Assignment-04

Sumanth Donthula

2022-10-22

Question 1)

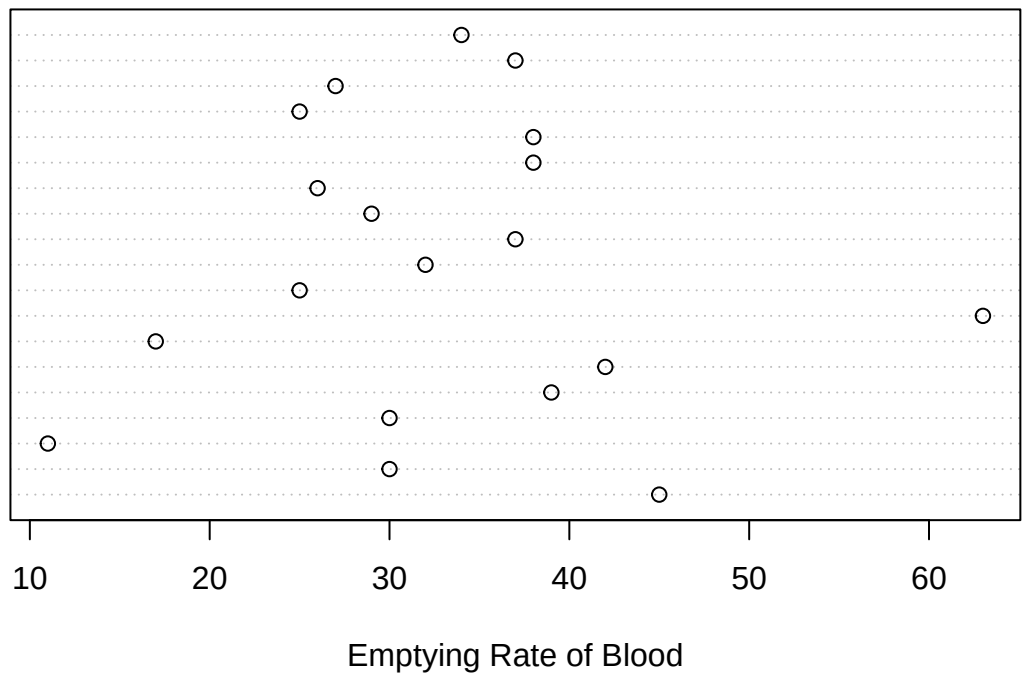
1.a)

X1 is almost Normally Distributed

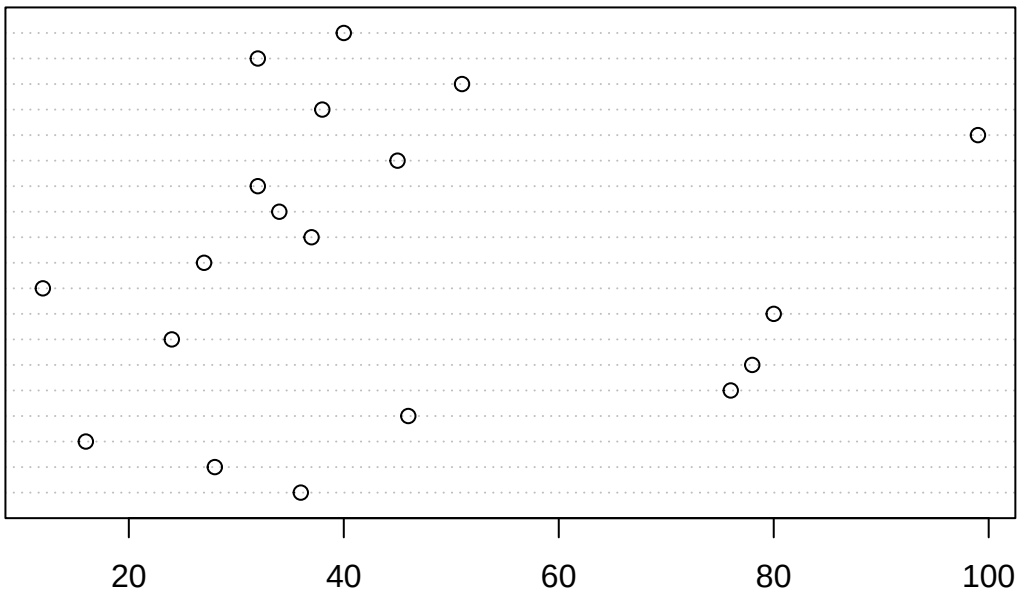
X2 is Right Skewed

X3 is Randomly distributed

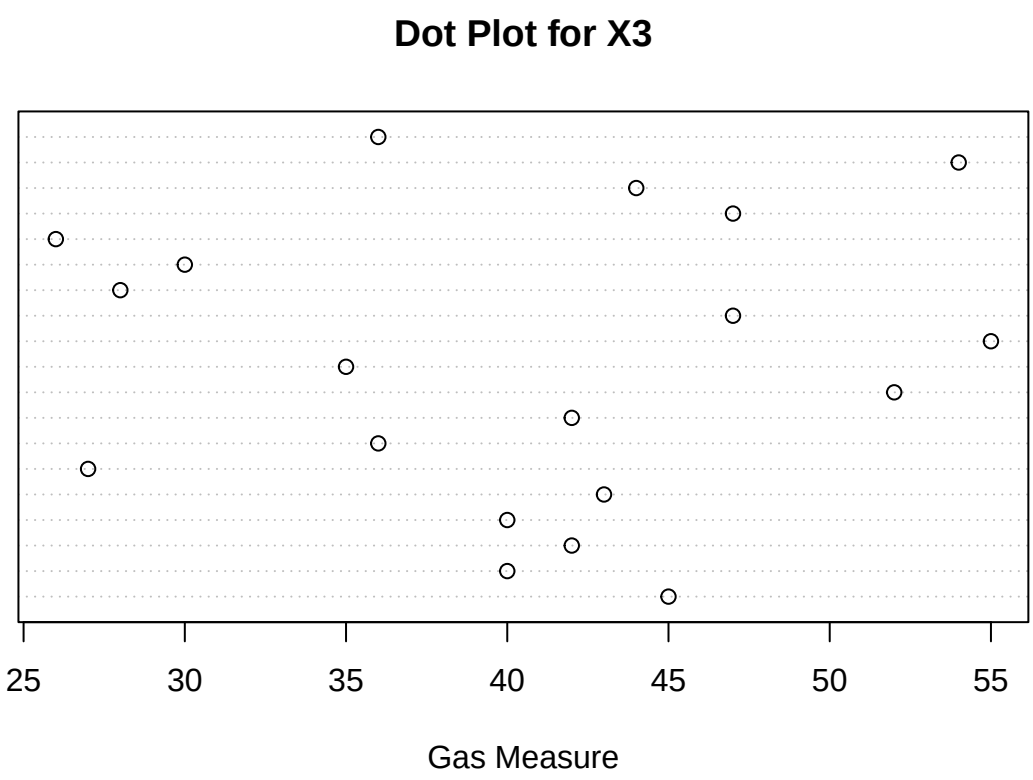
```
Data1=read.table("As4Q1.txt", header = FALSE, sep = "")
colnames(Data1)=c("Y","X1","X2","X3")
Y=Data1$Y
X1=Data1$X1
X2=Data1$X2
X3=Data1$X3

dotchart(X1, main="Dot Plot for X1",xlab="Emptying Rate of Blood")
```




```
dotchart(X2, main="Dot Plot for X2",xlab="Ejection Rate of Blood")
```




```
dotchart(X3, main="Dot Plot for X3",xlab="Gas Measure")
```



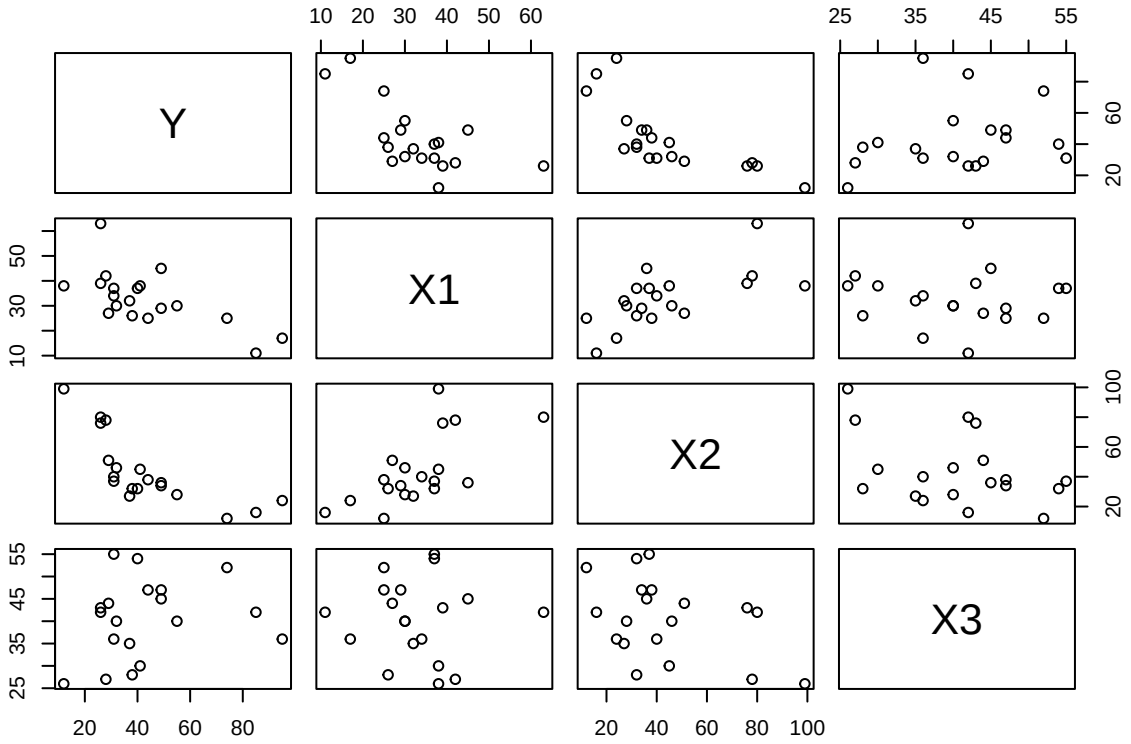
1.b)

Y is correlated with X1 and X2 but not correlated with X3 from scatter plot.

Based on correlation matrix, X1 and X2 have high colinearity. No correlation between X1 and X3 but X2 and X3 have less correlation between them.

so from the aboove inferences multicolineaity exists.

```
par(mfrow=c(3,2))
pairs(Data1)
```




```
cor(Data1[-1])
```

```
##             X1         X2          X3
## X1  1.00000000  0.6528513 -0.04613927
## X2  0.65285127  1.0000000 -0.42348025
## X3 -0.04613927 -0.4234803  1.00000000
```

1.c)

The value of coeficient of X3 is 0.07 which seems less significant.

```
Model1=lm(Y~X1+X2+X3)
Model1
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3)
##
## Coefficients:
## (Intercept)           X1           X2           X3
##    87.18750     -0.56448     -0.51315     -0.07196
```

```
summary(Model1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.075 -12.064  -0.988   7.707  32.315
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 87.18750   21.55246   4.045  0.00106 **
## X1          -0.56448    0.42791  -1.319  0.20691
## X2          -0.51315    0.22449  -2.286  0.03723 *
## X3          -0.07196    0.45457  -0.158  0.87633
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.42 on 15 degrees of freedom
## Multiple R-squared:  0.6141, Adjusted R-squared:  0.5369
## F-statistic: 7.957 on 3 and 15 DF,  p-value: 0.002083
```

Question 2)

2.a)

The three best hierarchical subset regression models are

```
library(leaps)
Subs = regsubsets(Y~X1 + X2 + X3 + I(X1^2)+I(X2^2) + I(X3^2) + I(X1*X2) + I(X1*X3)+ I(X2*X3), method =

SubsDf = data.frame(features = summary(Subs)$which, adjr2 = summary(Subs)$adjr2)

colnames(SubsDf) = c('Intercept', 'X1', 'X2', 'X3', 'I(X1^2)', 'I(X2^2)', 'I(X3^2)', 'I(X1 * X2)', 'I(X

SubsDf = SubsDf[order(-SubsDf$AdjR_2),][1:3,]
SubsDf
```

```
##   Intercept   X1    X2    X3 I(X1^2) I(X2^2) I(X3^2) I(X1 * X2) I(X1 * X3)
## 4      TRUE TRUE  TRUE FALSE    TRUE    TRUE   FALSE      FALSE      FALSE
## 3      TRUE TRUE  TRUE FALSE   FALSE   FALSE   FALSE       TRUE      FALSE
## 6      TRUE TRUE FALSE  TRUE   FALSE   FALSE    TRUE       TRUE       TRUE
##   I(X2 * X3)    AdjR_2
## 4      FALSE 0.7506701
## 3      FALSE 0.7506631
## 6       TRUE 0.7381101
```

2.b)

There is no much difference in R Adjusted squared values.

Question 3)

3.a)

The regression model is $Y = 1.023 + 0.965 * X1 + 0.629 * X2 + 0.676 * X3$

```
Data2=read.table("As4Q2.txt", header = FALSE, sep = "")
colnames(Data2)=c("Y","X1","X2","X3");
Y=Data2$Y
X1=Data2$X1
X2=Data2$X2
X3=Data2$X3

Mod2=lm(Y~X1+X2+X3)
Mod2
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3)
##
## Coefficients:
## (Intercept)           X1           X2           X3
##      1.0233       0.9657       0.6292       0.6760
```

3.b)

Hypothesis test

H0: Beta1=Beta2=Beta3=0

Ha: At least one of the coeficient is not 0

Since the P value from the summary of model is 7.82e-12 which is less than 0.05 we reject null hypothesis. So, there is regreession relation between sales and the predictors.

```
summary(Mod2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4217 -0.9115  0.0703  1.1420  3.5479
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0233     1.2029   0.851   0.4000
## X1            0.9657     0.7092   1.362   0.1809
## X2            0.6292     0.7783   0.808   0.4237
## X3            0.6760     0.3557   1.900   0.0646 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.825 on 40 degrees of freedom
## Multiple R-squared:  0.7417, Adjusted R-squared:  0.7223
## F-statistic: 38.28 on 3 and 40 DF,  p-value: 7.821e-12
```

3.c)

From the summary of model

Hypothesis test

H0 : Betak = 0

Ha : Betak <> 0

Since the T values of all coeficeints are greater than Ttest(2.021) we note that the null hypothesis is false and all the coeficeints are significant.

The conclusions of this test does not correspond to the one obtained in part (b).

```
summary(Mod2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4217 -0.9115  0.0703  1.1420  3.5479
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0233     1.2029   0.851   0.4000
## X1            0.9657     0.7092   1.362   0.1809
## X2            0.6292     0.7783   0.808   0.4237
## X3            0.6760     0.3557   1.900   0.0646 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.825 on 40 degrees of freedom
## Multiple R-squared:  0.7417, Adjusted R-squared:  0.7223
## F-statistic: 38.28 on 3 and 40 DF,  p-value: 7.821e-12
```

```
t_ratio <- qt(0.975, nrow(Data2) - 4)
t_ratio
```

```
## [1] 2.021075
```

3.d)

The correlation matrix is as follows

```
cor(Data2[-1])
```

```
##           X1        X2        X3
## X1 1.0000000 0.9744313 0.3759509
## X2 0.9744313 1.0000000 0.4099208
## X3 0.3759509 0.4099208 1.0000000
```

3.e)

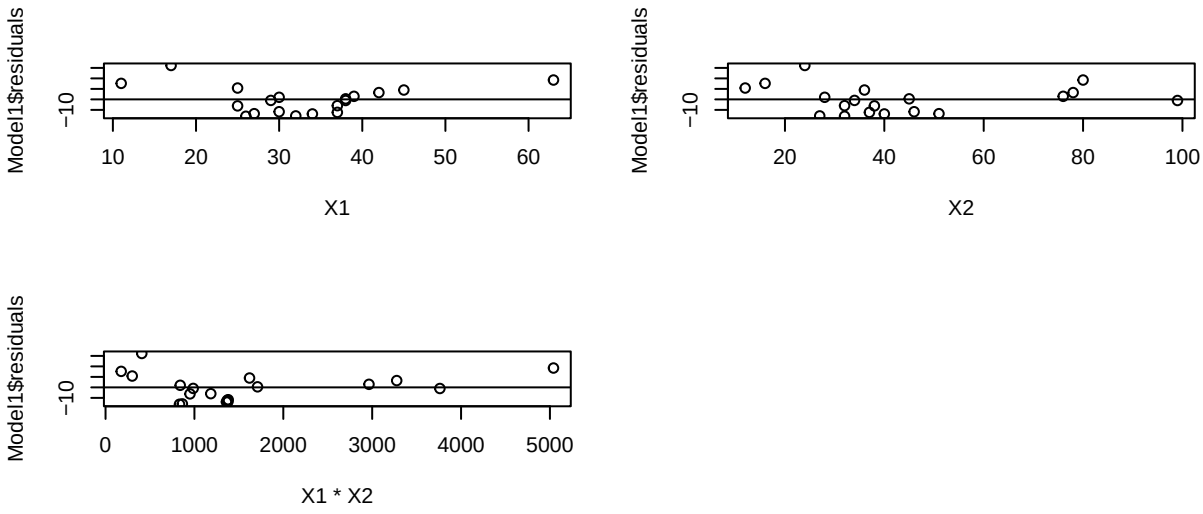
From b,c and d we found that there is correlation in data and increase in X1 by 1000 wont be a good thing keeping other predictors constant.

Question 4)

4.a)

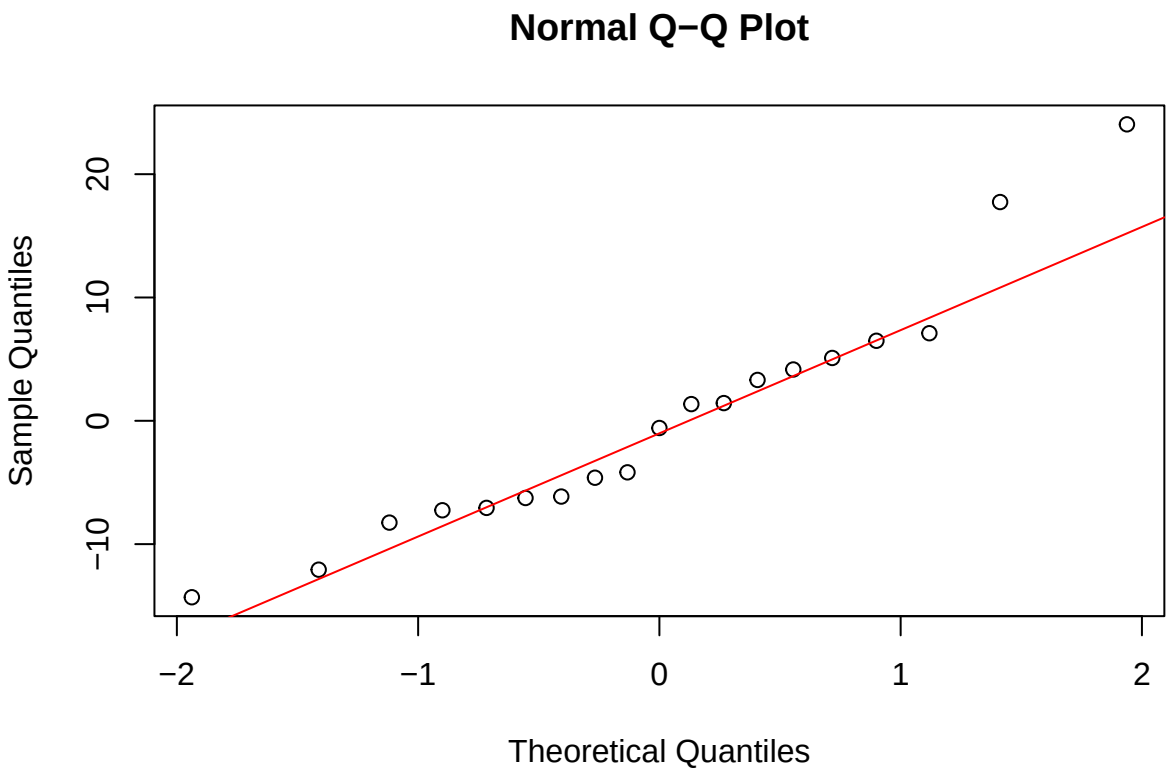
From the plots we observe that the residues are spread over zero line with some variance and it can also be seen there are some outliers which are at extreme values of X.

```
Y=Data1$Y
X1=Data1$X1
X2=Data1$X2
X3=Data1$X3

Model3=lm(Y~X1+X2+X1*X2)

par(mfrow=c(3,2))

plot(X1,Model1$residuals)
abline(0,0)
plot(X2,Model1$residuals)
abline(0,0)
plot(X1*X2,Model1$residuals)
abline(0,0)
```



4.b) The residuals seems to be normally distributed.

```
qqnorm(Model3$residuals)
qqline(Model3$residuals, col = "red")
```



4.c)

since vif of all predictors are >1 there will be multicolinearity.

```
library(car)
```

```
## Loading required package: carData
```

```
vif(Model3)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif

##        X1        X2     X1:X2
##  5.431477 11.639560 22.474469
```

4.d) Hypothesis test:

H0 : Observation is an Outlier

Ha : Observation is not an Outlier

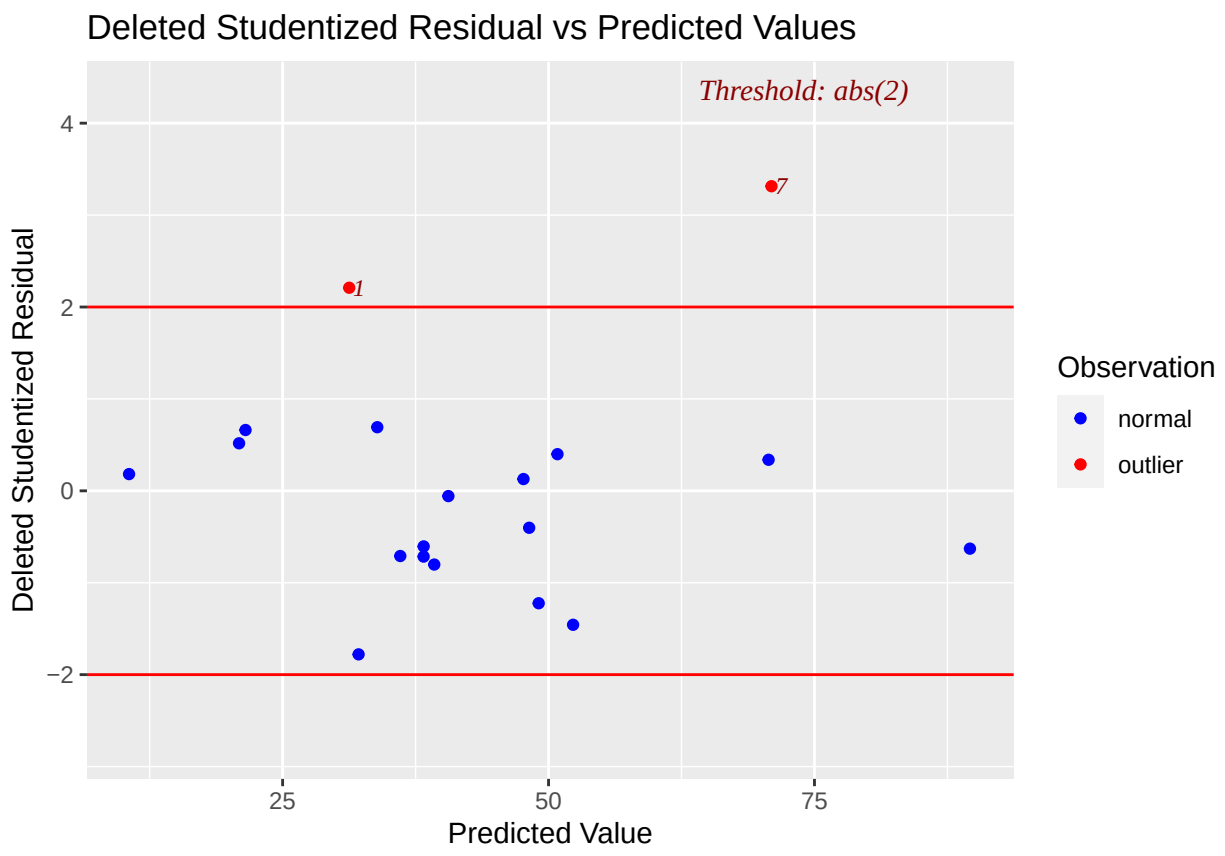
We find that no observations are outliers with a level of significance of alpha = 0.05

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
Dr =ols_plot_resid_stud_fit(model = Model3)
```




```
Dr = Dr$data[,'dsr']
Ttest= qt(1-0.05/(2*nrow(Data2)),25-3)
Dr[Dr>Ttest]
```

```
## numeric(0)
```

4.e)

The observations below has high leverage and values are greater than 2xmean(Hmat) and are outliers. The results are consistent with 9.13 a. Because these values are located far on dot plot and have large values of X1 and X2.

```
Hmat <- influence(Model3)
Hmat$hat[(Hmat$hat > mean(Hmat$hat)*2)]
```

```
##         3         8        15
## 0.5388667 0.8782787 0.4798210
```

4.f)

Conclusion:

Dffits:

7 and 8 have high Dffits greater than 1 which says they are more influential.

Dfbeta:

7 and 8 have high Dfbetas greater than 1 which says they are more influential.

Cooks Distance:

The percentile value for 8 is significant in cooks distance which is more influetial.

```
library(car)

dffits(Model3)[c(3,7,8,15)]
```

```
##          3          7          8         15
## -0.6801824  1.7485509 -4.7797848  0.1748573
```

```
dfbetas(Model3)[c(3,7,8,15),]
```

```
##    (Intercept)          X1          X2       X1:X2
## 3   -0.6519371  0.59191342  0.43337176 -0.48191103
## 7    1.4541305 -1.27760852 -0.74151968  0.84752328
## 8   -1.5469080  1.18662253  3.16226530 -3.28579003
## 15  -0.0155059 -0.03525106  0.07714703 -0.01569977
```

```
dfbetas(Model3)[c(3,7,8,15),]
```

```
##    (Intercept)          X1          X2       X1:X2
## 3   -0.6519371  0.59191342  0.43337176 -0.48191103
## 7    1.4541305 -1.27760852 -0.74151968  0.84752328
## 8   -1.5469080  1.18662253  3.16226530 -3.28579003
## 15  -0.0155059 -0.03525106  0.07714703 -0.01569977
```

```
pf(cooks.distance(Model3)[c(3,7,8,15)],4,nrow(Data2)-4 )
```

```
##            3            7            8           15
## 0.0256289784 0.2346505878 0.9976684825 0.0001385187
```