Bangabandhu Sheikh Mujibur Rahman Agricultural University
EDGE_Batch-06
Project Report Marks: 25
Name: Md. Mohaimenul Islam
Reg. No: 2019-05-5035, Dept: BGE

Note: Submit the completed file as *pdf* to nazmol.stat.bioin@bsmrau.edu.bd with subject *EDGE_06_Project_Your registration number_ Department by 26th of December, 2024.*

Problem# 1:

A split-plot design was conducted considering tree blocks, three levels/treatments of variety in the main plot, and five levels/treatments of nitrogen in the split-plot. Afterward, the yield of certain plant characteristics was observed. The data regarding this experiment were given in the file "Split_Plot_Design". Answer the following question using this data.

a) 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.
c) Perform a post-hoc test for the interaction effect (variety × nitrogen) and draw a bar diagram with lettering.

Problem# 2:

a) What is principal component analysis?
b) What are the main purposes of principle component analysis in your study area?
c) Compute the eigenvalue and eigenvector using the iris data based on R programming.
d) Construct a scree plot and interpret how many principal components should be retained to interpret the iris dataset.
e) Construct a bi-plot for the iris data based on R programming and interpret the results.

ANSWER:

Solution 01:

a) Construction of an ANOVA table using the mentioned dataset based on R programming is given below:

```
# Load  data

data <- read.csv("Split_Plot_Design.csv")
```

```
# Fit the model
REPLICATE <- as.factor(data$REPLICAT)
REPLICATE
VARIETY <- as.factor(data$VARIETY)
VARIETY
NITROGEN <- as.factor(data$NITROGEN)
NITROGEN
# Assuming REPLICAT is a random effect and VARIETY and NITROGEN are fixed effects
model <- aov(YIELD ~ VARIETY * NITROGEN + Error(REPLICAT/VARIETY), data = data)
model
# Display the ANOVA table
summary(model)
```

Result:

```
summary(model)
Error: REPLICAT
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  1   1.24    1.24

Error: REPLICAT:VARIETY
          Df Sum Sq  Mean Sq
VARIETY  1 0.4944   0.4944
```

Anova Table:

Factors	Df	Sum Sq	Mean Sq	F value	Pr(>F)
VARIETY	1	0.47	0.47	0.764	0.387
NITROGEN	1	50.15	50.15	80.918	4.69e-11 ***
VARIETY:NITROGEN	1	0.01	0.01	0.010	0.922
Residuals	39	24.17	0.62		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

b) The null hypothesis of all possible effects and interpretion of the results based on the ANOVA table is given below:

Main effect of VARIETY:

Null hypothesis: There is no significant difference in the means of YIELD between the different levels of VARIETY.

Interpretation: The p-value for VARIETY is 0.387, which is greater than the commonly used significance level of 0.05. Therefore, we fail to reject the null hypothesis and conclude that there is no significant effect of VARIETY on YIELD.

Main effect of NITROGEN:

Null hypothesis: There is no significant difference in the means of YIELD across the different levels of NITROGEN.

Interpretation: The p-value for NITROGEN is very small (4.69e-11), which is less than 0.05. Therefore, we reject the null hypothesis and conclude that the levels of NITROGEN significantly affect YIELD.

Interaction effect of VARIETY and NITROGEN:

Null hypothesis: There is no significant interaction effect between VARIETY and NITROGEN on YIELD.

Interpretation: The p-value for the interaction (VARIETY:NITROGEN) is 0.922, which is much greater than 0.05. Therefore, we fail to reject the null hypothesis and conclude that there is no significant interaction between VARIETY and NITROGEN on YIELD.

Summary :

VARIETY does not significantly affect YIELD (p = 0.387).

NITROGEN has a significant effect on YIELD (p = 4.69e-11).

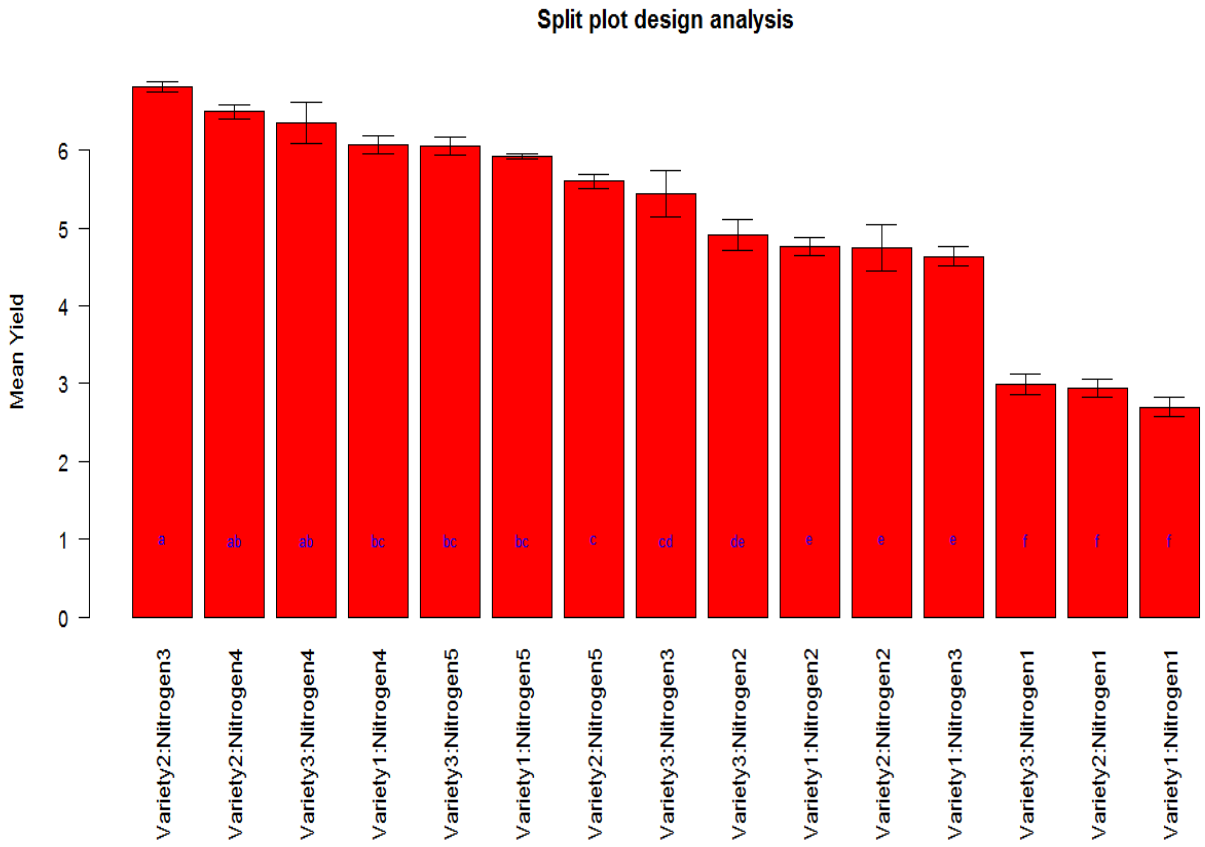
There is no significant interaction between VARIETY and NITROGEN (p = 0.922).

c) Perform a post-hoc test for the interaction effect (variety × nitrogen) and draw a bar diagram with lettering.

###CODE

```
# Load necessary packages

install.packages("emmeans")

library(emmeans)


# Load your data (ensure the correct file path)

data <- read.csv("Split_Plot_Design.csv")


# Fit the model (assuming the data has a structure for split-plot design)

model <- aov(YIELD ~ VARIETY * NITROGEN + Error(REPLICAT/VARIETY), data = data)


# Perform the post-hoc test for the interaction effect (VARIETY × NITROGEN)

post.hoc.test.yield <- with(data, HSD.test(YIELD, VARIETY:NITROGEN,

                                    DFerror = 39, MSerror = 0.62))

post.hoc.test.yield

# Barplot with lettering

mean.matrix <- post.hoc.test.yield$means
mean.matrix <- mean.matrix[order(mean.matrix$YIELD, decreasing = TRUE),]
mean.matrix

Mu_Tret <- mean.matrix$YIELD
Mu_Tret
SE_Tret <- mean.matrix$std/sqrt(mean.matrix$r)
SE_Tret
```

```
library(gplots)

par(mar = c(8, 6, 4, 4) + 1)
bar.plot <- barplot2(Mu_Tret, names.arg = rownames(mean.matrix),
                     xlab = ,
                     ylab = "Mean Yield", main = "Split plot design analysis"
,
                     plot.ci = T, # confidence interval
                     ci.l = Mu_Tret - SE_Tret, ci.u = Mu_Tret + SE_Tret,
                     las = 2, # for writing x axis names vertically for space
saving
                     col = "red")

text(bar.plot, 1, post.hoc.test.yield$groups[, 2],
     cex = .75, Pos = 3, col = "blue")



#RESULT:

YIELD groups

# Variety2:Nitrogen3 6.806667    a

# Variety2:Nitrogen4 6.490000    ab

# Variety3:Nitrogen4 6.346667    ab

# Variety1:Nitrogen4 6.070000    bc

# Variety3:Nitrogen5 6.056667    bc

# Variety1:Nitrogen5 5.923333    bc

# Variety2:Nitrogen5 5.596667    c

# Variety3:Nitrogen3 5.443333    cd

# Variety3:Nitrogen2 4.910000    de

# Variety1:Nitrogen2 4.760000    e

# Variety2:Nitrogen2 4.743333    e

# Variety1:Nitrogen3 4.636667    e

# Variety3:Nitrogen1 2.993333    f

# Variety2:Nitrogen1 2.936667    f

# Variety1:Nitrogen1 2.696667    f
```



Solution 02:

a). Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical technique used to simplify complex datasets by reducing their dimensions while retaining most of the original information. It transforms correlated variables into a smaller number of uncorrelated variables called **principal components**, which capture the maximum variance in the data.

Key Points:

1. **Dimensionality Reduction**: Makes large datasets easier to analyze and visualize.
2. **Variance Focus**: The first few components capture the most important patterns in the data.

3. **Applications**: Used for pattern recognition, data visualization, feature selection, and noise reduction.

b). The main purposes of principle component analysis in my study area-

My study area is Biotechnology in Agriculture. Principal Component Analysis (PCA) is a powerful statistical technique that finds widespread application in agricultural biotechnology. Here's a breakdown of its main purposes:

1. Dimensionality Reduction:

* Simplifying Complex Data: Agricultural datasets often involve numerous variables (e.g., soil properties, climate factors, crop yields, genetic markers). PCA transforms these variables into a smaller set of uncorrelated variables called principal components.

* Identifying Key Trends: By focusing on the most important principal components, researchers can identify the underlying patterns and relationships within the data, even with a large number of variables.

2. Data Visualization:

* Visualizing High-Dimensional Data: PCA allows researchers to visualize complex datasets in a lower-dimensional space (often 2D or 3D). This helps to identify clusters, outliers, and trends that might not be apparent in the original high-dimensional space.

* Understanding Relationships: By plotting data points based on their principal component scores, researchers can gain insights into the relationships between different variables and samples.

3. Feature Selection:

* Identifying Important Variables: PCA can help identify the most important variables that contribute to the overall variability in the data. This information can be used to select a subset of variables for further analysis or modeling.

* Improving Model Performance: By selecting only the most relevant features, researchers can improve the performance of machine learning models, such as those used for crop yield prediction or disease diagnosis.

4. Genotype and Phenotype Analysis:

* Understanding Genetic Variation: PCA can be used to analyze genetic data (e.g., single nucleotide polymorphisms) to identify patterns of genetic variation within and between populations.

* Linking Genotypes to Phenotypes: By combining genetic and phenotypic data, PCA can help to identify the genetic factors that contribute to specific traits of interest, such as disease resistance or yield potential.

5. Quality Control:

* Identifying Outliers: PCA can help to identify outliers or unusual samples in a dataset, which may be due to measurement errors or other factors.

* Monitoring Process Variability: PCA can be used to monitor the variability of processes in agricultural production, such as fermentation or food processing, to identify potential problems and improve quality.

By effectively utilizing PCA, agricultural biotechnologists can gain deeper insights into complex datasets, make more informed decisions, and ultimately improve agricultural productivity and sustainability.

c). Computation of the eigenvalue and eigenvector using the iris data based on R programming is given below-

Code:

```
# Load the data

iris_data <- read.csv("iris_Data.csv")


# Extract numerical columns (exclude the species column)

numeric_data <- iris_data[, 1:4]


# Compute the covariance matrix

cov_matrix <- cov(numeric_data)


# Compute eigenvalues and eigenvectors

eigen_results <- eigen(cov_matrix)


# Display the eigenvalues

cat("Eigenvalues:\n")

print(eigen_results$values)


# Display the eigenvectors

cat("\nEigenvectors:\n")

print(eigen_results$vectors)
```

Result:

Eigenvalues:

[1] 4.22824171 0.24267075 0.07820950 0.02383509

Eigenvectors:

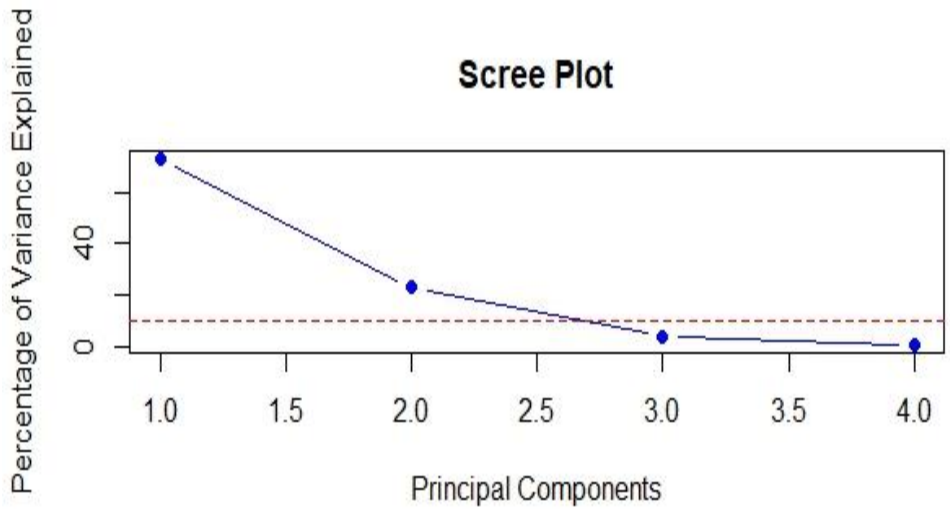
```
           [,1]       [,2]       [,3]       [,4]
 [1,]  0.36138659 -0.65658877  0.58202985  0.3154872
 [2,] -0.08452251 -0.73016143 -0.59791083 -0.3197231
 [3,]  0.85667061  0.17337266 -0.07623608 -0.4798390
 [4,]  0.35828920  0.07548102 -0.54583143  0.7536574
```

d). Construction of a scree plot and interpretation of how many principle components should be retained to interpret the iris dataset is given below:

#Code:

```
# Load the data
iris_data <- read.csv("iris_Data.csv")

# Extract numerical columns (exclude the species column)
numeric_data <- iris_data[, 1:4]

# Perform PCA
pca_result <- prcomp(numeric_data, scale. = TRUE) # Scale the data for standardization

# Compute the proportion of variance explained
explained_variance <- (pca_result$sdev^2) / sum(pca_result$sdev^2) * 100
```

```r
# Cumulative variance explained
cumulative_variance <- cumsum(explained_variance)

# Create a scree plot
plot(
  explained_variance,
  type = "b",
  xlab = "Principal Components",
  ylab = "Percentage of Variance Explained",
  main = "Scree Plot",
  pch = 19,
  col = "blue"
)
abline(h = 10, col = "red", lty = 2) # Optional: threshold for significance
# Add cumulative variance interpretation (optional)
cat("Explained Variance by Principal Components:\n")
print(explained_variance)
cat("\nCumulative Variance:\n")
print(cumulative_variance)
```



pca_result

Standard deviations (1, .., p=4):

[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation (n x k) = (4 x 4):

	PC1	PC2	PC3	PC4
Sepal.Length	0.5210659	-0.37741762	0.7195664	0.2612863
Sepal.Width	-0.2693474	-0.92329566	-0.2443818	-0.1235096
Petal.Length	0.5804131	-0.02449161	-0.1421264	-0.8014492
Petal.Width	0.5648565	-0.06694199	-0.6342727	0.5235971

Explained Variance by Principal Components:

[1] 72.9624454 22.8507618 3.6689219 0.5178709

Cumulative Variance:

[1] 72.96245 95.81321 99.48213 100.00000

Interpretation:

Scree Plot Insight:

In the scree plot, observed a sharp drop in variance explained from PC1 to PC2, and then the curve flattens after PC2. This suggests that **two principal components** would be adequate to interpret the dataset.

It can be choosen to retain **two components** for dimensionality reduction, as this will capture most of the variance without losing much information.

##The scree plot shows the **percentage of variance explained** by each principal component (PC):

1. **PC1** (first component):
    - Explains the largest variance (around 72.96% as per your data).
    - Represents the most significant pattern in the dataset.
2. **PC2** (second component):
    - Adds a significant amount of variance (around 22.85%, bringing the cumulative variance to 95.81%).
    - Together, PC1 and PC2 capture the majority of the information (approximately 96%).
3. **PC3 and PC4**:
    - Contribute very little additional variance (3.67% and 0.52%, respectively).
    - These components are not significant for explaining the variability in the data.

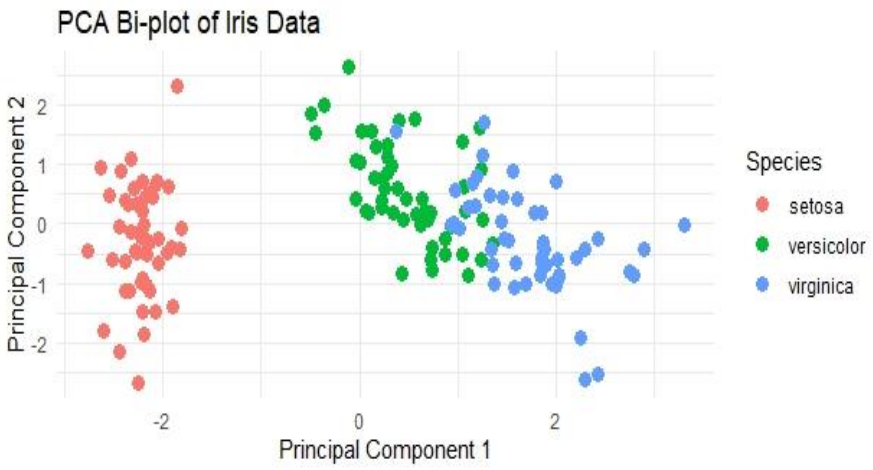
Retain PC1 and PC2: These two components explain around **96% of the total variance**, which is sufficient to summarize the dataset effectively.
Discard PC3 and PC4: These components add minimal new information and can be ignored in most analyses.

e). Construction a bi-plot for the iris data based on R programming and interpretion of the the results is given below:

```
# Load the iris dataset

data(iris)


# Perform PCA on the numerical columns of the iris dataset (excluding the Species column)

pca_result <- prcomp(iris[, 1:4], center = TRUE, scale. = TRUE)
```

```
# Plot the bi-plot
biplot(pca_result, main = "Bi-plot of Iris Data")

# Optionally, you can customize the plot with different colors for each species
library(ggplot2)
pca_data <- data.frame(pca_result$x, Species = iris$Species)

# Plot with ggplot2 for better customization
ggplot(pca_data, aes(PC1, PC2, color = Species)) +
  geom_point(size = 3) +
  labs(title = "PCA Bi-plot of Iris Data", x = "Principal Component 1", y = "Principal Component 2") +
  theme_minimal()
```

Ans:



Interpretion:

- **Species Labels**: Each point is labeled with its species (setosa, versicolor, or virginica), making it easy to see how the species are distributed along the principal components.
- **Cluster Separation**: To observe clear separation of points between species (e.g., setosa may cluster in one part of the plot while versicolor and virginica cluster in other parts), this suggests

that the principal components (PC1 and PC2) capture the variation that distinguishes these species.

- **Principal Components**: The arrows in the bi-plot represent the loadings of the original variables (sepal length, sepal width, petal length, and petal width) on the principal components.