Bangabandhu Sheikh Mujibur Rahman Agricultural University
EDGE_Batch-11
Mid Exam
Marks: 20 Time: 90 minutes
Name: Nusrat Jahan Naba……
Reg. No:…2018-05-4821……Dept.…Agricultural economics……..

Note: Submit the completed file to ***nazmol.stat.bioin@bsmrau.edu.bd*** and ***keyadas57@bsmrau.edu.bd*** with subject ***EDGE11_Mid_Your registration number_ Dept.***

1. Short Questions (5*1=05)

1. When comparing the means of two related groups (e.g., pre-test and post-test), the paired t..... test is used, assuming the data is normally distributed.
2. In regression analysis, the t test..... test is used to determine if the slope of a regression line is significantly different from zero, assuming normally distributed residuals.
3. In testing for normality, the Shapiro..... test is used to check if a data set follows a normal distribution, assuming that the data are parametric.
4. The Kruskal Wallis..... non-parametric test is used when comparing three or more independent groups.
5. The Spearman's rank..... correlation measures the degree of association between two variables when both are measured at the ordinal level.

2. For the given data set "Reg1",
   a) Present a correlation plot among independent variables using corrplot package.
   b) Check the assumptions and fit a multiple linear regression model.
   c) Apply forward selection method (stepwise regression) to find best subset of the independent variables.

3. A randomized complete block design was conducted considering four blocks, seven levels/treatments. Afterward, the yield of certain plant characteristics was observed. The data regarding this experiment were given in the file "RBDdata". 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 the treatment effects and interpret the results based on the ANOVA table.
   c) Perform a post-hoc test for the treatments and draw a bar diagram with lettering.

Answer to the question no. (2)

a)Code

```
install.packages("corrplot")
library("corrplot")
data<-read.csv('Reg1.csv')
correlations <- cor(data[,1:5])
cor_matrix <- cor(data[, c("x1", "x2", "x3", "x4")])
corrplot(cor_matrix, method = "square", type = "upper", lower.col = "black", number.cex = .7)
corrplot.mixed(cor_matrix,lower = "number", upper = "square" )
corrplot.mixed(cor_matrix, lower.col = "blue", number.cex = .7)
```

Result:

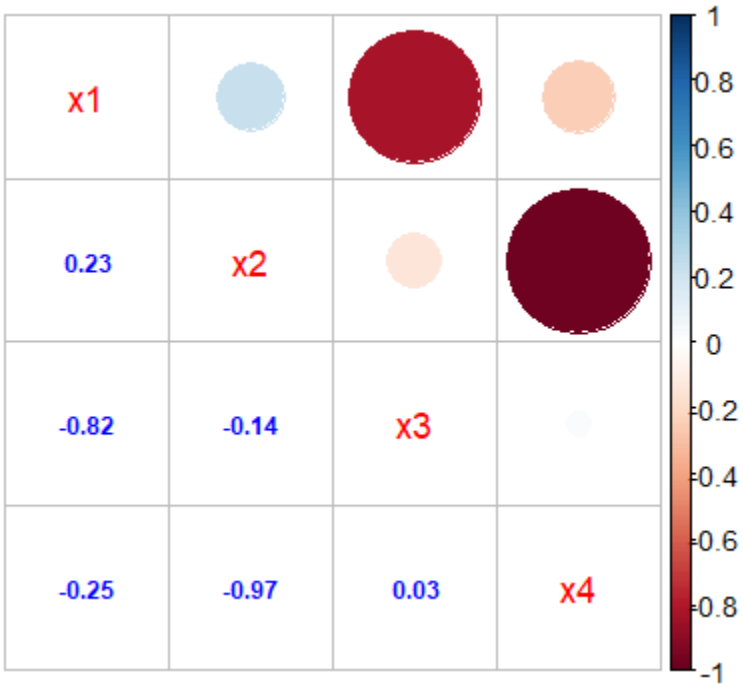


Fig : Corplot

b) Code

fit a multiple linear regression model

```
model <- lm(y ~ x1 + x2 + x3+x4, data=data)

summary(model)

AIC(model)

abline<-lm(y~.,data=data)

plot(model)
```

Return

Result:

Coefficient(intercept)	x1	x2	x3	x4
62.4054	1.5511	0.5102	0.1019	-0.1441

This table shows that, Intercept β_0 =62.4054. That's mean , if the other factor are zero, y will be 62.4054 unit.

β_1 = 1.5511, if other things remaining the same , then 1 unit increase in x1 will cause 1.5511 unit increase in y.

β_2 = 0.5102, if other things remaining the same , then 1 unit increase in x1 will cause 0.5102 unit increase in y.

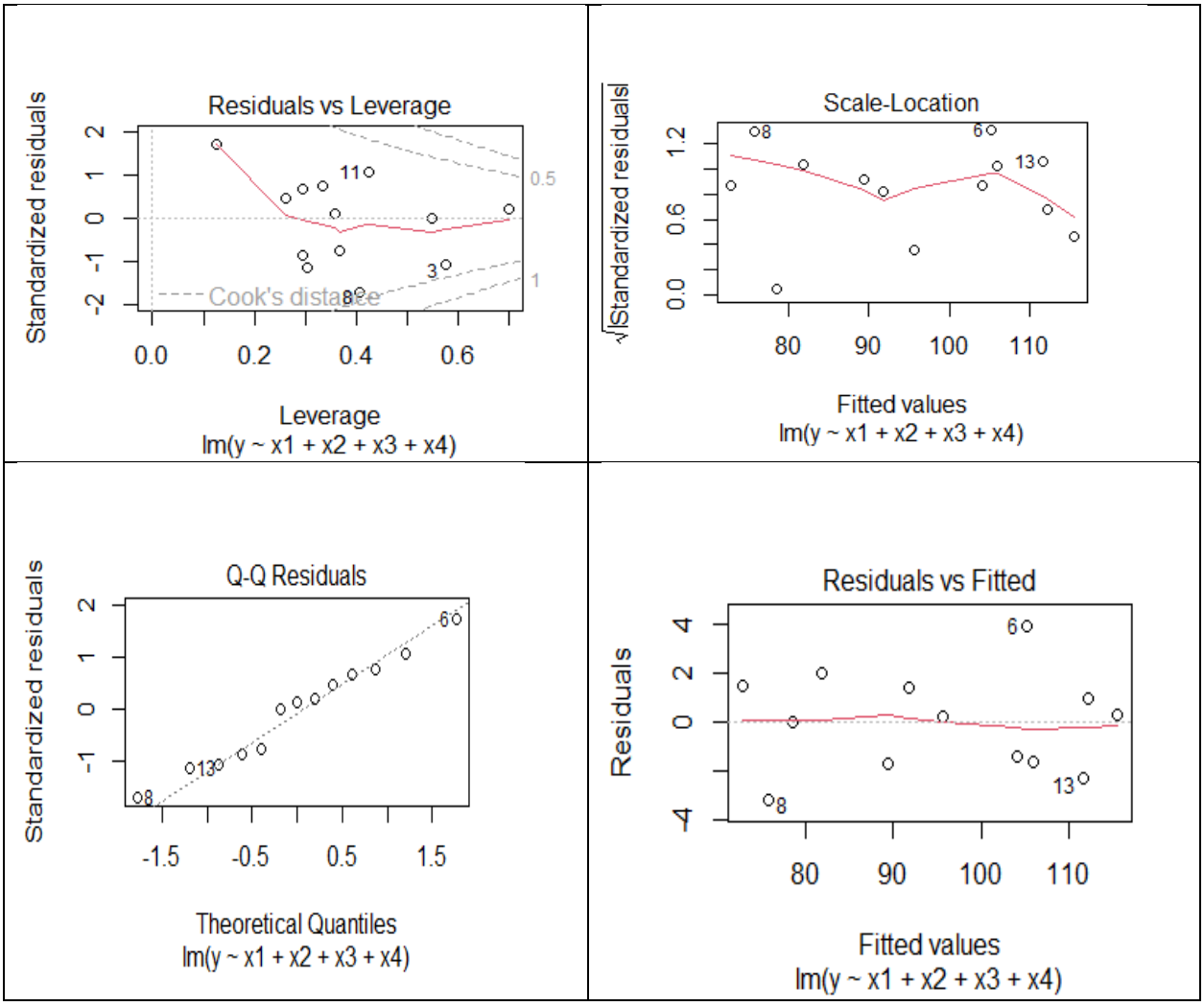
β_3 = 0.1019, if other thing remaining the same , then 1 unit increase in x1 will cause 0.1019 unit increase in y.

β_4 = -0.1441, if other thing remaining the same , then 1 unit increase in x1 will cause 0.1441 unit decrease in y.

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.1750 -1.6709  0.2508  1.3783  3.9254


oefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
x1            1.5511     0.7448   2.083   0.0708 .
x2            0.5102     0.7238   0.705   0.5009
x3            0.1019     0.7547   0.135   0.8959
x4           -0.1441     0.7091  -0.203   0.8441
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared:  0.9824,   Adjusted R-squared:  0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```



The residuals exhibit slight patterns, indicating potential non-linearity in the data. While the points generally align with the diagonal line, suggesting approximate normality, there are deviations at the extremes, particularly for observation 13. This suggests slight non-normality in the tails of the residuals.The red trend line shows a slight curvature, and the spread of residuals is not uniform across fitted values, indicating possible heteroscedasticity. Observation 13 has high leverage and standardized residuals, marking it as an influential point. Observation 11 is also near the Cook's distance threshold, suggesting it might have a notable influence as well.

c)Code:

```
library(MASS)

stepwise_model <- stepAIC(lm(y ~ 1, data = data),

                          scope = list(lower = ~1, upper = ~x1 + x2 + x3 + x4),
```

```
                direction = "forward")

summary(stepwise_model)
```

Result:

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.0919 -1.8016  0.2562  1.2818  3.8982
```

Estimate	Std. Error	t value	Pr(>|t|)
(Intercept) 71.6483	14.1424	5.066	0.000675
x4 -0.2365	0.1733	-1.365	0.205395
x1 1.4519	0.117	12.41	5.78E-07
x2 0.4161	0.1856	2.242	0.051687

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

Residual standard error: 2.309 on 9 degrees of freedom
Multiple R-squared:  0.9823,   Adjusted R-squared:  0.9764
F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08
```

Interpretation:

Intercept β_0 =71.6483. That’s mean, if the other factor (x1, x2, x4) is zero, y will be 71.6483 unit.

• β_1 = -0.2365; Holding x1, x2, x4 constant, a one-unit increase in x4 is associated with an average decrease of -0.2365 units in y.

• β_2 = 1.4519; Holding x2, x4 constant, a one-unit increase in x1 is associated with an average increase of 1.4519 units in y.

• β_3 = 0.4161; Holding x1, x4 constant, a one-unit increase in x2 is associated with an average increase of 0.4161 units in y.

•The adjusted R-squared value of 0.9764 indicates that 97.64% of the variance in the dependent variable (y) is explained by the independent variables in the model, after accounting for the number of predictors. This high value suggests that the model fits the data very well, and the predictors are highly effective in explaining the variability in y.

Null Hypothesis (H0): All regression coefficients are zero ($\beta1=\beta2=\beta3 =0$).

Alternative Hypothesis (Ha): At least one regression coefficient is not zero ($\beta j\neq0\beta j=0$ for at least one jj).

Since the p-value is extremely small ($p<0.001p < 0.001p<0.001$), we reject the null hypothesis (H0) at any reasonable significance level (e.g., 0.05, 0.01, 0.001). This indicates that the overall regression model is statistically significant, and at least one of the predictors (x1, x2, x4) has a significant effect on the dependent variable y.

Answer to the question no. (3)

a)Code:

```
Data.RCBD<-read.csv("RBDdata.CSV")

Data.RCBD<-Data.RCBD[,2:4]

Rep<-c("Rep1","Rep2","Rep3","Rep4")

Treat<-c("Treat1","Treat2","Treat3","Treat4","Treat5","Treat6","Treat7")

r<-length(Rep)

t<-length(Treat)

Block<-gl(r,t,r*t,factor(Rep))

Treat<-gl(t,1,r*t,factor(Treat))

ANOVA.RCBD<-aov(YIELD~Block+Treat,

        data=Data.RCBD)

summary(ANOVA.RCBD)
```

Result:

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Block	3	1742	580.7	29.61	3.55e-07***
Treat	6	12148	2024.6	103.24	5.96e-13***
Residuals	18	353	19.6		

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

b)The null hypothesis based on treatment effects:
H_0: $\mu1=\mu2=\mu3=\cdots=\mu t$ & the alternative hypothesis is H1 which is opposite to H_0.
H_1: $\mu1 \neq \mu2 \neq \mu3 \neq \ldots\ldots \neq \mu t$

Interpretation:

From the ANOVA table, p value is less than 0.05. So, the null hypothesis is rejected. It indicates that there is significant differences among treatments effects.

c)Code:

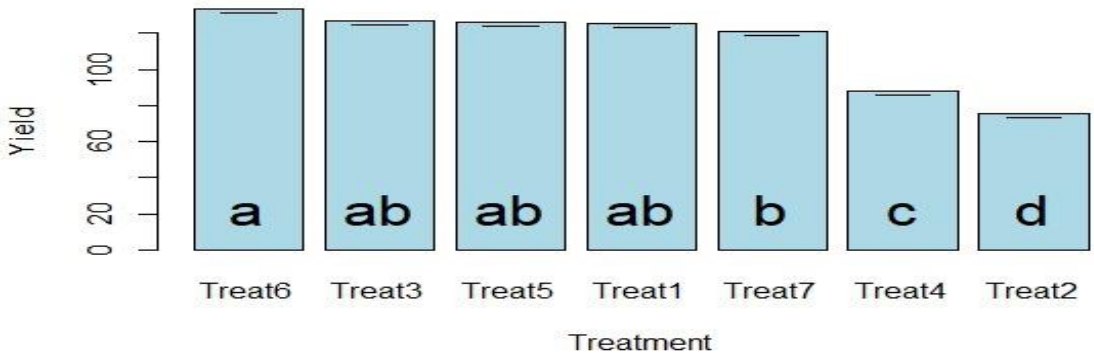
```
library(agricolae)
PostHoc.Test<-with(Data.RCBD,HSD.test(YIELD,Treat,DFerror=18,MSerror=19.6))
Mutplcom.TreatFact<-with(Data.RCBD,HSD.test
                  (YIELD,Treat,DFerror=18,MSerror=19.6))
library(gplots)
Treat.SE.Mat<-Mutplcom.TreatFact$means[,"se"]
Treat.Mean<-Mutplcom.TreatFact$groups
Mean.Mat<-Mutplcom.TreatFact$means
Mean.Mat<-Mean.Mat[order(-Mean.Mat$YIELD)]
Treat.Treat.Mean<-Treat.Mean$YIELD
Treat.SE<-Mean.Mat[, "se"]
Treat.SE.Mat<-Mutplcom.TreatFact$means[order(Mutplcom.TreatFact$means[,"se"])]
Barplot.Se<-barplot2(Treat.Treat.Mean,
                  names.arg = rownames(Treat.Mean),
                  xlab="Treatment",ylab="Yield",
                  horix=F,plot.ci = T,
                  ci.l=Treat.Treat.Mean-Treat.SE,
                  ci.u=Treat.Treat.Mean-Treat.SE,
                  col="lightblue")
text(Barplot.Se, 7,Treat.Mean$groups, cex=2,
    pos = 3, col= "black"
```

Result:

YIELD	groups	Lettering
Treat6	**133.25**	**a**
Treat3	**127.00**	**ab**
Treat5	**125.75**	**ab**
Treat1	**125.00**	**ab**
Treat7	**121.00**	**b**
Treat4	**87.75**	**c**
Treat2	**75.25**	**d**

Interpretation:

From this test, I can say that Treat6, Treat3, Treat5, Treat1 get the same letter which is let tering ‘a’. So, these four treatments are better for getting better yield but Treat6 is the bes t in all of these treatments.


Yield
0
20
60
100
a
ab
ab
ab
b
c
d
Treat6
Treat3
Treat5
Treat1
Treat7
Treat4
Treat2
Treatment