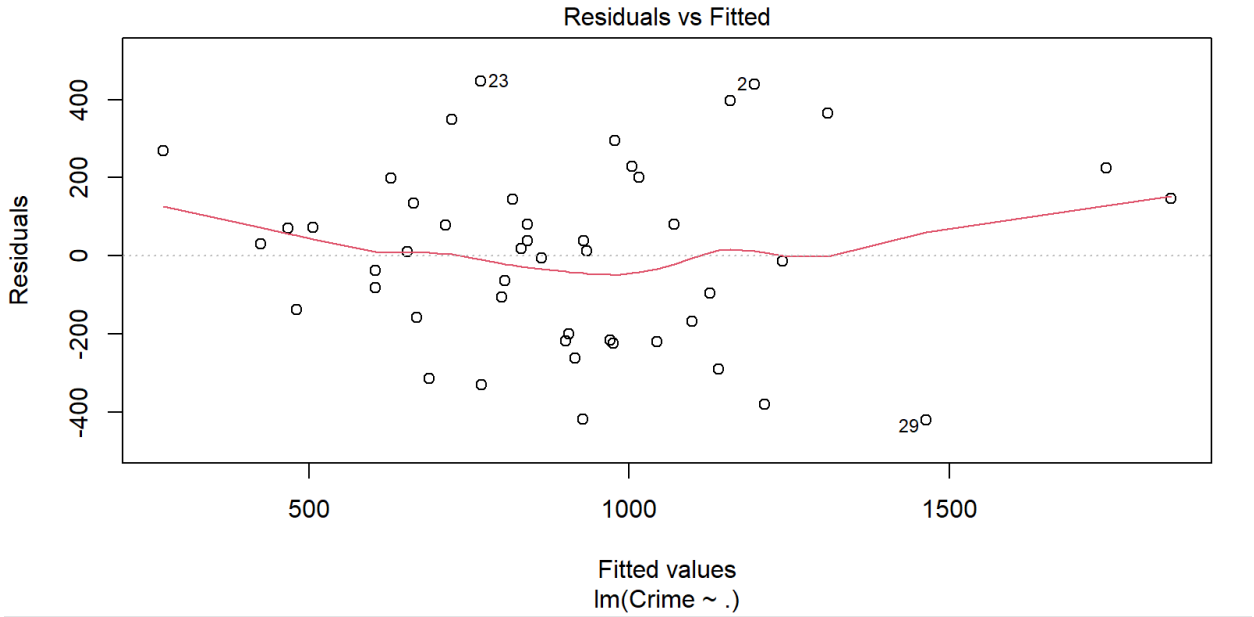
```
Residuals:
    Min      1Q  Median      3Q     Max
-420.79 -185.01   12.21  146.24  447.86

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   905.09      35.59  25.428  < 2e-16 ***
PC1            65.22      14.67   4.447 6.51e-05 ***
PC2           -70.08      21.49  -3.261  0.00224 **
PC3            25.19      25.41   0.992  0.32725
PC4            69.45      33.37   2.081  0.04374 *
PC5          -229.04      36.75  -6.232 2.02e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 244 on 41 degrees of freedom
Multiple R-squared:  0.6452,    Adjusted R-squared:  0.6019
F-statistic: 14.91 on 5 and 41 DF,  p-value: 2.446e-08
```

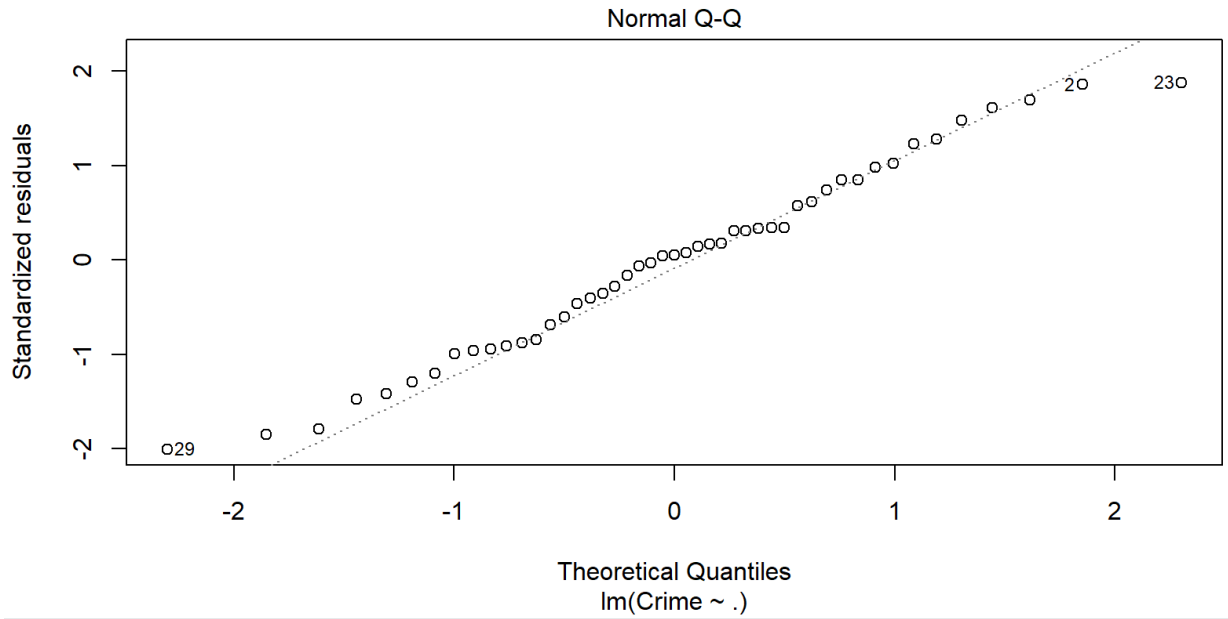
```
> lm_pred
       1        2        3        4        5        6        7        8        9       10       11       12
713.6803 1195.7066  506.4008 1744.8151 1004.3223  901.3083  817.7618 1158.0158  862.6600  906.1942 1309.8473  831.7397
      13       14       15       16       17       18       19       20       21       22       23       24
668.7175  653.8079  663.3242  933.7860  467.7924 1097.8331  975.2212 1238.8452  805.7895  769.6724  768.1369  928.9523
      25       26       27       28       29       30       31       32       33       34       35       36
604.2355 1845.7567  480.4270 1015.0839 1463.7936  801.6455  687.8542  969.6941  722.6822  841.7013  914.9564  977.8353
      37       38       39       40       41       42       43       44       45       46       47
1211.6890  604.2928  627.6148 1069.8938  841.4929  272.2545 1043.4520 1126.3430  425.4541  927.1627 1139.3538
```

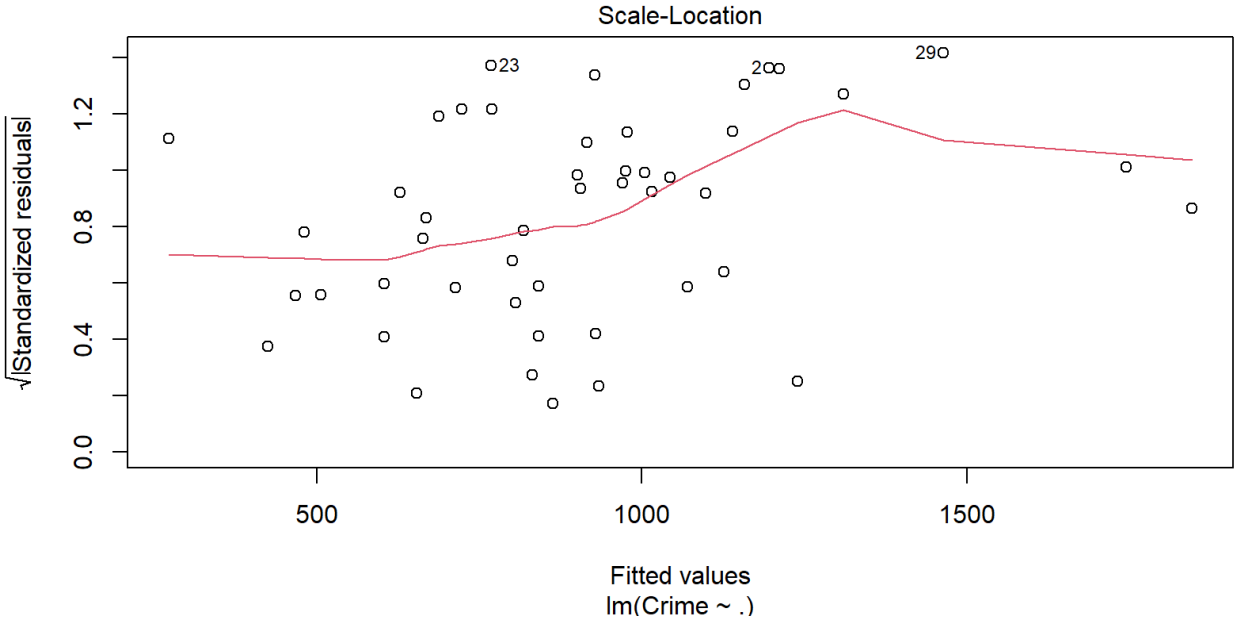
7. Compare
Compare which model fits the data best with the minimum sum of squared errors.

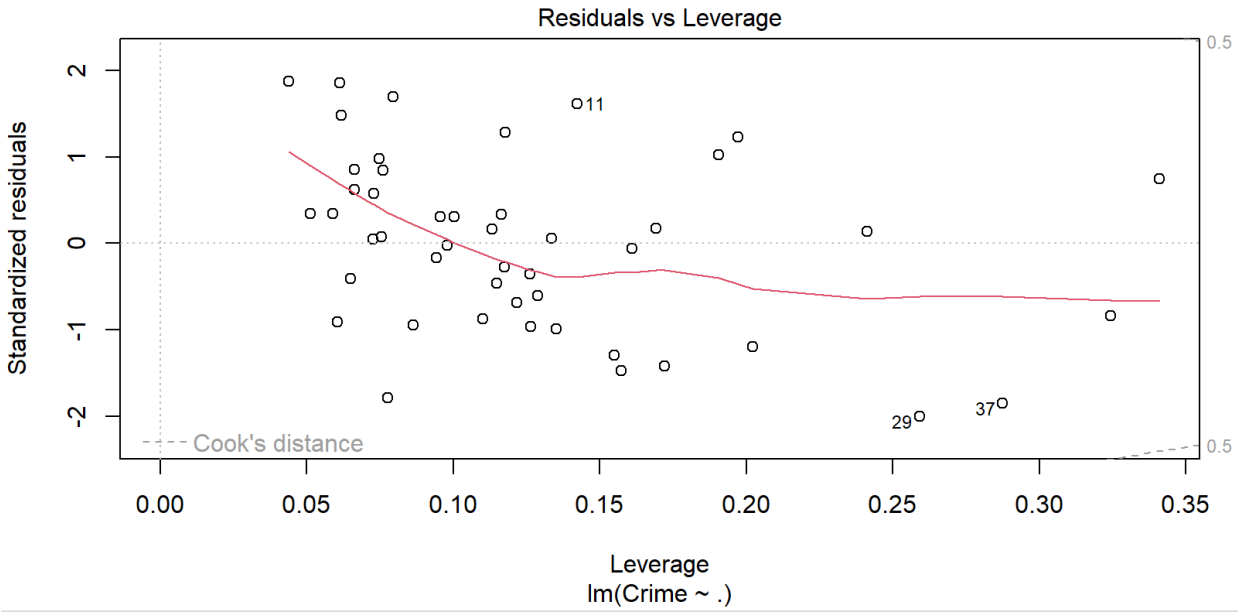
Raw Code in R

```r
# Principal Component Analysis
rm(list=ls())
crime_df <- read.table("uscrime.txt", stringsAsFactors = FALSE, header = TRUE)

PCA <- prcomp(crime_df[1:15], scale=TRUE)
plot(PCA)

selection <- data.frame(PCA$x[,1:5])
selection$Crime <- crime_df$Crime

lin_model <- lm(Crime~., selection)
summary(lin_model)
lin_model
library(Metrics)
lm_pred <- predict(lin_model, selection)
summary(lm_pred)
rmse(actual = selection$Crime, predicted = as.numeric(lm_pred))

center.PCA <- PCA$center
scale.PCA <- PCA$scale
scaled.PCA <- PCA$x

reverse.scale <- as.vector(center.PCA)*as.matrix(scaled.PCA)+as.vector(scale.PCA)
reverse.scale

selection.unscale <- reverse.scale[,1:5]
coef <= as.vector(lin_model$coefficients)
test <- coef * t(selection.unscale)
```




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Crime ~ .)
29
2
23


Scale-Location
√|Standardized residuals|
Fitted values
lm(Crime ~ .)
23
2
29


Residuals vs Leverage
Standardized residuals
2
1
0
-1
-2
11
29
37
Cook's distance
0.5
0.5
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Leverage
lm(Crime ~ .)