**#Topic-Group Assignment**

**#Subject- R Programming**

**#Group -P**

**#GROUP MEMBERS:-**

Atul Kumar, Roll No. -21227707058

Arkajeet Das, Roll No.-21227707049

SK Jane Alam,Roll No.-21227707254

Shreya Ghosh,Roll No.-21227707240

**#Submitted to-Dr. Devesh Birwal**

**#step1 : import dataset**

library(readxl)

group\_p <- read\_excel("group\_p.xlsx")

View(group\_p)

**#step2 : attach dataset, view and summarize the data**

attach(group\_p)

View(group\_p)

summary(group\_p)

**#step3 : extracting data\_3 using rows and columns**

**#row**

group\_p[10,]

group\_p[3,]

**#column**

group\_p[,5]

group\_p[,2]

**#both row & column**

group\_p[5,7]

**#extract column/object by names**

group\_p$Age

group\_p$AnnualIncome

group\_p$kids\_6

group\_p$FamilyMember

group\_p$MilkConsumptionltrs

dim(group\_p)

## 92 19

ncol(group\_p)

## 19

nrow(group\_p)

## 92

head(group\_p,5)

tail(group\_p,4)

str(group\_p)

**#step4 : finding mean,median,mode,sd,var,max,min**

mean(group\_p$Age)

## NA

mean(group\_p$INCOME)

## 7.73913

mean(group\_p$FamilyMember)

## 5.271739

mean(group\_p$MilkConsumptionltrs)

## 2.98913

mode(group\_p$Age)

mode(group\_p$INCOME)

mode(group\_p$FamilyMember)

mode(group\_p$MilkConsumptionltrs)

median(group\_p$Age)

## NA

median(group\_p$INCOME)

## 5.5

median(group\_p$FamilyMember)

## 4.5

median(group\_p$MilkConsumptionltrs)

## 2

sd(group\_p$Age)

## NA

sd(group\_p$INCOME)

## 6.088847

sd(group\_p$FamilyMember)

## 6.088847

sd(group\_p$MilkConsumptionltrs)

## 1.699997

var(group\_p$Age)

## NA

var(group\_p$INCOME)

## 37.07406

var(group\_p$FamilyMember)

## 5.167105

var(group\_p$MilkConsumptionltrs)

## 2.88999

max(group\_p$Age)

## NA

max(group\_p$INCOME)

## 18

max(group\_p$FamilyMember)

## 16

max(group\_p$MilkConsumptionltrs)

## 8

min(group\_p$Age)

##NA

min(group\_p$INCOME)

## 18

min(group\_p$FamilyMember)

## 16

min(group\_p$MilkConsumptionltrs)

## 8

**#step5 :simple linear regression in group\_p**

names(group\_p)

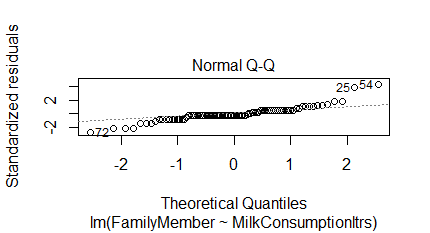
class(FamilyMember)

## "numeric"

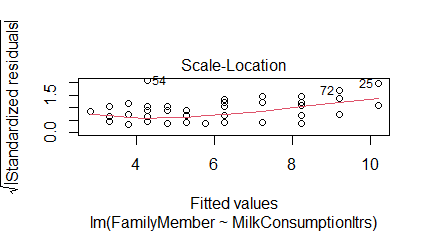
class(MilkConsumptionltrs)

## "numeric"

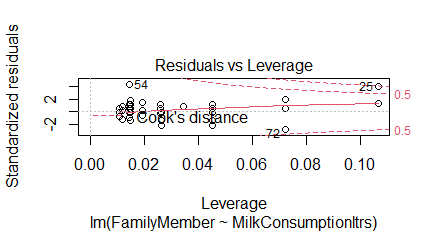
plot(FamilyMember,MilkConsumptionltrs,main="Scatterplot")



cor(FamilyMember,MilkConsumptionltrs)



mod<-lm(FamilyMember~MilkConsumptionltrs)



summary(mod)

## Call:

lm(formula = FamilyMember ~ MilkConsumptionltrs)

Residuals:

Min 1Q Median 3Q Max

-4.2107 -0.3004 -0.2824 0.6996 6.6996

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.33623 0.32849 7.112 2.64e-10

MilkConsumptionltrs 0.98206 0.09565 10.267 < 2e-16

(Intercept) \*\*\*

MilkConsumptionltrs \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.551 on 90 degrees of freedom

Multiple R-squared: 0.5394, Adjusted R-squared: 0.5343

F-statistic: 105.4 on 1 and 90 DF, p-value: < 2.2e-16

attributes(mod)

## "coefficients" "residuals" "effects"

[4] "rank" "fitted.values" "assign"

[7] "qr" "df.residual" "xlevels"

[10] "call" "terms" "model"

$class

[1] "lm"

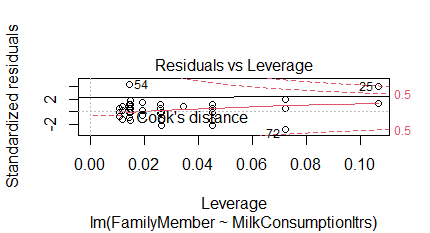
mod$coefficients

## (Intercept) MilkConsumptionltrs

2.3362265 0.9820624

abline(mod)

abline(mod, col=2, lwd=3)



confint(mod)

## 2.5 % 97.5 %

(Intercept) 1.6836157 2.988837

MilkConsumptionltrs 0.7920276 1.172097

confint(mod, level=0.99)

## 0.5 % 99.5 %

(Intercept) 1.4717730 3.200680

MilkConsumptionltrs 0.7303407 1.233784

anova(mod)

## Analysis of Variance Table

Response: FamilyMember

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

MilkConsumptionltrs 1 253.64 253.639 105.41 < 2.2e-16

Residuals 90 216.57 2.406

MilkConsumptionltrs \*\*\*

Residuals

---

Signif. codes:

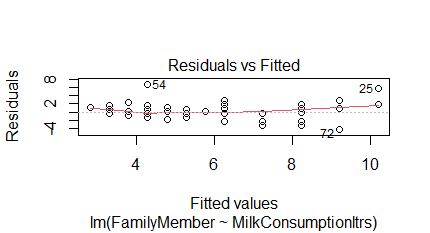
0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**#step6 : checking linear regression in group\_p**

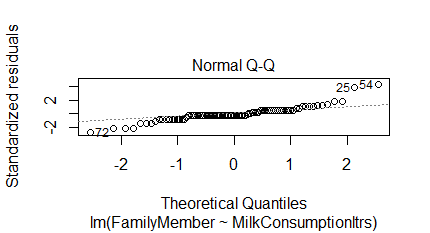
plot(INCOME,MilkConsumptionltrs)

plot(mod)

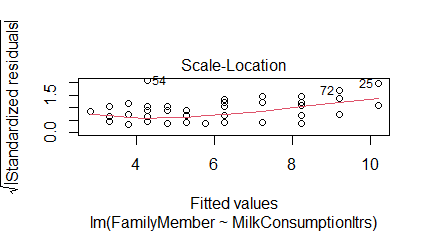
par(mfrow=c(1,1))



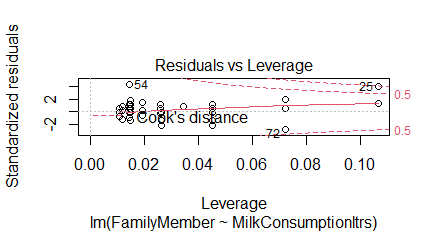
plot(FamilyMember,MilkConsumptionltrs)



mod2<-lm(FamilyMember,MilkConsumptionltrs)



abline(mod2)



**#step7 : hypothesis testing in group\_p**

FamilyMember = rnorm(10)

MilkConsumptionltrs = rnorm(10)

t.test(FamilyMember,MilkConsumptionltrs)

## data: FamilyMember and MilkConsumptionltrs

t = 0.43743, df = 17.231, p-value = 0.6672

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.889638 1.355620

sample estimates:

mean of x mean of y

0.18516094 -0.04782985

t.test(FamilyMember,MilkConsumptionltrs, var.equal=TRUE)

## data: FamilyMember and MilkConsumptionltrs

t = 0.43743, df = 18, p-value = 0.667

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.8860471 1.3520287

sample estimates:

mean of x mean of y

0.18516094 -0.04782985

t.test(FamilyMember, mu =5)

## data: FamilyMember

t = -14.395, df = 9, p-value = 1.613e-07

alternative hypothesis: true mean is not equal to 5

95 percent confidence interval:

-0.5715114 0.9418333

sample estimates:

mean of x

0.1851609

t.test(MilkConsumptionltrs, mu=5)

## data: MilkConsumptionltrs

t = -12.178, df = 9, p-value = 6.793e-07

alternative hypothesis: true mean is not equal to 5

95 percent confidence interval:

-0.9855266 0.8898669

sample estimates:

mean of x

-0.04782985

t.test(FamilyMember,mu=5, alternative= 'greater')

## data: FamilyMember

t = -14.395, df = 9, p-value = 1

alternative hypothesis: true mean is greater than 5

95 percent confidence interval:

-0.4279998 Inf

sample estimates:

mean of x

0.1851609

t.test(MilkConsumptionltrs,mu=5, alternative= 'greater')

## data: MilkConsumptionltrs

t = -12.178, df = 9, p-value = 1

alternative hypothesis: true mean is greater than 5

95 percent confidence interval:

-0.8076816 Inf

sample estimates:

mean of x

-0.04782985

**#TWO SAMPLE U-TEST**

wilcox.test(FamilyMember,MilkConsumption)

## data: FamilyMember and MilkConsumptionltrs

W = 54, p-value = 0.7959

alternative hypothesis: true location shift is not equal to 0

**#ONE SAMPLE U-TEST**

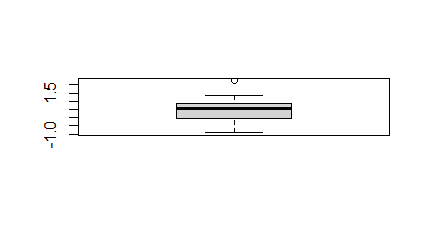
wilcox.test(MilkConsumptionltrs, exact=FALSE)

## data: MilkConsumptionltrs

V = 25, p-value = 0.8385

alternative hypothesis: true location is not equal to 0

boxplot(FamilyMember)



**#Ho:mu<8**

**#one-sided 95% confidence interval for mu**

t.test(MilkConsumptionltrs, mu=8,alternative="less",conf.level=0.95)

## data: MilkConsumptionltrs

t = -19.415, df = 9, p-value = 5.897e-09

alternative hypothesis: true mean is less than 8

95 percent confidence interval:

-Inf 0.7120219

sample estimates:

mean of x

-0.04782985

**#two-sided**

t.test(FamilyMember , mu=8,alt="two.sided",conf=0.99)

## data: FamilyMember

t = -23.363, df = 9, p-value = 2.296e-09

alternative hypothesis: true mean is not equal to 8

99 percent confidence interval:

-0.9018815 1.2722034

sample estimates:

mean of x

0.1851609

TEST<-t.test(FamilyMember, mu=8,alt="two.sided",conf=0.99)

TEST

## data: FamilyMember

t = -23.363, df = 9, p-value = 2.296e-09

alternative hypothesis: true mean is not equal to 8

99 percent confidence interval:

-0.9018815 1.2722034

sample estimates:

mean of x

0.1851609

attributes(TEST)

TEST$conf.int

## -0.9018815 1.2722034

attr(,"conf.level")

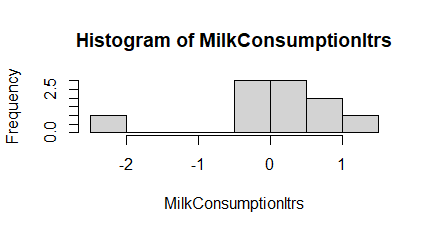
[1] 0.99

TEST$p.value

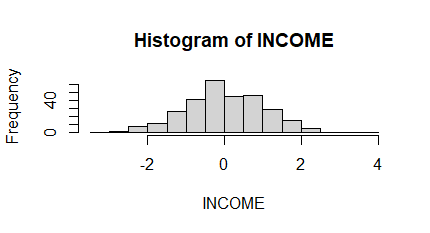
## 2.296139e-09

**#step8: plotting of graphs of group\_p**

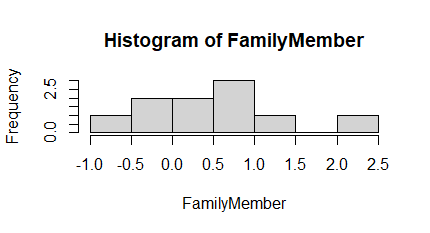
hist(MilkConsumptionltrs)



hist(INCOME)



hist(FamilyMember)



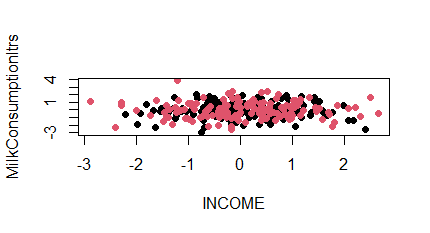
set.seed(1)

INCOME<-rnorm(300)

MilkConsumptionltrs<-rnorm(300)

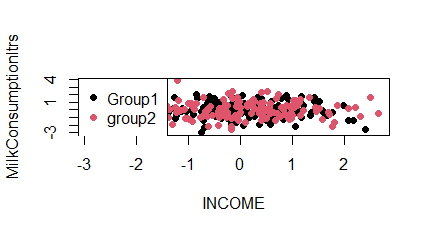
group<-rep(1:2,each=150)

plot(INCOME,MilkConsumptionltrs, col=group ,pch=16)



legend("topleft",

legend=c("Group1","group2"),col=1:2,pch=16)

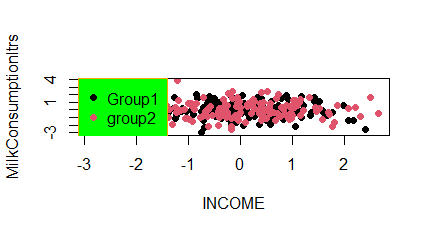


plot(FamilyMember,MilkConsumptionltrs, col=group ,pch=16)

legend("topleft",

legend=c("Group1","group2"),col=1:2,pch=16,

box.col="orange",bg="green")

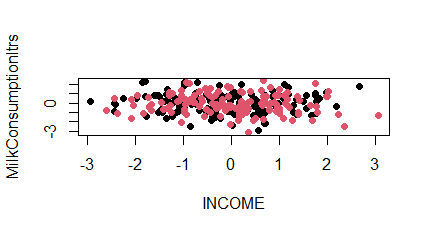


INCOME<-rnorm(300)

MilkConsumptionltrs<-rnorm(300)

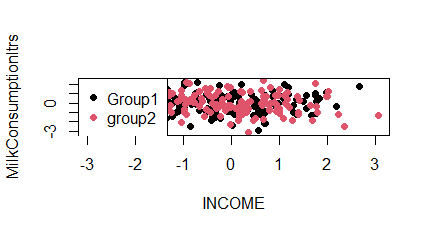
group<-rep(1:2,each=150)

plot(INCOME,MilkConsumptionltrs, col=group ,pch=16)



legend("topleft",

legend=c("Group1","group2"),col=1:2,pch=16)

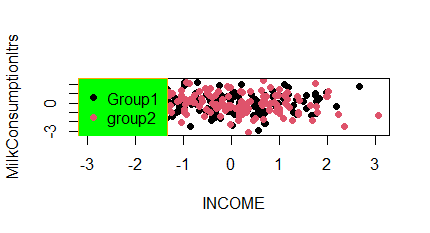


plot(INCOME,MilkConsumptionltrs, col=group ,pch=16)

legend("topleft",

legend=c("Group1","group2"),col=1:2,pch=16,

box.col="orange",bg="green")



#Analysis

**R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Adjusted R2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs. R2 tends to optimistically estimate the fit of the linear regression. In our data we find the Adjusted R Square to be 0.51 indicating that only 50% of the variability in the outcome data can be explained by the model**