Department of Statistics

Jahangirnagar University

Professional Masters in Applied Statistics and Data Science (ASDS)

Course Title: Multivariate Analysis

Course No.: PM-ASDS06

Assignment 1, Summer 2021

Submit to –

Dr. Rumana Rois

Associate Professor, Department of Statistics

Professional Masters in Applied Statistics and Data Science (ASDS)

Jahangirnagar University

Submitted by –

Md. Shamimul Islam

ID:20204012

Section-B, 4 Batch

Professional Masters in Applied Statistics and Data Science (ASDS)

Jahangirnagar University

Code for Append the data

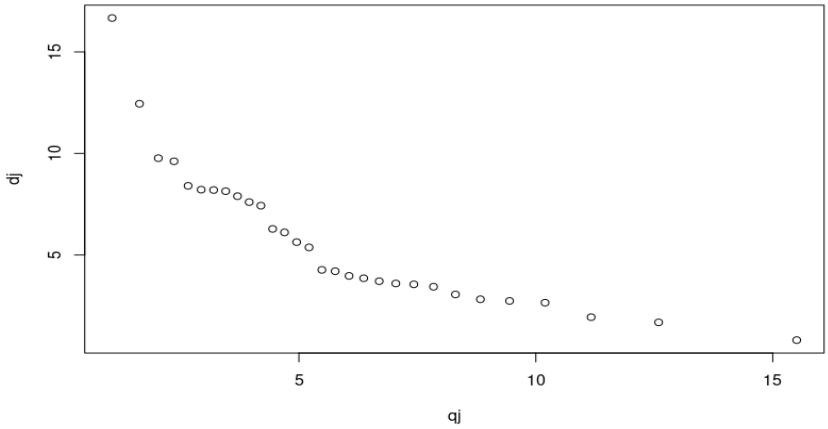
```
 X=read.table("/media/shamim/New Volume/Data Science 2nd Semester/Multivariate analysis/Assignment/T1-8.DAT",header = FALSE)
new_row_26 <- c(0.8912, 0.812, 1.7912, 1.712, 0.712, 0.612)
new_row_27 <- c(0.9112, 0.912, 1.8112, 1.812, 0.612, 0.812)
new_row_28 <- c(0.8412, 0.912, 1.8912, 1.712, 0.712, 0.712)
new_row_29 <- c(0.8812, 0.812, 1.9112, 1.812, 0.812, 0.612)
new_row_30 <- c(0.8312, 0.912, 1.9412, 1.912, 0.612, 0.812)
X <- rbind(X, new_row_26)
X <- rbind(X, new_row_27)
X <- rbind(X, new_row_28)
X <- rbind(X, new_row_29)
X <- rbind(X, new_row_30)


X=as.matrix(X)
```

1. Examine the multivariate normality of the observations on six different variables of the mineral content of three bones on the dominant and nondominant sides of older women. Also, detect the outliers

   Code:

   ```
   X=as.matrix(X)
   p = 6   #numbers of variables
   X_bar=colMeans(X)
   S = cov(X)
   n=nrow(X)  #number of observations
   Dsq=matrix(0,nrow=n,ncol=1)
   for (i in 1:n) {
     Dsq[i] = t(X[i,]-X_bar)%*%solve(S)%*%(X[i,]-X_bar)
   }
   dj=sort(Dsq, decreasing = F)
   j=matrix(seq(1,n), nrow=n, ncol=1)
   qj = qchisq((n-j+0.5)/n, p)
   plot(qj,dj)
   ```

   Result Analysis:



The observations on six different variables of the mineral content of three bones on the dominant and nondominant sides of older women are not hold Normality.

There are two outliers.

2. Evaluate T^2 of the six variables (x1, x2, x3,.....,x6) for testing H0 : μ = 0 . 80 0 . 80 1 . 70 1 . 70 0 . 70 0 . 70 at 5% level of significance. Hence, find out the sampling distribution of T^2 .

Code:

```
x = X
mu0=matrix(c(0.80, 0.80, 1.70, 1.70, 0.70, 0.70),6,1) ### From the q
xbar=colMeans(x)
S=cov(x)
S_inv=solve(S)
n=nrow(x)  #number of observations
T2=n*t(xbar-mu0)%*%S_inv%*%(xbar-mu0)
p=ncol(x)  #numbers of variables


F_value=qf(.95, ncol(x), nrow(x)-ncol(x))### alpha=.05, 1-alpha=1-.05=.95
critical_value=((n-1)*p)/(n-p)*F_value
critical_value
T2 ### If T2 > critical_value then H0 will be rejected at alpha
```

Result interpretation:

T^2 = 17.65282
Critical Value = 18.18437

H0 isn't rejected at 5% level of significance.

3. Construct the sample covariance matrix S for the above data matrix. Hence, determine the sample principal components and their variances for the covariance matrix S . How many principal components will be retained in this analysis?

Code:

```
########## Principal Component Analysis ##############
turtles=log(X[1:30,1:6]) ### turtles is the data set
xbar=colMeans(turtles)
S=cov(turtles)
fit <- prcomp(S) #princomp(S)
fit ### To get the PC after rotation
summary(fit) ## To get the proportions
screeplot(fit, npcs = 6, type = "lines")
eigen(S) ### To get the PC without rotation
var_explained = fit$sdev^2 / sum(fit$sdev^2)
library(ggplot2)

qplot(c(1:6), var_explained) +
  geom_line() +
  xlab("Principal Component") +
  ylab("Variance Explained") +
  ggtitle("Scree Plot") +
  ylim(0, 1)
```

Result interpretation:

Sample Covariance Matrix

	V1	V2	V3	V4	V5	V6
V1	0.01889908	0.01437302	0.015846722	0.01271358	0.01370695	0.011680826
V2	0.01437302	0.01737135	0.012539357	0.01371694	0.01217332	0.015641990
V3	0.01584672	0.01253936	0.024631079	0.01968804	0.01230013	0.009500464
V4	0.01271358	0.01371694	0.019688036	0.02017717	0.01205229	0.012628956
V5	0.01370695	0.01217332	0.012300127	0.01205229	0.02198224	0.011720309
V6	0.01168083	0.01564199	0.009500464	0.01262896	0.01172031	0.022259261

eigen() decomposition

$values

[1] 0.0878637462 0.0163014702 0.0110387191 0.0064795217 0.0027145134
0.0009222111

$vectors

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	-0.4057505	0.03630578	-0.2156633	0.74564760	0.21490229	0.43053974
[2,]	-0.3969767	-0.27095915	0.1657710	0.27901828	-0.71184761	-0.39614694
[3,]	-0.4461545	0.62733388	0.1241836	-0.02346706	0.32667853	-0.53357962
[4,]	-0.4268221	0.29948784	0.2937628	-0.42232207	-0.31031408	0.60595591
[5,]	-0.3872591	-0.18051559	-0.8121103	-0.39055065	-0.01789568	-0.07121332
[6,]	-0.3828518	-0.63987217	0.4059163	-0.18633841	0.49371655	-0.02724029

Var_explained = 6.223659e-01 2.682958e-01 9.124324e-02 1.607073e-02
2.024330e-03 6.926089e-34

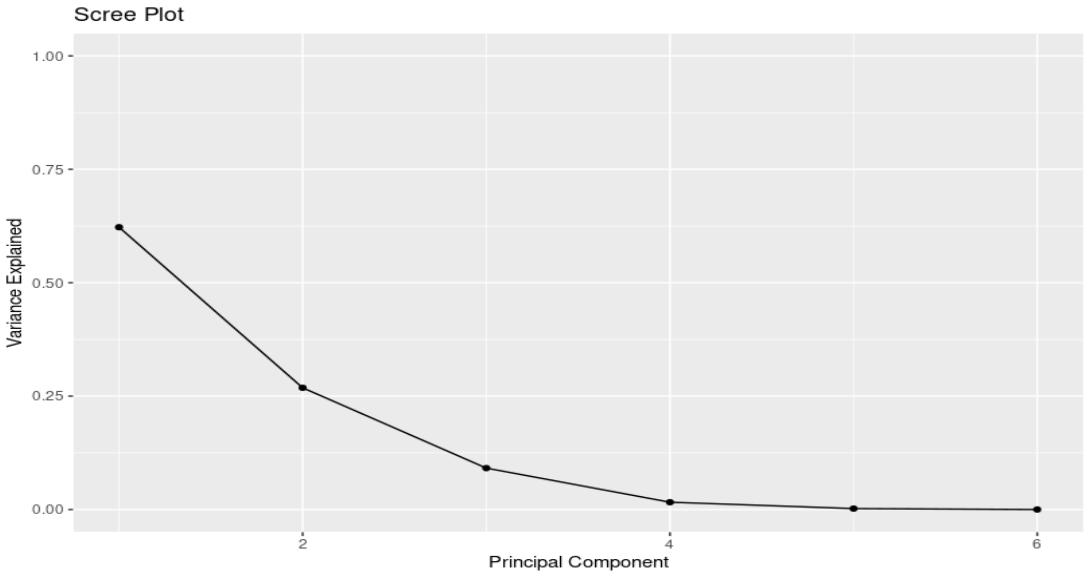


Figure: Scree Plot

Three principal components are retained in this analysis.

4. Conduct the factor analysis with these 6 variables and m = 2 common factors using maximum likelihood procedure and find the followings:
   (i)
   Find the estimated factor loadings and communalities.
   (ii)
   What proportion of the total population variance is explained by the first common factors? And by the 2 nd common factor.
   (iii) Check whether the 2 factors are adequate for our model?

Code:

```
fac <- factanal(X, factors=2, method='mle', scale=T, center=T)

factanal(X, factors=2, method='PCA', scale=T, center=T)

factanal(x = X, factors = 2, method = "PCA", scale = T, center = T)
```

Result interpretation:

i)

	Factor1	Factor2
V1	0.684	0.511
V2	0.904	0.377
V3	0.273	0.960
V4	0.483	0.790
V5	0.545	0.386
V6	0.802	0.209

Communality for the Variable Dominant radius = (0.684)^2 + (0.511)^2 = 0.728977
Communality for the Variable Radius = (0.904)^2 + (0.377)^2 = .959345
Communality for the Variable Dominant humerus = (0.273)^2 + (0.960)^2 = .996129
Communality for the Variable Humerus = (0.483)^2 + (0.790)^2 = .857389
Communality for the Variable Dominant ulna = (0.545)^2 + (0.386)^2 = .446021
Communality for the Variable Ulna = (0.802)^2 + (0.209)^2 = .686885

ii) What proportion of the total population variance is explained by the first common factors? And by the 2 nd common factor.
The first common factor explains the 0.422 proportion of the total population variance.
The Second common factor explains the 0.357 proportion of the total population variance.

iii) Check whether the 2 factors are adequate for our model?

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 9.89 on 4 degrees of freedom.
The p-value is 0.0422

5. Calculate the Euclidean distances between six different variables of the mineral content of three bones on the dominant and nondominant sides of older women. Cluster the six variables using the single linage and complete linkage hierarchical methods. Draw the dendrograms and compare the results.

Code:

```
############### Cluster Analysis ###############
data1=X
#To remove any missing value that might be present in the data, type this:
data1 <- na.omit(data1)
#As we don't want the clustering algorithm to depend to an arbitrary variable unit,
#we start by scaling/standardizing the data using the R function
sdata1 <- scale(data1)
# Dissimilarity matrix
d <- dist(t(sdata1), method = "euclidean") ### t for variable-wise


# Hierarchical clustering using Complete Linkage
hc1 <- hclust(d, method = "single" ) #"average", "single", "complete", "ward"


# Plot the obtained dendrogram
plot(hc1, cex = 0.6, hang = -1)

# Hierarchical clustering using Complete Linkage
hc1 <- hclust(d, method = "complete" ) #"average", "single", "complete", "ward"


# Plot the obtained dendrogram
plot(hc1, cex = 0.6, hang = -1)
```

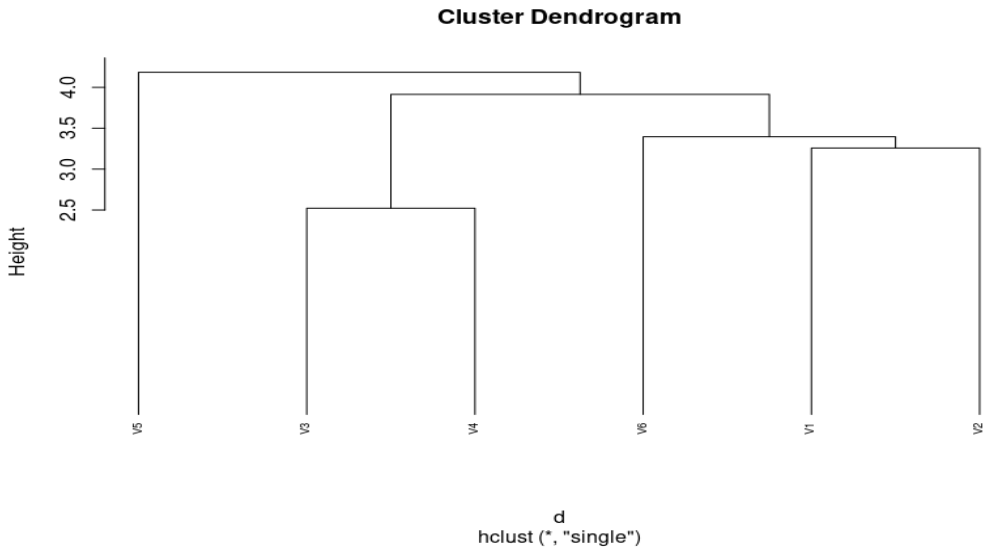
Result interpretation:

Euclidean distances between six different variables of the mineral content of three bones on the dominant and nondominant sides of older women

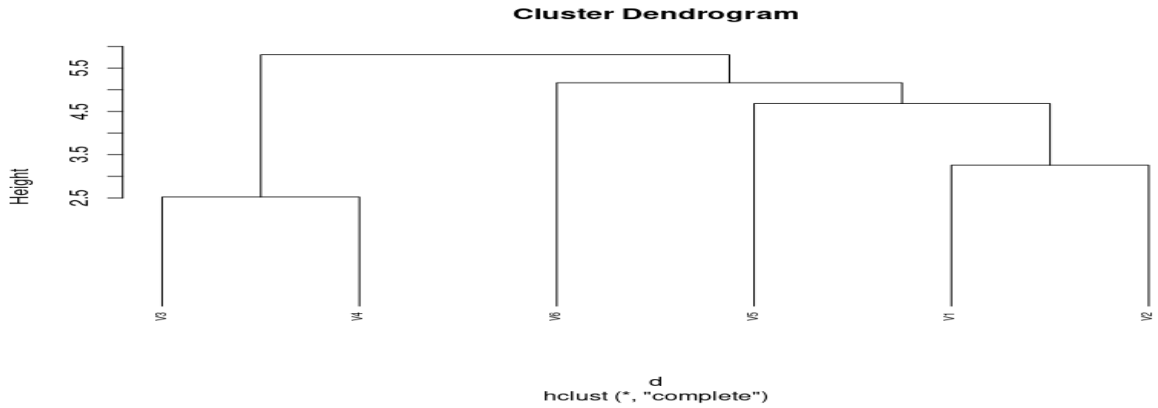
	V1	V2	V3	V4	V5
V2	3.258055				
V3	4.321270	4.770605			

V4 4.461456 3.914452 2.522297
V5 4.184305 4.681984 5.283251 5.007780
V6 4.706802 3.395765 5.809538 4.810075 5.160707

Cluster the six variables using the single linkage



Cluster the six variables using Complete Linkage



In Single linkage, there is an outlier in this analysis. In height 2.5, V3 and V4 build a cluster. Then 3.4 height, V1 and V2 construct a cluster. In distance 3.5, V1, V2 and V5 build a cluster. In height 3.9, V1, V2, V3, V4, V5 construct a cluster. Finally, V3, V4, V1, V2, V5 and V6 construct a cluster.

In complete linkage, there is no outlier in this cluster. In height 2.5, V3 and V4 build a cluster, then 3.3 height, V1 and V2 construct a cluster. In height 4.5, V5, V1 and V2 construct a cluster. In Height 5, V6, V5, V1, V2 construct a cluster. Finally in height 5.5 V3, V4, V6, V5, V1, V2 construct a cluster.