Assignment 8

Dipesh Poudel

1/5/2022

# Assignment 8

**Use the built-in “mtcars” data of R and do as follows:**  1. Check the data with head(mtcars) and save a new data as mtcars.subset after dropping two non-numeric (binary) variables for PCA analysis

head(mtcars)

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

Removing the column with binary values

mtcars.subset<-subset(mtcars,select = -c(vs,am))

2.Fit PCA in the as mtcars.pca matcars.subset data with cor = TRUE and scores = TRUE)

mtcars.pca<-princomp(mtcars.subset,cor=TRUE,scores = TRUE)

1. Get summary of mtcars.pca and interpret standard deviation, proportion of variance carefully

summary(mtcars.pca)

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Standard deviation 2.3782219 1.4429485 0.71008086 0.5148082 0.42797037  
## Proportion of Variance 0.6284377 0.2313445 0.05602387 0.0294475 0.02035096  
## Cumulative Proportion 0.6284377 0.8597822 0.91580607 0.9452536 0.96560453  
## Comp.6 Comp.7 Comp.8 Comp.9  
## Standard deviation 0.3518426 0.32413257 0.241896155 0.148964367  
## Proportion of Variance 0.0137548 0.01167355 0.006501528 0.002465598  
## Cumulative Proportion 0.9793593 0.99103287 0.997534402 1.000000000

1. Get eigenvalue of the components using standard deviation of mtcars.pca and chose the number of components based on Kaiser’s criteria

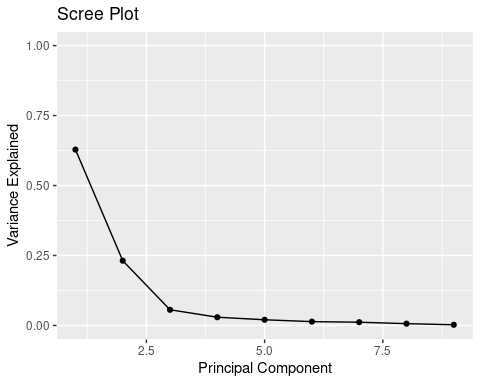
# Eigenvalues are square of standard deviation  
(eigenvalues<-mtcars.pca$sdev^2)

## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7   
## 5.65593947 2.08210029 0.50421482 0.26502753 0.18315864 0.12379319 0.10506192   
## Comp.8 Comp.9   
## 0.05851375 0.02219038

According to the Kaiser’s rule, PC with Eigenvalue >= 1 (SD-square) must be used/retained for the latent variable. So, PC1 and PC1 are two components that we should retain.

1. Get scree plot and chose the number of components best on “first bend” of this plot

library(ggplot2)  
# calculate total variance explained by each principal component  
var\_explained<-mtcars.pca$sdev^2/sum(mtcars.pca$sdev^2)  
  
# Scree Plot  
  
qplot(c(1:9),var\_explained)+geom\_line()+xlab("Principal Component")+ylab("Variance Explained")+ggtitle("Scree Plot")+ylim(0,1)



In the plot, the first bend is at 2 so 2 is the number of components to choose.

1. Write how many components must be retained based on Kaiser’s rule and/or scree plot Based on the scree plot and Kaiser’s rule we have to retain two(2) components.
2. Fit the final PCA model based on the retained components and interpret it carefully

library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

final\_pca<-psych::principal(mtcars.subset,nfactors = 2,rotate = "none")

final\_pca

## Principal Components Analysis  
## Call: psych::principal(r = mtcars.subset, nfactors = 2, rotate = "none")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## PC1 PC2 h2 u2 com  
## mpg -0.93 0.04 0.88 0.124 1.0  
## cyl 0.96 0.02 0.92 0.083 1.0  
## disp 0.94 -0.13 0.91 0.091 1.0  
## hp 0.87 0.39 0.91 0.087 1.4  
## drat -0.74 0.49 0.79 0.207 1.7  
## wt 0.89 -0.25 0.85 0.150 1.2  
## qsec -0.53 -0.70 0.77 0.227 1.9  
## gear -0.50 0.79 0.88 0.120 1.7  
## carb 0.58 0.70 0.83 0.173 1.9  
##   
## PC1 PC2  
## SS loadings 5.66 2.08  
## Proportion Var 0.63 0.23  
## Cumulative Var 0.63 0.86  
## Proportion Explained 0.73 0.27  
## Cumulative Proportion 0.73 1.00  
##   
## Mean item complexity = 1.4  
## Test of the hypothesis that 2 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.05   
## with the empirical chi square 6.04 with prob < 1   
##   
## Fit based upon off diagonal values = 0.99

The first variable can explain 63% and the second variable can explain 23% variance. These both can explain 86% of the variance.

1. Get the head of the saved loadings of mtcars.pca and interpret the values carefully

head(final\_pca$loadings)

## PC1 PC2  
## mpg -0.9349924 0.03973679  
## cyl 0.9573620 0.02266837  
## disp 0.9449932 -0.12825603  
## hp 0.8730011 0.38875010  
## drat -0.7415688 0.49298721  
## wt 0.8882114 -0.24810498

The head shows only six rows of data.

The loadings variable provides the coefficients of linear combination to compute pc1 and pc2 from the data. Positive loadings values indicates that the variable and the component are positively correlated. Negative loadings values indicate a negative correlation between the variable and the component.

1. Retain two components, get their loadings and interpret them carefully

final\_pca$loadings

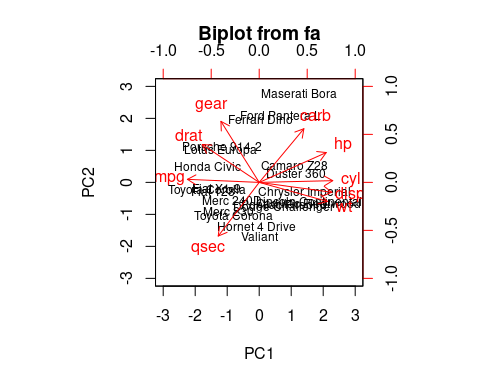
##   
## Loadings:  
## PC1 PC2   
## mpg -0.935   
## cyl 0.957   
## disp 0.945 -0.128  
## hp 0.873 0.389  
## drat -0.742 0.493  
## wt 0.888 -0.248  
## qsec -0.534 -0.698  
## gear -0.498 0.795  
## carb 0.582 0.699  
##   
## PC1 PC2  
## SS loadings 5.656 2.082  
## Proportion Var 0.628 0.231  
## Cumulative Var 0.628 0.860

Here we are looking at the loadings for all the variables. The loadings variable provides the coefficients of linear combination to compute pc1 and pc2 from the data. Positive loadings values indicates that the variable and the component are positively correlated. Negative loadings values indicate a negative correlation between the variable and the component. The larger value of cyl indicates the strong effect on PC1.

We can also see that PC1 explains 62% and PC2 explains 23% of variance in the data.

1. Get biplot of these two component loadings and interpret it carefully

biplot(final\_pca,labels=rownames(mtcars.subset))



We can see that the mpg and cyl are moving in opposite direction as they have high value but different sign while calculating loadings.

1. Get the head of the saved scores of mtcars.pca and interpret carefully

head(mtcars.pca$scores)

## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Mazda RX4 0.67485176 1.1922239 0.2075865 0.12803392 -0.76404073  
## Mazda RX4 Wag 0.64739389 0.9925769 -0.1125504 0.08704801 -0.66719589  
## Datsun 710 2.33653412 -0.3318150 0.2135122 0.11036335 0.07744294  
## Hornet 4 Drive 0.21874167 -2.0084411 0.3347400 0.31299155 0.24782081  
## Hornet Sportabout -1.61236724 -0.8419890 1.0495215 -0.14974247 0.22626754  
## Valiant -0.05039885 -2.4858332 -0.1135663 0.88549481 0.12776087  
## Comp.6 Comp.7 Comp.8 Comp.9  
## Mazda RX4 -0.12706898 0.43035095 0.003311314 -0.169724100  
## Mazda RX4 Wag -0.06725355 0.45566941 -0.057549593 -0.072737641  
## Datsun 710 -0.57600804 -0.39230244 0.205268522 0.116337227  
## Hornet 4 Drive 0.08516563 -0.03352155 0.024093561 -0.147579996  
## Hornet Sportabout 0.18572940 0.05886501 -0.154780223 -0.157120566  
## Valiant -0.23411734 -0.22810775 -0.100241845 -0.004301647

The original dataset is projected into the 8 principle componenets.

12 .Get the head of the scores of first two components of mtcars.pca and intepret it carefully

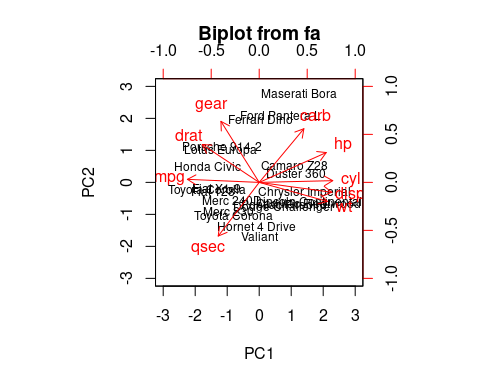
head(final\_pca$scores)

## PC1 PC2  
## Mazda RX4 -0.27929417 0.8132290  
## Mazda RX4 Wag -0.26793045 0.6770476  
## Datsun 710 -0.96699807 -0.2263347  
## Hornet 4 Drive -0.09052843 -1.3699797  
## Hornet Sportabout 0.66729435 -0.5743299  
## Valiant 0.02085807 -1.6956141

The original dataset is projected into the final two prinicpal componenet.

1. Get biplot of these two component scores and interpret it carefully

biplot(final\_pca,rownames(mtcars.subset))



Here we can observe that hp, cyl, disp and wt contribute to PC1 with higher values. And mpg which has negative loadings is in opposite direction to PC1 with higher values. Gear and carb has higher contribution to PC2 with positive values and qsec has negative value.

1. Get dissimilar distance of all the variables of mtcars data as mtcars.dist

mtcars.dist<-dist(mtcars.subset)

1. Fit classical multi-dimensional scaling model with the mtcars.dist in 2-dimensional state as cars.mds.2d

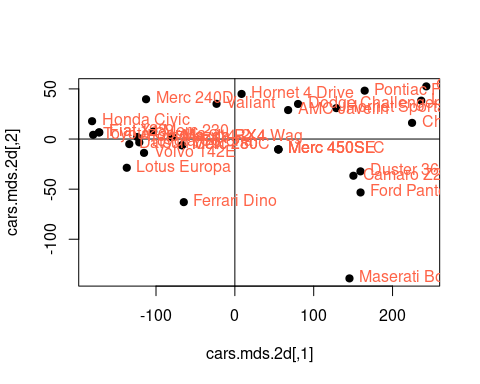
cars.mds.2d<-cmdscale(mtcars.dist)

summary(cars.mds.2d)

## V1 V2   
## Min. :-181.07 Min. :-139.047   
## 1st Qu.:-116.69 1st Qu.: -10.373   
## Median : -43.99 Median : 2.144   
## Mean : 0.00 Mean : 0.000   
## 3rd Qu.: 132.85 3rd Qu.: 29.375   
## Max. : 242.81 Max. : 52.503

1. Plot the cars.mds.2d and compare it with the biplot of mtcars.pca and interpret it carefully

plot(cars.mds.2d,pch=19)  
abline(h=0,v=0)  
text(cars.mds.2d, pos = 4, labels =rownames(mtcars.subset), col ='tomato')



From the graph we can see that Hornet 4 Drive, Chry etc are in the positive qudrantt whch means they have positive contribution to the first component.

1. Fit classical multi-dimensional scaling model with the mtcars.dist in 3-dimensional state as cars.mds.3d

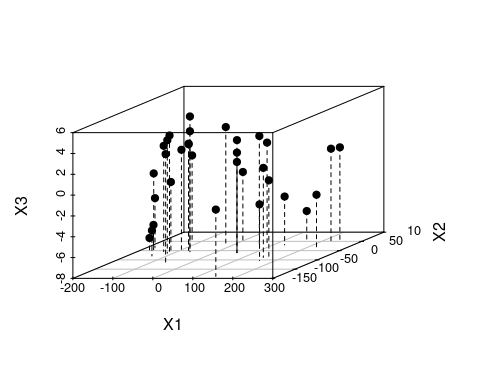
cars.mds.3d<-cmdscale(mtcars.dist,k=3)

summary(cars.mds.3d)

## V1 V2 V3   
## Min. :-181.07 Min. :-139.047 Min. :-6.8611   
## 1st Qu.:-116.69 1st Qu.: -10.373 1st Qu.:-1.8374   
## Median : -43.99 Median : 2.144 Median : 0.8492   
## Mean : 0.00 Mean : 0.000 Mean : 0.0000   
## 3rd Qu.: 132.85 3rd Qu.: 29.375 3rd Qu.: 2.2806   
## Max. : 242.81 Max. : 52.503 Max. : 5.0029

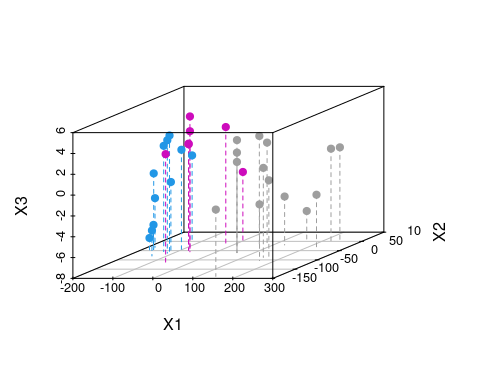
1. Create a 3-d scatterplot of cars.mds.3d with type = “h”, pch=20 and lty.hplot=2 and interpret it carefully

library(scatterplot3d)  
cars.mds.3d <- data.frame(cmdscale(mtcars.dist, k = 3))  
scatterplot3d(cars.mds.3d, type = "h", pch = 19, lty.hplot = 2)



1. Create a 3-d scatterplot of cars.mds.3d with type = “h”, pch=20, lty.hplot=2 and color=mtcars$cyl and interpret it carefully

library(scatterplot3d)  
cars.mds.3d <- data.frame(cmdscale(mtcars.dist, k = 3))  
scatterplot3d(cars.mds.3d, type = "h", pch = 19, lty.hplot = 2, color = mtcars$cyl)



The plot shows the principle component in 3 dimensions.

1. Write a summary comparing PCA and MDS fits done above for mtcars data

PCA takes original dataset as an input whereas MDS takes pairwise distance between the data points as input. For the mtcars, it shows that two latent variables can be generated from the given dataset that can explain about 86% of variance in the data.