Module 4 Assignment

Benjan

10/31/2022

Lets read the file from github.

urlRemote <- "https://raw.githubusercontent.com/"  
pathGithub <- "EricBrownTTU/ISQS6350/main/"  
filename <- "crime.csv"  
crime <- read.csv(paste0(urlRemote, pathGithub, filename))

# 1.

Lets conduct a principle component analysis on the covariance matrix.

s <- cov(crime[-1]) # the first column is not numeric  
crime\_pca <- princomp(covmat=s)

Now, we can see what proportion of variance are explained by the components.

summary(crime\_pca)

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4  
## Standard deviation 819.8170152 252.30867827 155.61516542 79.457309918  
## Proportion of Variance 0.8735949 0.08274478 0.03147604 0.008206235  
## Cumulative Proportion 0.8735949 0.95633968 0.98781572 0.996021957  
## Comp.5 Comp.6 Comp.7  
## Standard deviation 54.933145307 6.097760e+00 2.381806e+00  
## Proportion of Variance 0.003922339 4.833002e-05 7.373758e-06  
## Cumulative Proportion 0.999944296 9.999926e-01 1.000000e+00

As we can see in the ‘cumulative proportion’ that 87 percent of the variability is explained by the first component, the first component represents a reasonable proportion of variability.

Let us check this with the mean variance. We will consider the component only if its variance is greater than mean variance.

crime\_pca$sdev^2 > mean(crime\_pca$sdev^2)

## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7   
## TRUE FALSE FALSE FALSE FALSE FALSE FALSE

Our decision is confirmed by preceding calculation.

# 2.

The first principle component is:

crime\_pca$loadings[,1]

## MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY   
## 0.0008640289 0.0087729919 0.0569930688 0.0591963208 0.4653457248 0.8728626273   
## AUTO   
## 0.1213839527

As we can see, the first principle component is most influenced by **Burglary** and **Larceny** and it is a weighted sum of these two variables.

crime\_pca$loadings[,2]

## MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY   
## 0.007076731 0.011476788 0.165921433 0.174242756 0.774438600 -0.481781175   
## AUTO   
## 0.331752162

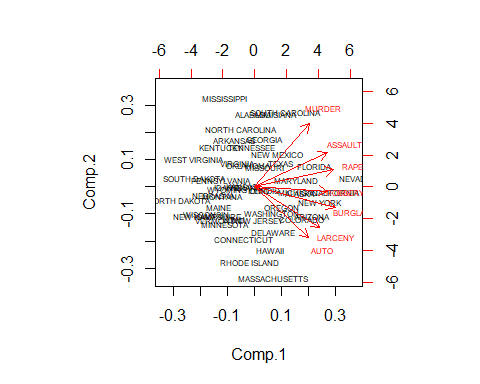
The second principle component is most influenced by **Burglary** and **Larceny** and **Auto** with **Larceny** contributing in opposite direction.

# 3.

First, let’s set the rownames of the dataset by the state name.

rownames(crime) <- crime[,1]

crime.pca <- princomp(crime[,-1],cor=TRUE)  
biplot(crime.pca, col = c("black","red"), cex = 0.5)



From the biplot, MISSISSIPPI has strong second component, NEVADA has strong first component and HAWAII has very low second component. Similarly, NEVADA is towards the RAPE arrow, so it must have a high RAPE score. The AUTO and LARCENCY arrows point opposite to MISSISSIPPI, so it should have very low values in these variables.

Lets look at the z-scores for validation.

round(scale(crime[,-1]),2)

## MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY AUTO  
## ALABAMA 1.75 -0.05 -0.31 0.67 -0.36 -1.09 -0.50  
## ALASKA 0.87 2.40 -0.31 0.73 0.09 0.96 1.94  
## ARIZONA 0.53 0.79 0.16 1.01 2.44 2.47 0.32  
## ARKANSAS 0.35 0.17 -0.46 -0.08 -0.74 -1.11 -1.00  
## CALIFORNIA 1.05 2.20 1.84 1.46 1.96 1.14 1.48  
## COLORADO -0.30 1.51 0.53 0.81 1.49 1.70 0.51  
## CONNECTICUT -0.84 -0.83 0.06 -0.79 0.13 -0.07 1.12  
## DELAWARE -0.37 -0.08 0.37 -0.17 0.90 1.39 0.46  
## FLORIDA 0.71 1.29 0.72 2.37 1.31 1.61 -0.14  
## GEORGIA 1.10 0.50 0.19 0.45 0.14 -0.69 -0.41  
## HAWAII -0.06 -0.02 0.04 -1.47 1.43 1.72 0.58  
## IDAHO -0.50 -0.59 -0.96 -0.39 -0.56 -0.10 -0.72  
## ILLINOIS 0.64 -0.37 0.99 -0.02 -0.48 0.22 0.78  
## INDIANA -0.01 0.07 -0.01 -0.58 -0.48 -0.24 0.00  
## IOWA -1.33 -1.41 -0.94 -1.21 -1.11 0.02 -0.82  
## KANSAS -0.22 -0.35 -0.26 -0.31 -0.05 0.09 -0.69  
## KENTUCKY 0.69 -0.62 -0.49 -0.88 -0.97 -1.39 -0.68  
## LOUISIANA 2.08 0.48 0.21 1.24 -0.29 -0.28 -0.21  
## MAINE -1.30 -1.14 -0.97 -0.41 -0.09 -0.44 -0.68  
## MARYLAND 0.14 0.84 1.90 1.47 0.25 0.70 0.26  
## MASSACHUSETTS -1.12 -0.46 0.51 0.20 0.56 -0.50 3.94  
## MICHIGAN 0.48 1.22 1.56 0.63 0.53 0.67 0.87  
## MINNESOTA -1.23 -0.58 -0.43 -1.25 -0.36 -0.15 -0.18  
## MISSISSIPPI 1.77 -0.57 -0.66 -0.22 -0.87 -1.97 -1.21  
## MISSOURI 0.56 0.24 0.73 0.22 0.06 -0.34 0.00  
## MONTANA -0.53 -0.84 -0.96 -0.54 -1.13 0.14 -0.35  
## NEBRASKA -0.92 -0.71 -0.67 -0.98 -1.23 -0.49 -0.66  
## NEVADA 2.16 2.17 2.25 1.43 2.69 2.12 0.94  
## NEW HAMPSHIRE -1.10 -1.40 -1.14 -1.35 -0.58 -0.45 -0.43  
## NEW JERSEY -0.48 -0.44 0.64 -0.26 0.33 0.14 0.69  
## NEW MEXICO 0.35 1.24 -0.16 1.32 0.29 0.46 -0.61  
## NEW YORK 0.84 0.34 3.94 1.08 1.01 0.15 1.90  
## NORTH CAROLINA 0.82 -0.81 -0.71 1.07 -0.32 -0.87 -0.96  
## NORTH DAKOTA -1.69 -1.56 -1.25 -1.67 -1.96 -1.14 -1.20  
## OHIO 0.09 0.15 0.75 -0.30 -0.18 0.04 0.12  
## OKLAHOMA 0.30 0.32 -0.57 -0.06 -0.01 -0.61 -0.26  
## OREGON -0.66 1.32 0.00 0.75 0.80 1.15 0.06  
## PENNSYLVANIA -0.48 -0.63 0.07 -0.83 -0.96 -1.44 -0.23  
## RHODE ISLAND -0.99 -1.42 -0.43 -0.10 0.46 0.24 2.14  
## SOUTH CAROLINA 1.15 0.68 -0.21 2.73 0.74 -0.45 -0.68  
## SOUTH DAKOTA -1.41 -1.14 -1.20 -0.55 -1.67 -1.33 -1.19  
## TENNESSEE 0.69 0.37 0.25 -0.07 -0.07 -1.23 -0.33  
## TEXAS 1.51 0.75 0.32 -0.03 0.72 0.44 0.10  
## UTAH -1.02 -0.51 -0.63 -0.64 -0.28 0.46 -0.22  
## VERMONT -1.56 -0.91 -1.06 -1.10 0.13 -0.65 -0.58  
## VIRGINIA 0.40 -0.23 -0.36 -0.45 -0.71 -0.21 -0.78  
## WASHINGTON -0.81 1.29 -0.20 0.13 0.73 0.99 -0.09  
## WEST VIRGINIA -0.37 -1.16 -0.93 -1.20 -1.61 -1.83 -1.11  
## WISCONSIN -1.20 -1.19 -0.81 -1.47 -1.03 -0.08 -0.81  
## WYOMING -0.53 -0.36 -0.96 -0.37 -1.11 0.14 -0.49  
## attr(,"scaled:center")  
## MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY AUTO   
## 7.444 25.734 124.092 211.300 1291.904 2671.288 377.526   
## attr(,"scaled:scale")  
## MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY AUTO   
## 3.866769 10.759630 88.348567 100.253049 432.455711 725.908707 193.394418

We can see that our observations from the biplots are validated by the z-scores.