HW6

2022-10-05

The first step is to set up the enviroment and load the data.

```
#Cealr environment
rm(list=ls())
set.seed(33)

#Load Data
data_df<- read.table("/uscrime.txt",header = T)
```

I then applied scaling to the predictors using the R-function prcomp(). The output summary is depicted in Figure 1 below. The Proportion of Variance for each Principle Component, or how much of the data each factor explains, may be seen from the summary.

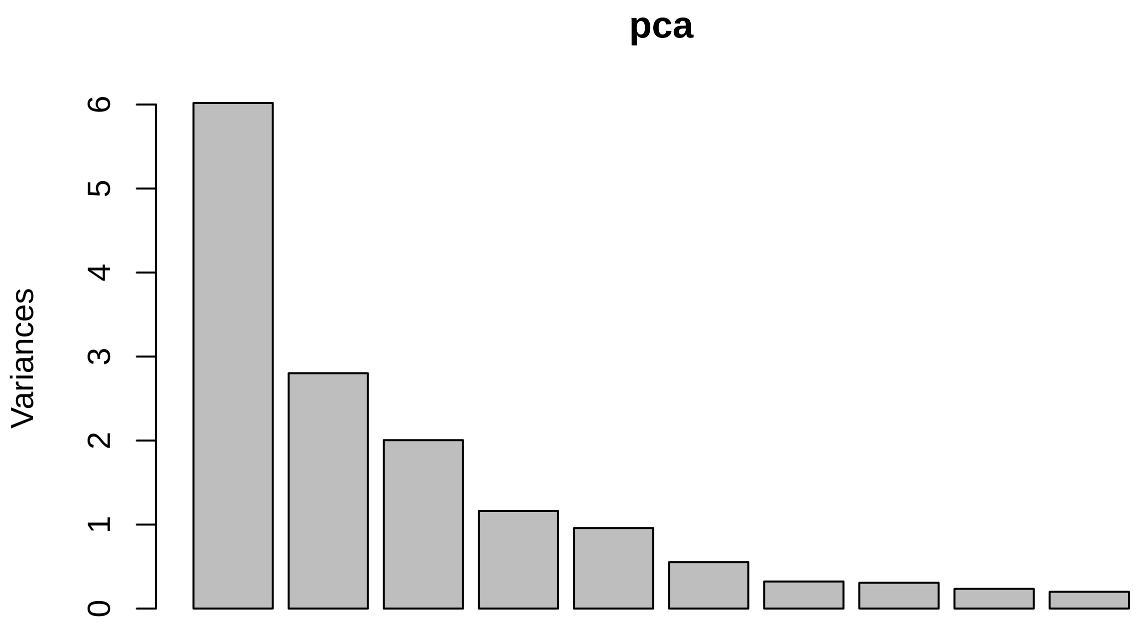
```
#Perform PCA
pca<-prcomp(data_df[,1:15], scale= T)
summary (pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.4534 1.6739 1.4160 1.07806 0.97893 0.74377 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    PC13   PC14
## Standard deviation     0.55444 0.48493 0.44708 0.41915 0.35804 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
```

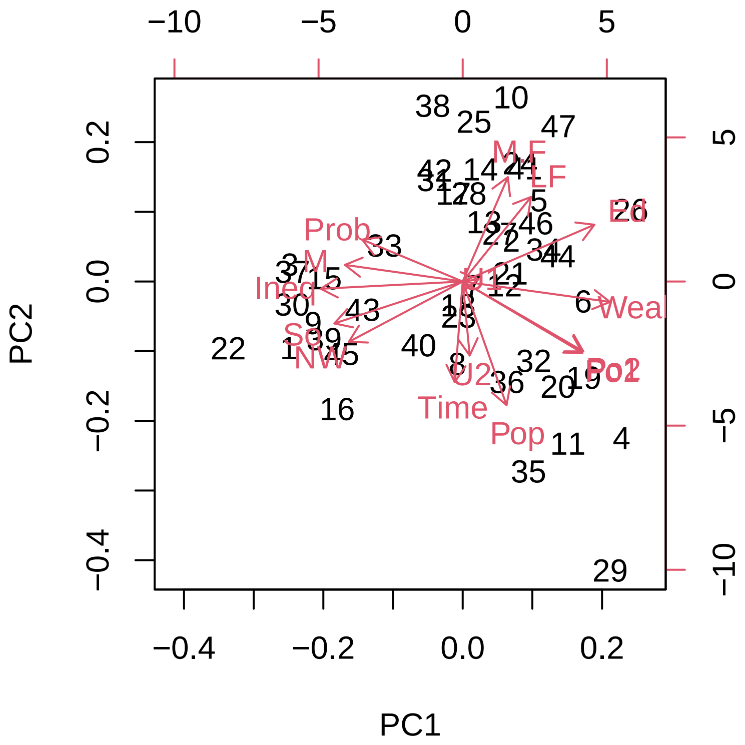
To determine the number of PCs to include in my regression model, I first produced a scree plot. We may clearly observe a diminishing return by graphing the proportion of variances against the PC number on the scree plot. The summary indicates that the first 7 PCs account for 92.1% of the values, therefore I'll pick those. Other values are possible, but in my opinion, this is where the curve starts to noticeably flatten down.

I created a biplot of the Eigen vectors of the first two PCs to better comprehend their fundamental structure. We can observe from the biplot that PC1 is probably a function of Wealth, Ineq, M, and So, and PC2 is probably a function of Time, Pop, M.F, and L.F. These have the highest differences in those specific scales and are therefore the easiest to observe because they are most parallel to the axes.

```
#Visualize PCA
plot(pca)
```



```
biplot(pca)
```



In order to use a new dataset for the linear regression model, I generate the PCA, extract the first four main components

```
#Extract First 4 principal components
pca_df<-data.frame(cbind(pca$x[,1:4],data_df$Crime))
names(pca_df)<-c('PC1','PC2','PC3','PC4','Crime')
```

Having created a new pca dataset, I create a model using lm():

```
model_pca<-lm(Crime~.,pca_df)
summary(model_pca)
```

```
##
## Call:
## lm(formula = Crime ~ ., data = pca_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -557.76 -210.91  -29.08  197.26  810.35
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   905.09      49.07  18.443  < 2e-16 ***
## PC1            65.22      20.22   3.225  0.00244 **
## PC2           -70.08      29.63  -2.365  0.02273 *
## PC3            25.19      35.03   0.719  0.47602
## PC4            69.45      46.01   1.509  0.13872
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 336.4 on 42 degrees of freedom
## Multiple R-squared:  0.3091, Adjusted R-squared:  0.2433
## F-statistic: 4.698 on 4 and 42 DF,  p-value: 0.003178
```

Next, convert the model_pca coefficients back to the original factors:

```
#convert
coefficients_converted <- (pca$rotation[,1:4]%*% model_pca$coefficients[2:5])/pca$scale

# Adject intercept based on pca$center
intercept<-model_pca$coefficients[1]-sum(coefficients_converted* pca$center)
```

With the original factors, try to predict the crime

```
#New data point that we'll predict Crime for
new_dp <- 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 )
# Manually calculate Crime for new_dp using coefficients_coverted and intercept
Crime<-sum(
coefficients_converted[1,1]%*%new_dp$M,
coefficients_converted[2,1]%*%new_dp$So,
coefficients_converted[3,1]%*%new_dp$Ed,
coefficients_converted[4,1]%*%new_dp$Po1,
coefficients_converted[5,1]%*%new_dp$Po2,
coefficients_converted[6,1]%*%new_dp$LF,
coefficients_converted[7,1]%*%new_dp$M.F,
coefficients_converted[8,1]%*%new_dp$Pop,
coefficients_converted[9,1]%*%new_dp$NW,
coefficients_converted[10,1]%*%new_dp$U1,
coefficients_converted[11,1]%*%new_dp$U2,
coefficients_converted[12,1]%*%new_dp$Wealth, coefficients_converted[13,1]%*%new_dp$Ineq,
coefficients_converted[14,1]%*%new_dp$Prob,
coefficients_converted[15,1]%*%new_dp$Time,
intercept
)
```

```
Crime
```

```
## [1] 1112.678
```

The HW5-Q2 model projected Crime for the new dp to be 1,304, while the PCA model forecasts Crime to be 1,113. The performance of the new pca-based model was significantly inferior than that of the model described in HW5 Q2, according to Adjusted R2.