**Bangabandhu Sheikh Mujibur Rahman Agricultural University**

**EDGE\_Batch-11**

**Project Report Marks: 25**

**Name:Fariya Akter**

**Reg. No:**19-05-5092 **Dept:**Genetics and Fish Breeding

**Note: Submit the completed file as pdf to** [**nazmol.stat.bioin@bsmrau.edu.bd**](mailto:nazmol.stat.bioin@bsmrau.edu.bd) **and** [**rabiulauwul@bsmrau.edu.bd**](mailto:rabiulauwul@bsmrau.edu.bd) **with subject: *EDGE\_11\_Project\_Your registration number\_ Department by 13th of January, 2025.***

**Problem# 1:** Choose a multivariate dataset (with at least 10 variables) in your subject area and solve the following issue. (***Attach your dataset in csv file to the email***)

Ans to the qu no 1(a)

1. Pre-process your dataset with imputing outliers and missing values.

data<-read.csv("Fish Project Data.csv")

My\_PCA\_Data<-na.omit(data[,-c(2,13,14)])

**#Missing value**

colSums(is.na(My\_PCA\_Data))

colSums(is.na(My\_PCA\_Data))

Fish\_ID Weight Length Age Total\_Fin\_Length

0 0 0 0 0

Girth Height Water\_Temperature Salinity Depth

0 0 0 0 0

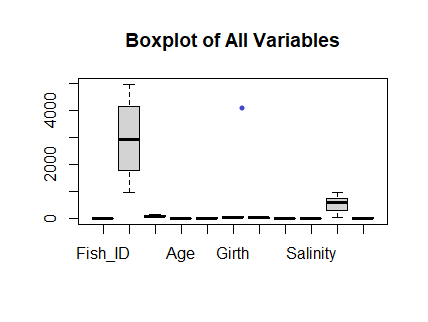
Catch\_Frequency

0

**There is no missing value in my dataset.**

**#Check outlier**

boxplot(My\_PCA\_Data, main = "Boxplot of All Variables")

 Figure:boxplot1

Here,Potential Outlier present in the Fish\_ID

**#Check outlier**

boxplot(My\_PCA\_Data$Fish\_ID,

main = "Boxplot of Fish\_ID",

ylab = "Fish\_ID")

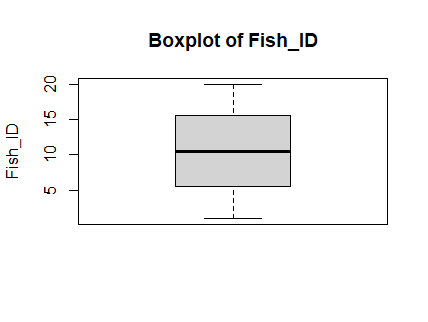


Figure:boxplot2

**# Calculate lower and upper bounds using MAD**

lower\_bound <- median(My\_PCA\_Data$Fish\_ID, na.rm = TRUE) -

3 \* mad(My\_PCA\_Data$Fish\_ID, na.rm = TRUE)

> lower\_bound

[1] -11.739

upper\_bound <- median(My\_PCA\_Data$Fish\_ID, na.rm = TRUE) +

3 \* mad(My\_PCA\_Data$Fish\_ID, na.rm = TRUE)

> upper\_bound

[1] 32.739

**# Identify indices of outliers**

outliers <- which(My\_PCA\_Data$Fish\_ID < lower\_bound |

My\_PCA\_Data$Fish\_ID > upper\_bound)

> outliers

integer(0)

**# Replace outliers with the calculated bounds**

My\_PCA\_Data$Fish\_ID [My\_PCA\_Data$Fish\_ID < lower\_bound] <- lower\_bound

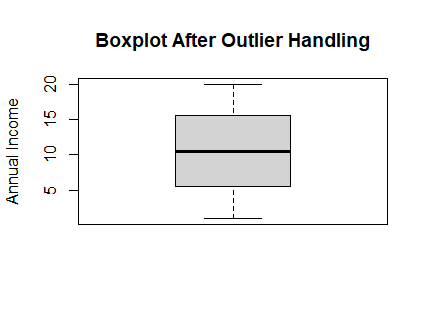
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

My\_PCA\_Data$Fish\_ID [My\_PCA\_Data$Fish\_ID > upper\_bound] <- upper\_bound

[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

boxplot(My\_PCA\_Data$Fish\_ID ,

main = "Boxplot After Outlier Handling",

 ylab = "Annual Income")

1. **Interpret how many principle components should be retained for your data with justification.**

# Perform PCA

correlation<- cor(My\_PCA\_Data)

mean(correlation)

eigen(correlation)

PCA\_result <- prcomp(My\_PCA\_Data, scale. = TRUE)

summary(PCA\_result)

> summary(PCA\_result)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Importance  of components | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 | PC11 |
| Standard deviation | 1.502 | 1.502 | 1.3231 | 1.1718 | 1.04814 | 0.93129 | 0.8035 | 0.66857 | 0.52 | 0.51 | 0.1961 |
| Proportion of Variance | 0.205 | 0.1799 | 0.1592 | 0.1248 | 0.09987 | 0.07885 | 0.0587 | 0.04063 | 0.02547 | 0.02418 | 0.0035 |
| Cumulative Proportion | 0.0035 | 0.3848 | 0.5440 | 0.6688 | 0.76868 | 0.84753 | 0.9062 | 0.94686 | 0.97233 | 0.99650 | 1.0000 |

install.packages("devtools")

library(devtools)

install\_github("vqv/ggbiplot")

ggscreeplot(PCA\_result)+aes(color="red"

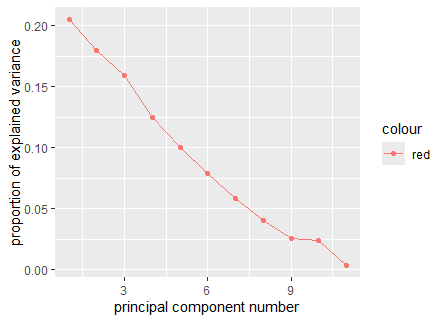


Figure:Screeplot

**Interpretation:**

To determine the number of principal components to retain, we evaluate the cumulative proportion of variance explained and examine the scree plot:

**Cumulative Proportion:** **Cumulative Proportion of Variance**:

* A common criterion is to select components until the cumulative variance explained exceeds 70% to 85%.
* From the provided table, the cumulative variance reaches **84.75%** by the first **six components** (PC1 to PC6).This level of variance is typically sufficient for retaining components, as it represents a significant portion of the dataset's variability.

**ScreePlot:**

The scree plot reveals an "elbow" after the six component, indicating a slower rate of decline in explained variance beyond this point. This supports the decision to retain six components.

1. **Construct a bi-plot with ggplot2 package for the selected principle components and describe the plots.**

#To draw bi-plot

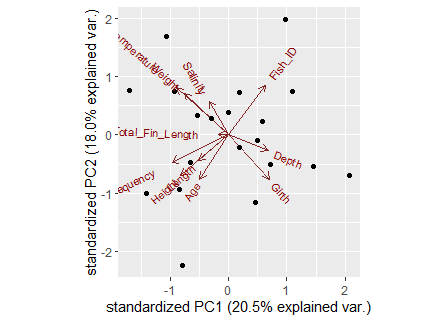
install.packages("devtools")

library(devtools)

install\_github("vqv/ggbiplot")

library(ggbiplot)

ggbiplot(PCA\_result)

****

### **Figure:Biplot**

### Key Features:

1. **Axes:**
   * The x-axis represents the first principal component (PC1), which explains 20.5% of the variation in the data.
   * The y-axis represents the second principal component (PC2), which explains 18.0% of the variation.
   * Combined, these two components capture 38.5% of the total variance.
2. **Variables (Red Arrows):**
   * Arrows represent the original variables, showing their contribution and direction in the PCA space.
   * The length of the arrows indicates the strength of the variable’s influence on the respective components.
   * For example:
     + **Total\_Fin\_Length** and **Temperature (possibly)** are strongly aligned with PC1.
     + **Depth**, **Girth**, and **Fish\_ID** are more aligned with PC2.
3. **Points:**
   * Each point represents an observation (individual fish or sample).
   * The positioning of points reflects their similarity: observations closer together have similar profiles based on the variables.
4. **Correlations:**
   * Variables pointing in the same direction are positively correlated (e.g., **Depth** and **Girth**).
   * Variables pointing in opposite directions are negatively correlated.
   * Variables perpendicular to each other are uncorrelated.

### Observations:

* **PC1 Key Drivers:**
  + Variables like **Total\_Fin\_Length**, **Temperature**, and **Weight** strongly influence PC1.
  + Observations to the right of PC1 have higher values for these variables, while those to the left have lower values.
* **PC2 Key Drivers:**
  + Variables like **Depth** and **Girth** strongly influence PC2.
  + Observations higher on PC2 have greater depth and girth compared to those lower on PC2.
* **Variable Clustering:**
  + Some variables (e.g., **Salinity**, **Age**) cluster together, suggesting they contribute similarly to the variance.

### Insights:

1. **Multivariate Structure:**
   * The data exhibits complexity, with no single variable dominating variance, as PC1 and PC2 explain moderate (not overwhelming) proportions of the variance.
2. **Biological Interpretation:**
   * The clustering of variables like **Depth**, **Girth**, and **Fish\_ID** along PC2 suggests morphological traits and identification metrics are linked.
   * Traits like **Total\_Fin\_Length** and **Temperature** influence a separate axis (PC1), perhaps related to environmental or growth factors.
3. **Potential Clusters of Observations:**
   * Observations group loosely in some areas, which may represent natural clusters of fish based on size, age, or habitat.
4. **Uncorrelated Variables:**
   * Variables perpendicular to each other (e.g., **Salinity** vs. **Depth**) suggest independent patterns, worth further investigation.
5. **Test whether your data is suitable for factor analysis or not.**

# Install required packages

install.packages("psych")

install.packages("GPArotation")

2.

# Load the libraries

library(psych)

library(GPArotation)

3.

data <- read.csv("Fish Project Data Final.csv")

library(psych)

# KMO Test

KMO(data)

> # KMO Test

> KMO(My\_PCA\_Data)

Kaiser-Meyer-Olkin factor adequacy

Call: KMO(r = My\_PCA\_Data)

Overall MSA = 0.21

MSA for each item =

Fish\_ID Weight Length Age Total\_Fin\_Length

0.15 0.16 0.48 0.16 0.21

Girth Height Water\_Temperature Salinity Depth

0.20 0.32 0.42 0.11 0.24

Catch\_Frequency

0.20

For Kaiser-Meyer-Olkin (KMO) test

• KMO > 0.9: Marvelous – Excellent suitability for factor analysis.

• KMO between 0.8 and 0.9: Great – Very good suitability.

• KMO between 0.7 and 0.8: Good – Adequate, acceptable for factor analysis.

• KMO between 0.6 and 0.7: Mediocre – Marginally acceptable, might need further checks.

• KMO < 0.6: Not suitable – Factor analysis may not be appropriate for this data.

Here, Overall KMO = 0.21.So, factor analysis may not be appropriate for this data.

# Bartlett’s Test

bartlett.test(My\_PCA\_Data)

> bartlett.test(My\_PCA\_Data)

Bartlett test of homogeneity of variances

data: My\_PCA\_Data

Bartlett's K-squared = 1171.8, df = 10, p-value < 2.2e-16

If the p-value is less than 0.05, we can conclude that the data is suitable for factor analysis. Here, the p-value is very small (< 0.05), which indicates that the correlation matrix is significantly different from an identity matrix. This suggests that the variables in the data are correlated enough to justify the use of factor analysis. In other words, Bartlett’s test indicates that factor analysis is appropriate for the data.

**e)Construct a suitable plot to visualize the factors with their loadings with factor analysis.**

#Perform factor analysis

fact\_result<-factanal(factors=2, covmat = cov(My\_PCA\_Data))

Rotation<-factanal(factors=2, covmat = cov(My\_PCA\_Data), rotation = "varimax")

print(fact\_result)

> print(fact\_result)

Call:

factanal(factors = 2, covmat = cov(My\_PCA\_Data))

Uniquenesses:

Fish\_ID Weight Length Age Total\_Fin\_Length

0.796 0.620 0.975 0.005 0.958

Girth Height Water\_Temperature Salinity Depth

0.588 0.866 0.295 0.881 0.865

Catch\_Frequency

0.855

Loadings:

Factor1 Factor2

Fish\_ID -0.445

Weight 0.610

Length -0.159

Age 0.918 -0.390

Total\_Fin\_Length 0.197

Girth -0.416 -0.488

Height 0.365

Water\_Temperature 0.306 0.782

Salinity 0.305 0.163

Depth -0.364

Catch\_Frequency 0.381

Factor1 Factor2

SS loadings 1.718 1.579

Proportion Var 0.156 0.144

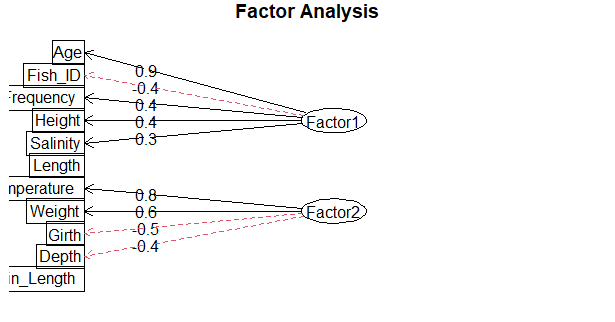
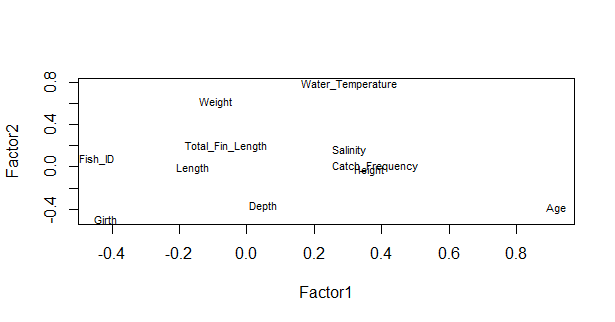
Cumulative Var 0.156 0.300

The degrees of freedom for the model is 34 and the fit was 3.2365

fa.diagram(plotloads)

#Plot

plot(plotloads,type="n")

text(plotloads,labels=names(My\_PCA\_Data), cex= .7)

### Observations:

1. **Factor 1:**
   * Variables **Age**, **Fish\_ID**, and **Frequency** have strong positive loadings (0.9, 0.9, and 0.4, respectively).
   * This suggests that **Factor 1** likely represents a dimension related to **individual identification and age characteristics** of fish.
2. **Factor 2:**
   * Variables **Weight** (0.8), **Girth** (0.6), and **Depth** (0.5) have high positive loadings.
   * This suggests **Factor 2** likely represents a **morphological or size-related factor** for the fish.
3. **Other Variables:**
   * **Temperature** and **Length** are moderately associated with Factor 2, while **Height** and **Salinity** have weaker relationships with both factors (loadings close to 0.3 or lower)

### Insights:

1. **Variable Groupings:**
   * Variables related to size and morphology (e.g., **Weight**, **Girth**, **Depth**) cluster under **Factor 2**.
   * Variables associated with identification or individual attributes (e.g., **Age**, **Fish\_ID**) cluster under **Factor 1**.
2. **Interpretation of Factors:**
   * **Factor 1:** Likely captures variability in demographic or inherent characteristics (e.g., age or ID).
   * **Factor 2:** Likely represents variability in physical traits like size or weight.

**Problem # 2:**  A two-factor factorial design was conducted considering tree blocks, three levels/treatments of variety, and five levels/treatments of nitrogen. Afterward, the yield of certain plant characteristics was observed. The data regarding this experiment were given in the file “Data\_Factorial\_Design”. Answer the following question using this data.

1. Construct an ANOVA table using the mentioned dataset based on R programming.

b) Write down the null hypothesis of all possible effects and interpret the results based on the ANOVA table

1. Perform a post-hoc test for the levels/treatments of nitrogen and draw a bar diagram with lettering.

**Ans to the qu no.2(a)**

# Loading the data

Data.factorial <- read.csv("Data\_Factorial\_Design.csv")

# Defining factors

block <- c("Block1", "Block2", "Block3")

variety <- c("Variety1", "Variety2", "Variety3")

nitrogen <- c("Nitrogen1", "Nitrogen2", "Nitrogen3", "Nitrogen4", "Nitrogen5")

# Determining the total number of blocks, varieties, and nitrogen levels

b <- length(block)

v <- length(variety)

n <- length(nitrogen)

# Generating factorial combinations

Block <- gl(b, v \* n, b \* v \* n, factor(block))

Varfact <- gl(v, n, b \* v \* n, factor(variety))

NitroFact <- gl(n, 1, b \* v \* n, factor(nitrogen)

# **Performing ANOVA for Randomized Complete Block Design (RCBD)**

ANOVA.twoFact.Factorial.RCBD <- aov(data = Data.factorial, YIELD ~ Varfact + Block + NitroFact + Varfact \* NitroFact)

summary(ANOVA.twoFact.Factorial.RCBD)

> summary(ANOVA.twoFact.Factorial.RCBD)

Df Sum Sq Mean Sq F value Pr(>F)

Varfact 2 1.93 0.963 22.09 1.75e-06 \*\*\*

Block 2 1.25 0.627 14.39 5.02e-05 \*\*\*

NitroFact 4 66.03 16.507 378.73 < 2e-16 \*\*\*

Varfact:NitroFact 8 6.10 0.763 17.50 5.23e-09 \*\*\*

Residuals 28 1.22 0.044

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Ans to the qu no.2(b)**

The null hypotheses are:

* **Main Effect of Block**: H0: μBlock1=μBlock2=μBlock3

Interpretation: Since p<0.05(table ),we can reject the null hypothesis by concluding that there are significance differences in all block levels.

* **Main Effect of Variety:** H0: μvariety1= μvariety2= μvariety3

Interpretation: Since p<0.05(table ),we can reject the null hypothesis by concluding that there are significance differences in all variety levels.

* **Main Effect of Nitrogen:** H0: μNitrogen1= μNitrogen2= μNitrogen3

Interpretation: Since p<0.05(table ),we can reject the null hypothesis by concluding that there are significance differences in all Nitrogen levels.

* **Interactiobn Effect (Variety\*Nitrogen):**

H0:(μvariety\*Nitrogen)ij= μVarietyi+μNitrogenj

Interpretation: Since p<0.05(table ),we can reject the null hypothesis by concluding that there are significance differences between Variety and Nitrogen levels.

**Ans to the Qu no (C)**

library(agricolae)

# Post-hoc test for Nitrogen levels

PostHoc.Test.nitrogen<-with(Data.factorial,HSD.test(YIELD,NITROGEN,DFerror = 28,MSerror = 0.044)

|  |  |  |
| --- | --- | --- |
| NITROGEN | YIELD | groups |
| 4 | 6.302222 | a |
| 5 | 5.858889 | b |
| 3 | 5.628889 | b |
| 2 | 4.804444 | c |
| 1 | 2.875556 | d |

From PostHoc test we can conclude that,

• Group a: Nitrogen level 4, highest yield, most distinct.

• Group b: Nitrogen levels 3 and 5, moderate yields.

• Group c: Nitrogen level 2, moderate-low yields

• Group d: Nitrogen level 1, lowest yield

#Barplot

Mutplcom.NitroFact<-with(Data.factorial,HSD.test (YIELD,NITROGEN,DFerror=28,MSerror=0.044))

Nitro.Mean <- Mutplcom.NitroFact$groups

Nitro.SE.Mat <- Mutplcom.NitroFact$means

Nitro.SE.Mat <- Mutplcom.NitroFact$means[, "se"]

Mean.Mat <- Mutplcom.NitroFact$means Mean.Mat <- Mean.Mat[order(-Mean.Mat$YIELD),]

Nitro.Nitro.Mean <- Nitro.Mean$YIELD

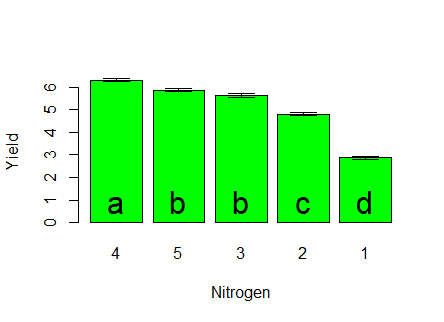
Nitro.SE <- Mean.Mat[, "se"]

Nitro.SE.Mat <- Mutplcom.NitroFact$means[order(Mutplcom.NitroFact$means[,”se”])]

library(gplots)

Barplot.SE <- barplot2(Nitro.Nitro.Mean, names.arg = rownames(Nitro.Mean), xlab = "Nitrogen", ylab = "Yield", horiz = F, plot.ci = T, ci.l = Nitro.Nitro.Mean - Nitro.SE, ci.u = Nitro.Nitro.Mean + Nitro.SE, col = "green")

text(Barplot.SE, 0,Nitro.Mean$groups , cex = 2, pos = 3, col = "black")



Page | 14