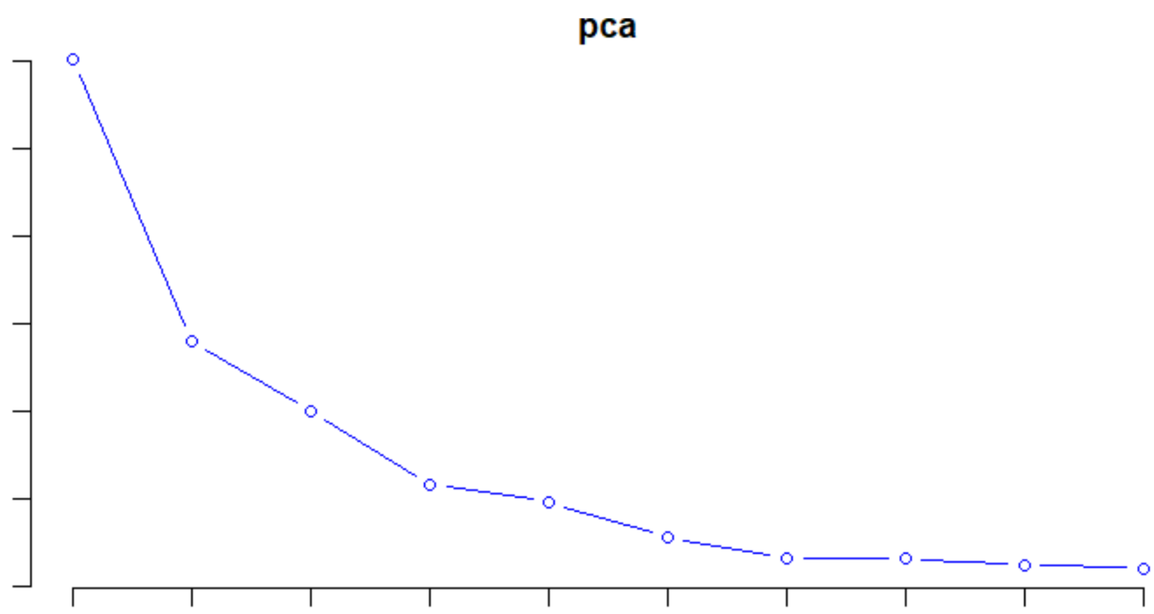
ISYE 6501 HW6

February 2021

1 Question 9.1

Since we are still using the uscrime data, I will not review data again. First, I applied prcomp function to the 15 predictors. This gives a prcomp object, which is the rotated matrix of the data. The ”rotation” element is the eigenvector for the rotation of the data matrix. I will need this again when converting matrix back to original coordination. I have plot the screeplot to see the trend of number of PC components. As figure showed below, the graph recommended a range of 3 to 5.



I fitted a linear regression on PC3 to PC5, PC3 gives an R squared value of 0.2208, PC4 gives an R squared of value 0.2433, and PC5 gives an R squared of value 0.6019. This suggest PC5 gives a better fit. Then, I take the coefficients of the fitted PC5 model, multiply by the first 5 columns from the eigenvector (rotation matrix)

```
Call:
lm(formula = V6 ~ ., data = as.data.frame(pcadata5))

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
```

The result coefficients are listed below, with an intercept of 905.09:

M	So	Ed	Po1	Po2
60.794349	37.848243	19.947757	117.344887	111.450787
LF	M.F	Pop	NW	U1
76.254902	108.126558	58.880237	98.071790	2.866783
U2	Wealth	Ineq	Prob	Time
32.345508	35.933362	22.103697	-34.640264	27.205022

1.1 My R code is:

```
library(GGally)
df <- read.table("uscrime.txt", sep = '\t', stringsAsFactors = FALSE, header = TRUE)

pca <- prcomp(df[,1:15], retx = TRUE, center = TRUE, scale. = TRUE)
summary(pca)

pca$rotation

screeplot(pca, type="lines", col="blue", nps=15)

# since this is ordered by importance, and the screeplot recommends 5 components
pcadf5 <- pca$x[,1:5]
pcadata5 <- cbind(pcadf5, df[,16])
fit5 <- lm(V6~., data=as.data.frame(pcadata5))
summary(fit5)

# try 4 components
pcadf4 <- pca$x[,1:4]
pcadata4 <- cbind(pcadf4, df[,16])
fit4 <- lm(V5~., data=as.data.frame(pcadata4))
summary(fit4)

# try 3 components
pcadf3 <- pca$x[,1:3]
pcadata3 <- cbind(pcadf3, df[,16])
fit3 <- lm(V4~., data=as.data.frame(pcadata3))
summary(fit3)

# PC5 gives the best R squared
beta.Z <- as.matrix(fit$coefficients[2:6])
V <- as.matrix(pca$rotation[,1:5])
beta.X <- V %*% beta.Z
beta.X
```