ISYE 6051 : Homework 6
2/3/2021

## Table of Contents



### 9.1 PCA and Regression

***Using the data from uscrime.txt, I ingested the data into a dataframe and viewed the first 6 records of data. After viewing the data (and wondering if PCA would work properly on the So variable since it is binary and not a continuous variable), I reviewed the summary level information of the PCA analysis on the uscrime data.***

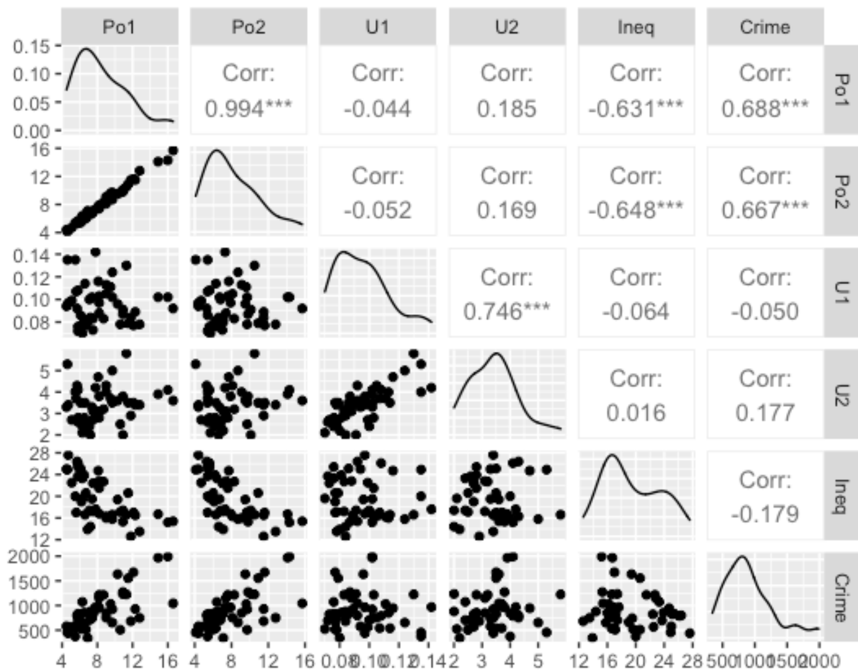
```
# Set the working directory
rm(list = ls())
set.seed(17)
setwd("~/Documents/ISYE6501 Intro to Analytics Modeling/FA_SP_hw6")

#library definition
require("knitr")
library("kernlab")
library("ggplot2")


#Read in Data
crimedf <- read.table("uscrime.txt", header = TRUE)

head(crimedf)
tail(crimedf)

#Perform PCA
compcrime <- prcomp(crimedf[,-16], scale = TRUE)
summary(compcrime)
```

```
> head(crimedf)
```

| | M | So | Ed | Po1 | Po2 | LF | M.F | Pop | NW | U1 | U2 | Wealth | Ineq | Prob | Time | Crime |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 15.11 | 1 | 9.1 | 5.8 | 5.6 | 0.510 | 95.0 | 33 | 30.1 | 0.108 | 4.1 | 3940 | 26.1 | 0.084602 | 26.2011 | 791 |
| 2 | 14.3 | 0 | 11.3 | 10.3 | 9.5 | 0.583 | 101.2 | 12 | 10.2 | 0.096 | 3.6 | 5570 | 19.4 | 0.029599 | 25.2999 | 1635 |
| 3 | 14.2 | 1 | 8.9 | 4.5 | 4.4 | 0.533 | 96.9 | 18 | 21.9 | 0.094 | 3.3 | 3180 | 25.0 | 0.083401 | 24.3006 | 578 |
| 4 | 13.6 | 0 | 12.1 | 14.9 | 14.1 | 0.577 | 99.4 | 157 | 8.0 | 0.102 | 3.9 | 6730 | 16.7 | 0.015801 | 29.9012 | 1969 |
| 5 | 14.1 | 0 | 12.1 | 10.9 | 10.1 | 0.591 | 98.5 | 18 | 3.0 | 0.091 | 2.0 | 5780 | 17.4 | 0.041399 | 21.2998 | 1234 |
| 6 | 12.1 | 0 | 11.0 | 11.8 | 11.5 | 0.547 | 96.4 | 25 | 4.4 | 0.084 | 2.9 | 6890 | 12.6 | 0.034201 | 20.9995 | 682 |

***Viewing correlation graphing data there appears to be a stronger correlation between Po1 and Po2 and a lesser correlation between U1 and U2. I chose these values to display simply because of the correlation in the naming convention. Ineq and Crime are used so that the full values of U1 and U2 are shown.***

```
ggpairs(crimedf, columns = c("Po1", "Po2", "U1", "U2",”Ineq”,”Crime”))
```




***Next, the Principal Component Analysis (PCA) was performed.***

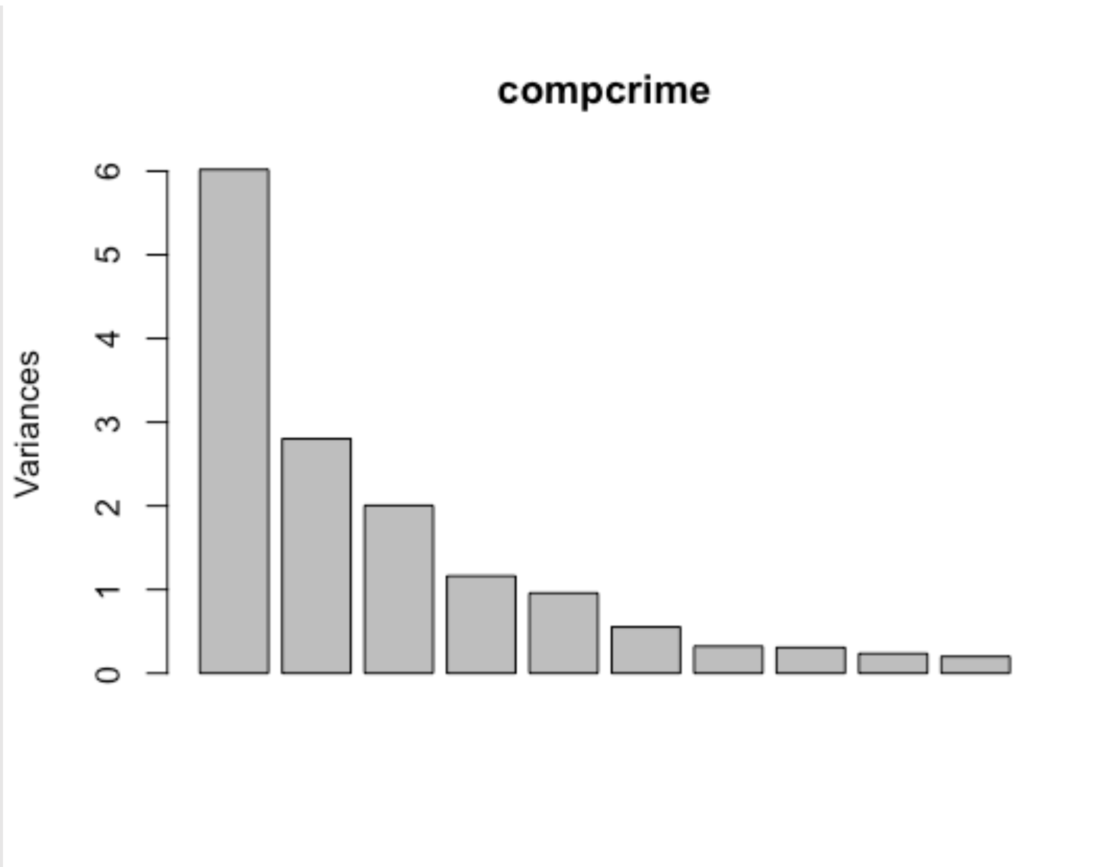
```
#Perform PCA
compcrime <- prcomp(crimedf[,-16], scale = TRUE)
summary(compcrime)
```

```
summary(compcrime)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     2.4534 1.6739 1.4160 1.07806 0.97893 0.74377
PC7
0.56729
Proportion of Variance 0.4013 0.1868 0.1337 0.07748 0.06389 0.03688 0.02145
Cumulative Proportion  0.4013 0.5880 0.7217 0.79920 0.86308 0.89996 0.92142
                          PC8     PC9     PC10    PC11    PC12
Standard deviation     0.55444 0.48493 0.44708 0.41915 0.35804
```

```
PC13       PC14
0.26333 0.2418
Proportion of Variance 0.02049 0.01568 0.01333 0.01171 0.00855 0.00462 0.0039
Cumulative Proportion  0.94191 0.95759 0.97091 0.98263 0.99117 0.99579 0.9997
                          PC15
Standard deviation     0.06793
Proportion of Variance 0.00031
Cumulative Proportion  1.00000
```

***My assumption on the ability for PCA to use binary data held true as only 15 important components were returned. Viewing the data shows that PC1 has a significant variance of 40% and the numbers go down considerably from there; PC2 = 18%, PC3 = 13%, PC4 = 7%.***

***The data is then plotted to provide additional visualization.***



***Eigenvalues for the first 6 principal components.***

```
              PC1         PC2           PC3         PC4         PC5          PC6
M      -0.30371194  0.06280357  0.1724199946 -0.02035537 -0.35832737 -
0.449132706
So     -0.33088129 -0.15837219  0.0155433104  0.29247181 -0.12061130 -
0.100500743
Ed      0.33962148  0.21461152  0.0677396249  0.07974375 -0.02442839 -
0.008571367
Po1     0.30863412 -0.26981761  0.0506458161  0.33325059 -0.23527680 -
0.095776709
Po2     0.31099285 -0.26396300  0.0530651173  0.35192809 -0.20473383 -
0.119524780
LF      0.17617757  0.31943042  0.2715301768 -0.14326529 -0.39407588
0.504234275
M.F     0.11638221  0.39434428 -0.2031621598  0.01048029 -0.57877443 -
0.074501901
Pop     0.11307836 -0.46723456  0.0770210971 -0.03210513 -0.08317034
0.547098563
NW     -0.29358647 -0.22801119  0.0788156621  0.23925971 -0.36079387
0.051219538
U1      0.04050137  0.00807439 -0.6590290980 -0.18279096 -0.13136873
0.017385981
U2      0.01812228 -0.27971336 -0.5785006293 -0.06889312 -0.13499487
0.048155286
Wealth  0.37970331 -0.07718862  0.0100647664  0.11781752  0.01167683 -
0.154683104
Ineq   -0.36579778 -0.02752240 -0.0002944563 -0.08066612 -0.21672823
0.272027031
Prob   -0.25888661  0.15831708 -0.1176726436  0.49303389  0.16562829
0.283535996
Time   -0.02062867 -0.38014836  0.2235664632 -0.54059002 -0.14764767 -
0.148203050
```

***After the first 6 principal components, the variance drops off considerably. Because data with a standard deviation of 1 and mean of 0 clusters around a normal distribution, I will focus on PC1 – PC5 and create a regression model using lm.***

```
regcrime <- cbind(compcrime$x[,1:5],crimedf[,16])
head(regcrime)
model1 <- lm(V6~., data = as.data.frame(regcrime))
summary(model1)
```

```
head(regcrime)
```

```
         PC1        PC2          PC3                PC4                 PC5
[1,] -4.199284 -1.0938312 -1.11907395  0.67178115  0.05528338  791
[2,]  1.172663  0.6770136 -0.05244634 -0.08350709 -1.17319982 1635
[3,] -4.173725  0.2767750 -0.37107658  0.37793995  0.54134525  578
[4,]  3.834962 -2.5769060  0.22793998  0.38262331 -1.64474650 1969
[5,]  1.839300  1.3309856  1.27882805  0.71814305  0.04159032 1234
[6,]  2.907234 -0.3305421  0.53288181  1.22140635  1.37436096  682
> model1 <- lm(V6~., data = as.data.frame(regcrime))
> summary(model1)

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

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 data will now be transformed (alpha & beta) in order to evaluate the use of principal components to determine if the prediction of the Crime rate is comparable to that in homework 8.2.***

```
beta0<-model1$coefficient[1]
betas<-model1$coefficients[2:6]
alphas <- compcrime$rotation[,1:5] %*% betas
alpha2 <- alphas/sapply(crimedf[,1:15], sd)
beta0 <- beta0 - sum(alphas*sapply(crimedf[,1:15], mean)/sapply(crimedf[,1:15], sd))
alphas
```

```
> alphas
           [,1]
M       60.794349
```

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

***The sum of squares is now ready to be evaluated in order to determine the predicted Crime.***

```
estimates <- as.matrix(crimedf[,1:15]) %*% alpha2 + beta2
SSE = sum((estimates - crimedf[,15])^2)
SSTot = sum((crimedf[,15] - mean(crimedf[,15]))^2)
R2 = 1- SSE/SSTot
R2_adjusted = R2-(1-R2)*4/(nrow(crimedf)-4-1)
R2_adjusted

R2_adjust <- R2 - (1-R2)*5/(nrow(crimedf)-5-1)
R2_adjust

testdata <- 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 = 0.040,Time = 39.0)

pred_df <- data.frame(predict(compcrime, testdata))

pred <- predict(model1, pred_df)
```

```
R2_adjusted
[1] -19279.34
> R2_adjust <- R2 - (1-R2)*5/(nrow(crimedf)-5-1)
> R2_adjust
[1] -19749.59
> testdata <- 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 = 0.040,Time = 39.0)
> pred_df <- data.frame(predict(compcrime, testdata))
> pred <- predict(model1, pred_df)
```

```
> pred
      1
1388.926
```

***Crime using the principal components of PCA is 1389 (rounded). The value from homework 8.2 was 1304. Since there wasn't a significant amount of data for either model while they appear to be similar in result, I cannot say definitively that either one is better than the other.***