K. J. Somaiya College of Engineering, Mumbai – 77

(Autonomous College Affiliated to University of Mumbai)

Associate Analytics I

Roll No: **1812054 A3** Name: **Karan Satish shah**

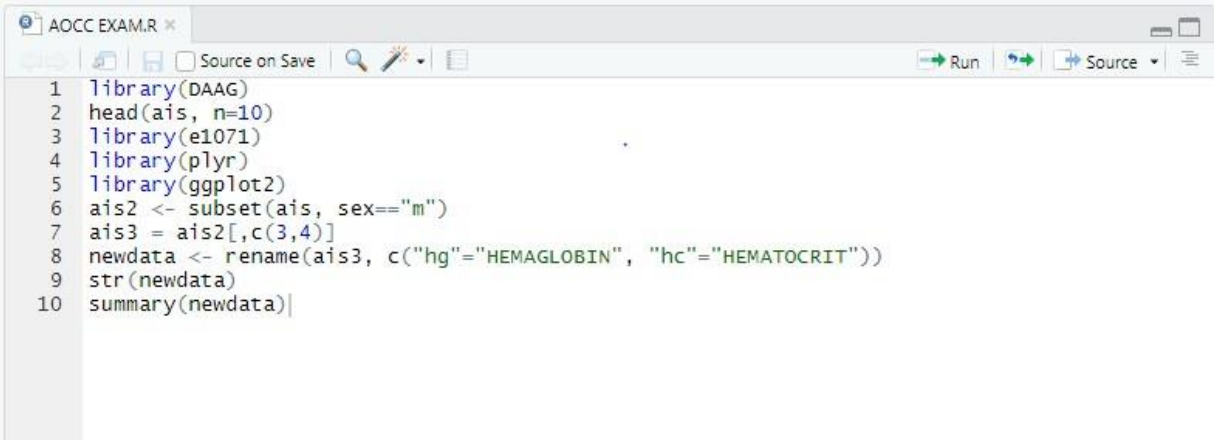
Branch: **ETRX**

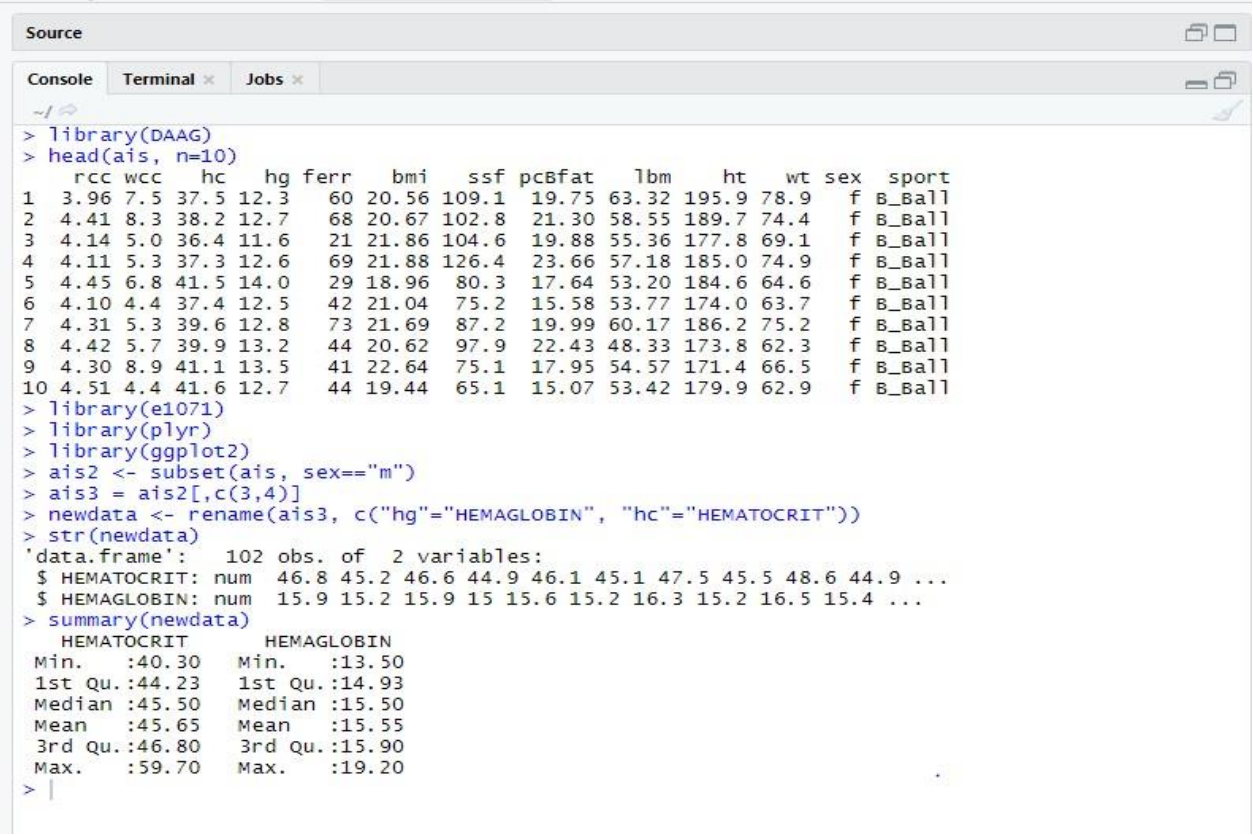
1812054	KARAN SHAH	Sports (athletic)

Q1.

Data Set Link (Standard web portal):

https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/csv/DAAG/ais.csv

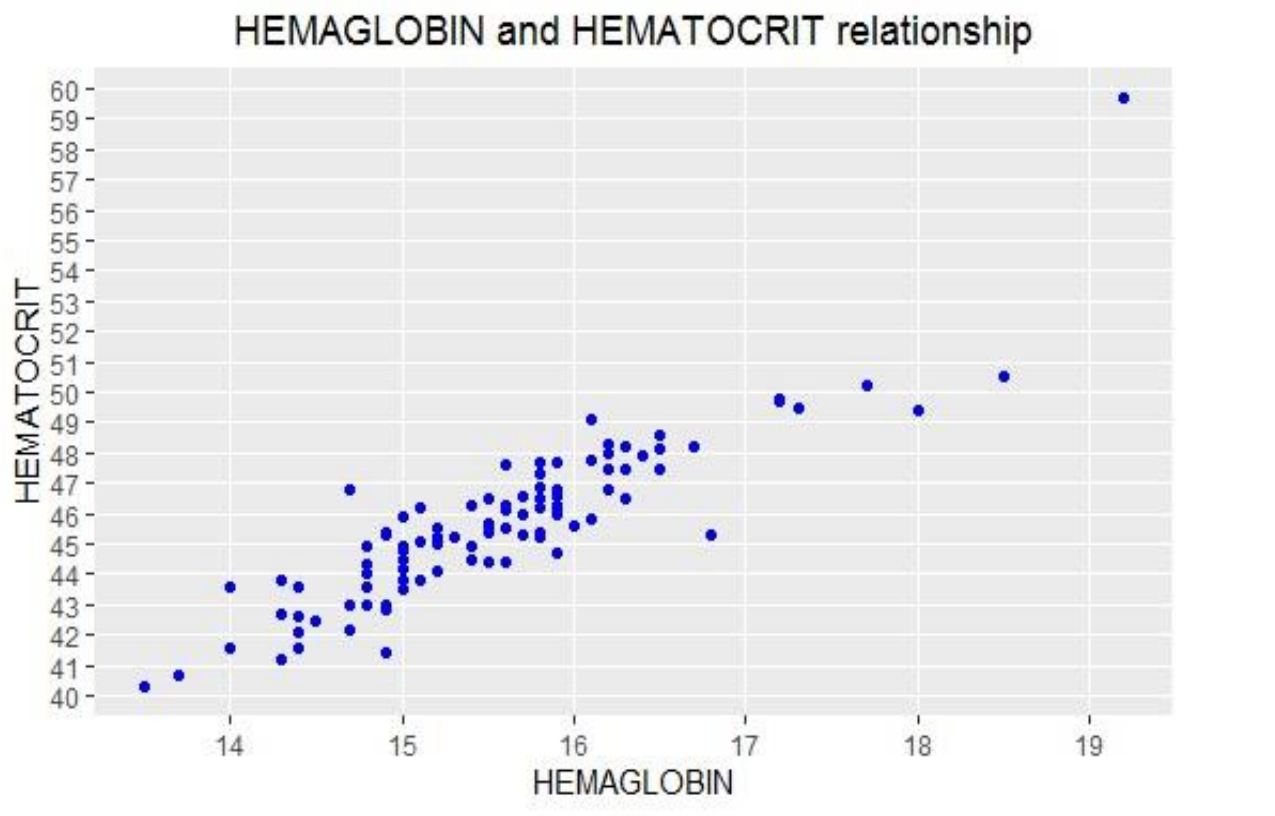




Graphical analysis:

AOCC EXAM.R*

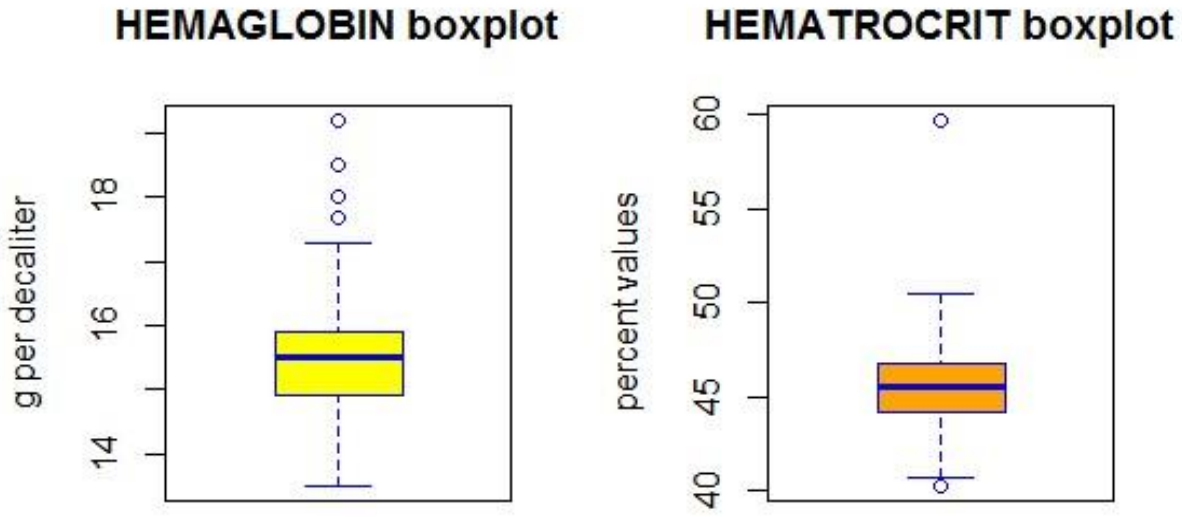
```
1  library(DAAG)
2  head(ais, n=10)
3  library(e1071)
4  library(plyr)
5  library(ggplot2)
6  ais2 <- subset(ais, sex=="m") # only male athletes
7  ais3 = ais2[,c(3,4)] # subset column number that correspond to "hg" and "hc"
8  newdata <- rename(ais3, c("hg"="HEMAGLOBIN", "hc"="HEMATOCRIT")) # rename variables
9  str(newdata)
10 colSums(is.na(newdata))
11 summary(newdata)
12 qplot(HEMAGLOBIN, HEMATOCRIT, data = newdata,
13       main = "HEMAGLOBIN and HEMATOCRIT relationship") +
14   theme(plot.title = element_text(hjust = 0.5)) +
15   geom_point(colour = "blue", size = 1.5) +
16   scale_y_continuous(breaks = c(30:65), minor_breaks = NULL) +
17   scale_x_continuous(breaks = c(10:25), minor_breaks = NULL)
18
```



Removable of Outlier:

AOCC EXAM.R*

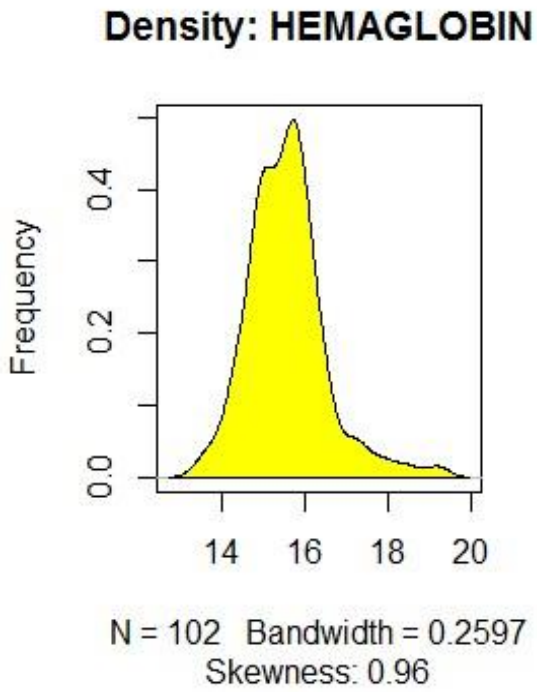
```
library(DAAG)
head(ais, n=10)
library(e1071)
library(plyr)
library(ggplot2)
ais2 <- subset(ais, sex=="m") # only male athletes
ais3 = ais2[,c(3,4)] # subset column number that correspond to "hg" and "hc"
newdata <- rename(ais3, c("hg"="HEMAGLOBIN", "hc"="HEMATOCRIT")) # rename variables
str(newdata)
colSums(is.na(newdata))
summary(newdata)
qplot(HEMAGLOBIN, HEMATOCRIT, data = newdata,
      main = "HEMAGLOBIN and HEMATOCRIT relationship") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_point(colour = "blue", size = 1.5) +
  scale_y_continuous(breaks = c(30:65), minor_breaks = NULL) +
  scale_x_continuous(breaks = c(10:25), minor_breaks = NULL)
par(mfrow=c(1, 2))  # it divides graph area in two parts
boxplot(newdata$HEMAGLOBIN, col = "yellow", border="blue",
        main = "HEMAGLOBIN boxplot",
        ylab = "g per decaliter")

boxplot(newdata$HEMATOCRIT, col = "orange", border="blue",
        main = "HEMATROCRIT boxplot",
        ylab = "percent values")
```

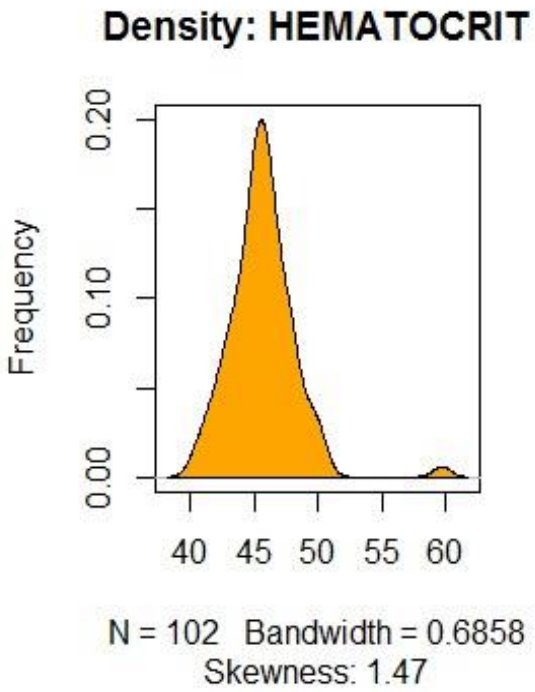


AOCC EXAM.R*

Source on Save | Run | Source

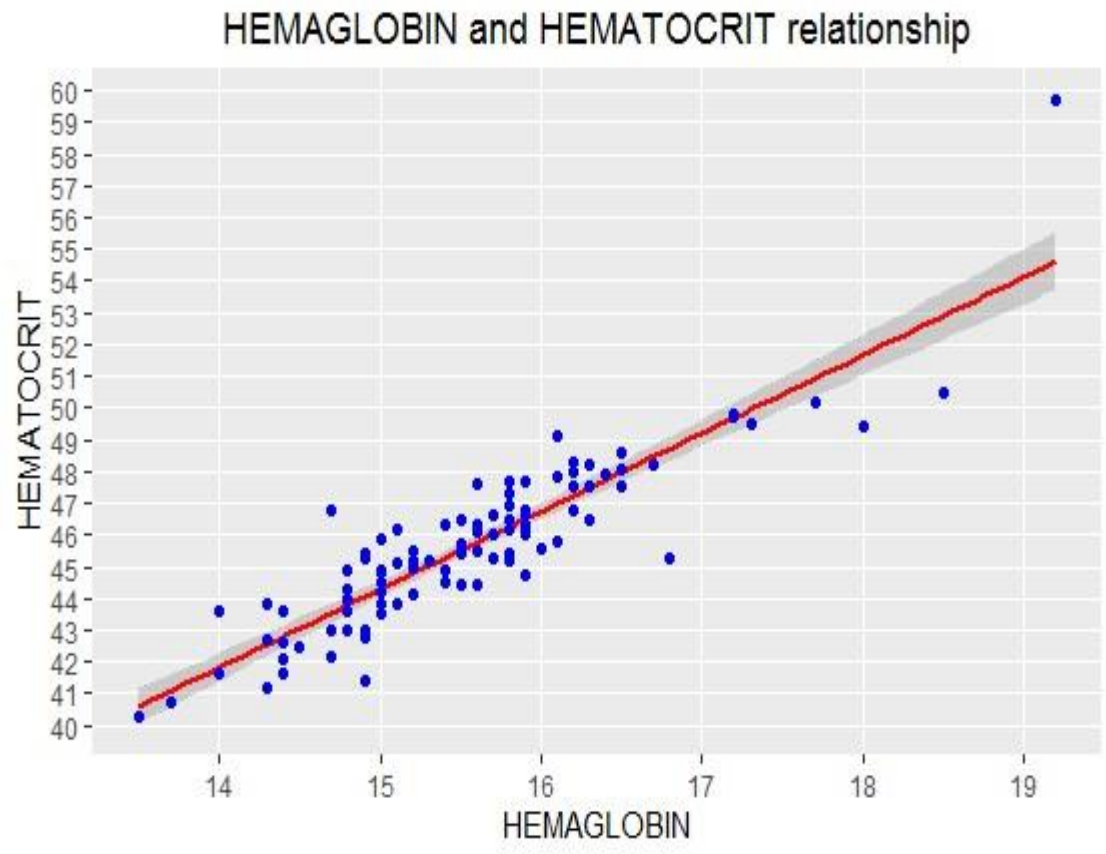
```
1  boxplot(newdata$HEMAGLOBIN, col = "yellow", border="blue",
2          main = "HEMAGLOBIN boxplot",
3          ylab = "g per decaliter")
4  boxplot(newdata$HEMATOCRIT, col = "orange", border="blue",
5          main = "HEMATROCRIT boxplot",
6          ylab = "percent values")
7  par(mfrow=c(1, 2))  # it divides graph area in two parts
8
9  plot(density(newdata$HEMAGLOBIN), main="Density: HEMAGLOBIN", ylab="Frequency",
10      sub=paste("Skewness:", round(e1071::skewness(newdata$HEMAGLOBIN), 2)))
11 polygon(density(newdata$HEMAGLOBIN), col="yellow")
12
13 plot(density(newdata$HEMATOCRIT), main="Density: HEMATOCRIT", ylab="Frequency",
14      sub=paste("Skewness:", round(e1071::skewness(newdata$HEMATOCRIT), 2)))
15 polygon(density(newdata$HEMATOCRIT), col="orange")
16
17
```





Graphical analysis after pre-processing:

```
plot(density(newdata$HEMAGLOBIN), main="Density: HEMAGLOBIN", ylab="Frequency",
     sub=paste("Skewness:", round(e1071::skewness(newdata$HEMAGLOBIN), 2)))
polygon(density(newdata$HEMAGLOBIN), col="yellow")

plot(density(newdata$HEMATOCRIT), main="Density: HEMATOCRIT", ylab="Frequency",
     sub=paste("Skewness:", round(e1071::skewness(newdata$HEMATOCRIT), 2)))
polygon(density(newdata$HEMATOCRIT), col="orange")
qplot(HEMAGLOBIN, HEMATOCRIT, data = newdata,
      main = "HEMAGLOBIN and HEMATOCRIT relationship") +
  theme(plot.title = element_text(hjust = 0.5)) +
  stat_smooth(method="lm", col="red", size=1) +
  geom_point(colour = "blue", size = 1.5) +
  scale_y_continuous(breaks = c(30:65), minor_breaks = NULL) +
  scale_x_continuous(breaks = c(10:25), minor_breaks = NULL)
```



K. J. Somaiya College of Engineering, Mumbai – 77

(Autonomous College Affiliated to University of Mumbai)

Regression Analysis :

AOCC EXAM.R*

Source on Save | Run | Source

```
qplot(HEMAGLOBIN, HEMATOCRIT, data = newdata,
      main = "HEMAGLOBIN and HEMATOCRIT relationship") +
  theme(plot.title = element_text(hjust = 0.5)) +
  stat_smooth(method="lm", col="red", size=1) +
  geom_point(colour = "blue", size = 1.5) +
  scale_y_continuous(breaks = c(30:65), minor_breaks = NULL) +
  scale_x_continuous(breaks = c(10:25), minor_breaks = NULL)
set.seed(123) # setting seed to reproduce results of random sampling
HEMAGLOBIN_CENT = scale(newdata$HEMAGLOBIN, center=TRUE, scale=FALSE) # center the variable
# Show the relationship with new variable centered, creating a regression line
qplot(HEMAGLOBIN_CENT, HEMATOCRIT, data = newdata,
      main = "HEMAGLOBIN_CENT and HEMATOCRIT relationship") +
  theme(plot.title = element_text(hjust = 0.5)) +
  stat_smooth(method="lm", col="red", size=1) +
  geom_point(colour = "blue", size = 1.5) +
  scale_y_continuous(breaks = c(30:65), minor_breaks = NULL) +
  scale_x_continuous(breaks = c(-2,-1.5,-1,-0.5,0,0.5,1,1.5,2,2.5,3,3.5,4), minor_breaks = NULL)
mod1 = lm(HEMATOCRIT ~ HEMAGLOBIN_CENT, data = newdata)
summary(mod1)
```

```
Call:
lm(formula = HEMATOCRIT ~ HEMAGLOBIN_CENT, data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4183 -0.7043 -0.0072  0.6049  5.0765

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     45.6500     0.1140  400.35   <2e-16 ***
HEMAGLOBIN_CENT  2.4605     0.1227   20.06   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.152 on 100 degrees of freedom
Multiple R-squared:  0.801,     Adjusted R-squared:  0.799
F-statistic: 402.4 on 1 and 100 DF,  p-value: < 2.2e-16
```

HEMATOCRIT

K. J. Somaiya College of Engineering, Mumbai – 77

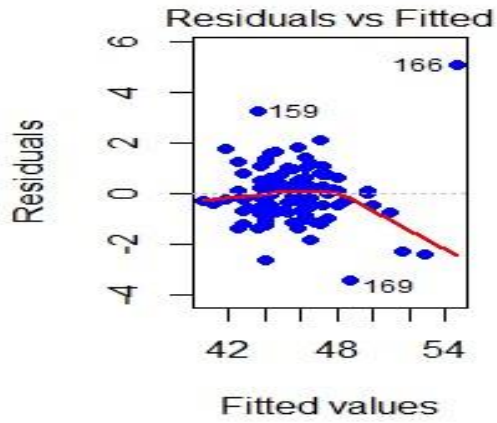
(Autonomous College Affiliated to University of Mumbai)

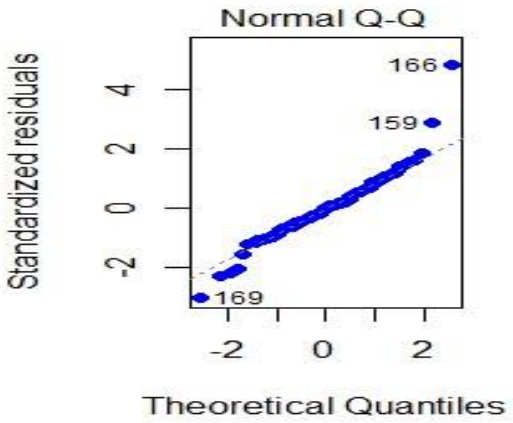
Criteria to prove analysis is correct

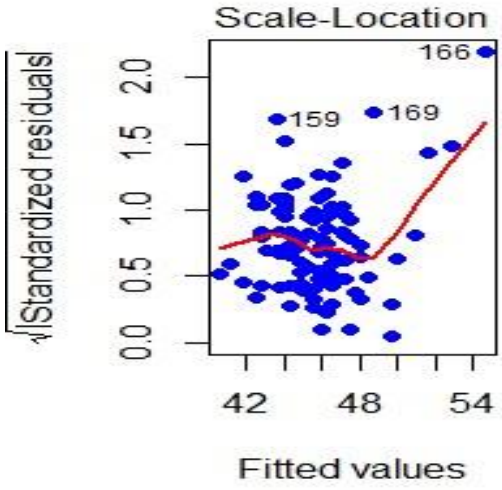
AOCC EXAM.R*

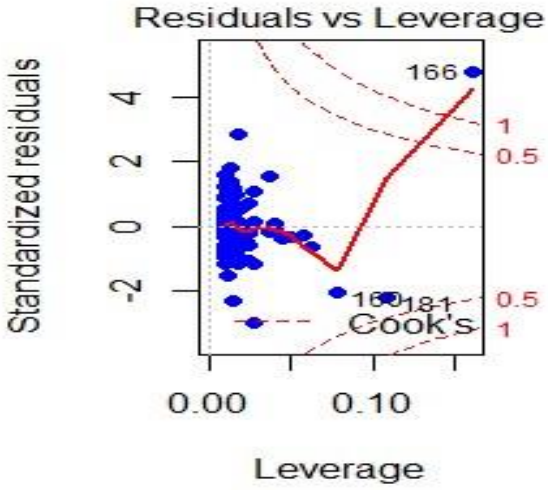
Source on Save | Run | Source

```
1  plot(mod1, pch=16, col="blue", lty=1, lwd=2, which=1)
2  plot(mod1, pch=16, col="blue", lty=1, lwd=2, which=2)
3  plot(mod1, pch=16, col="blue", lty=1, lwd=2, which=3)
4  plot(mod1, pch=16, col="blue", lty=1, lwd=2, which=5)
5  plot(mod1, pch=16, col="blue", lty=1, lwd=2, which=4)
6  par(mfrow = c(2,2)) # display a unique layout for all graphs
7  plot(mod2)
8
```

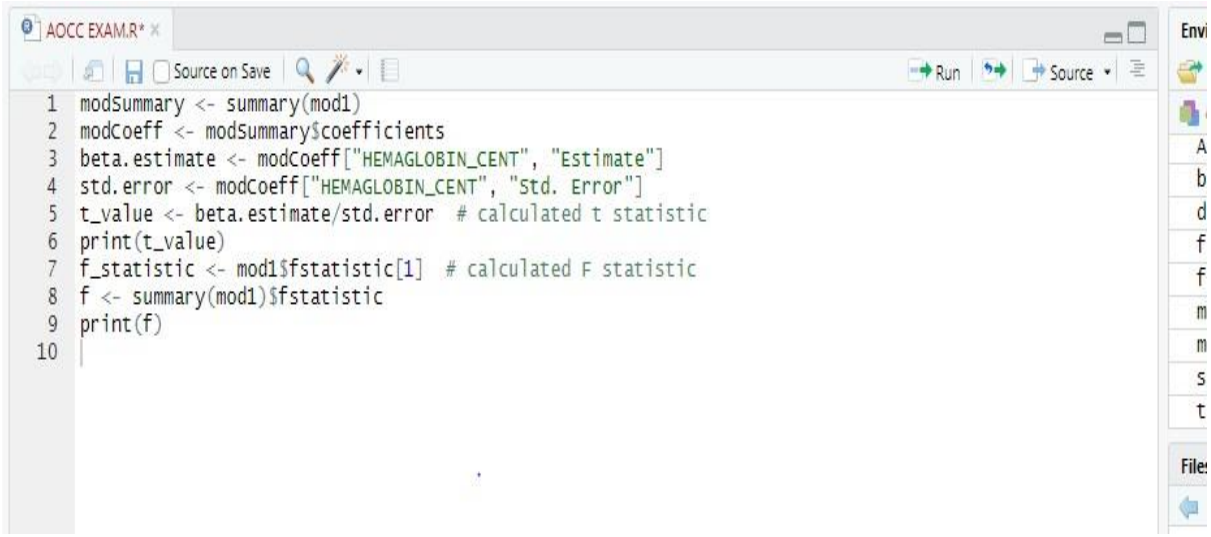


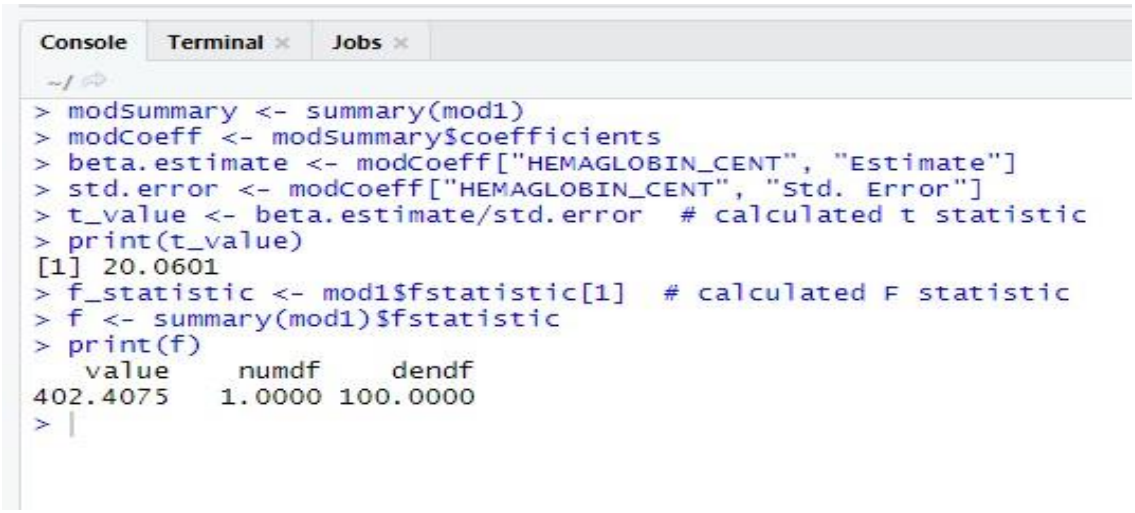




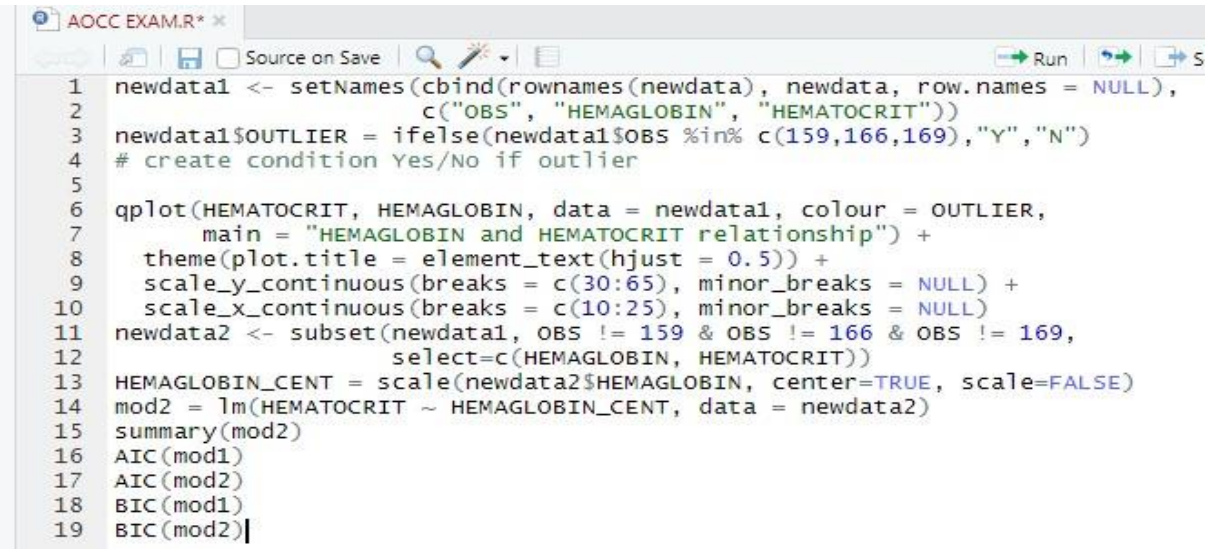


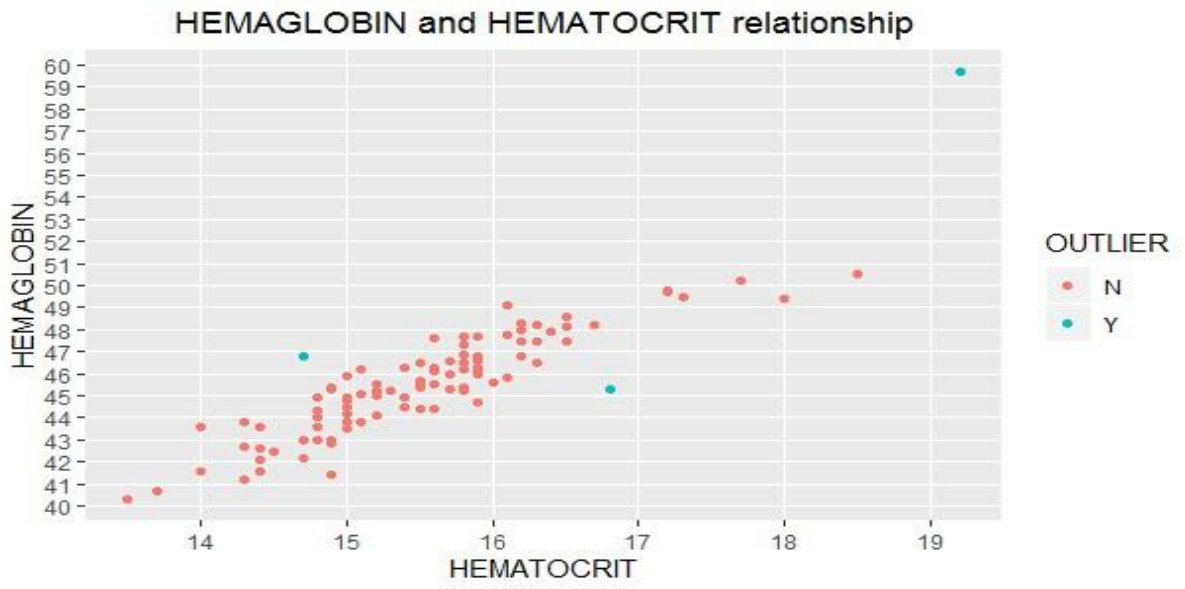
Calculation of t statistic and p-statistic value:





Model improvement:





Blue points represent the three outliers indentified. So a new dataset can be created excluding them:

```
Call:
lm(formula = HEMATOCRIT ~ HEMAGLOBIN_CENT, data = newdata2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83224 -0.20845 -0.00573  0.21535  1.19873

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    15.51212    0.03656  424.26   <2e-16 ***
HEMAGLOBIN_CENT  0.35783    0.01687   21.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3638 on 97 degrees of freedom
Multiple R-squared:  0.8227,    Adjusted R-squared:  0.8209
F-statistic: 450.1 on 1 and 97 DF,  p-value: < 2.2e-16

> AIC(mod1)
[1] 322.2403
> AIC(mod2)
[1] 84.71671
> BIC(mod1)
[1] 330.1153
> BIC(mod2)
[1] 92.50206
```

The model with the lower AIC and BIC is preferred. The model 2 is more accurate as we can see above.

Prediction through model:

AOCC EXAM.R*

Source on Save | Run | Source

```
1  set.seed(123)
2  trainingRowIndex <- sample(1:nrow(newdata2), 0.7*nrow(newdata2))
3  #  training and testing: 70/30 split
4  trainingData <- newdata2[trainingRowIndex, ]
5  testData  <- newdata2[-trainingRowIndex, ]
6  modTrain <- lm(HEMATOCRIT ~ HEMAGLOBIN, data=trainingData)  # I have build the model
7  predict <- predict(modTrain, testData)  # predicted values
8  summary(modTrain)
9  act_pred <- data.frame(cbind(actuals=testData$HEMATOCRIT, predicteds=predict))
10 # actuals_predicteds
11 cor(act_pred) # correlation_accuracy
12 head(act_pred, n=10)
```

Console | Terminal | Jobs

```
Call:
lm(formula = HEMATOCRIT ~ HEMAGLOBIN, data = trainingData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.85570 -0.20286  0.03222  0.20443  1.22379

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.43907    0.92072  -0.477    0.635
HEMAGLOBIN   0.35080    0.02024  17.335   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3711 on 67 degrees of freedom
Multiple R-squared:  0.8177,    Adjusted R-squared:  0.815
F-statistic: 300.5 on 1 and 67 DF,  p-value: < 2.2e-16

> act_pred <- data.frame(cbind(actuals=testData$HEMATOCRIT, predicteds=predict))
> # actuals_predicteds
> cor(act_pred) # correlation_accuracy
            actuals predicteds
actuals    1.0000000  0.9156139
predicteds 0.9156139  1.0000000
> head(act_pred, n=10)
   actuals predicteds
1     15.9   15.97825
2     15.2   15.41698
3     15.9   15.90810
10    15.4   15.31174
11    16.1   16.32905
19    15.4   15.17142
20    16.2   16.22381
24    15.5   15.87302
28    15.6   15.52222
35    13.7   13.83839
>
```

K. J. Somaiya College of Engineering, Mumbai – 77

(Autonomous College Affiliated to University of Mumbai)

MIN & MAX Accuracy Error Rates:

```
AOCC EXAM.R*
Source on Save | Run | Source
1  min_max <- mean(apply(act_pred, 1, min) / apply(act_pred, 1, max))
2  print(min_max) # show the result
3  mape <- mean(abs((act_pred$predicteds - act_pred$actuals))/act_pred$actuals)
4  print(mape) # show the result
```

```
4:30 (Top Level)                                   R Script
Console  Terminal  Jobs
~/
> min_max <- mean(apply(act_pred, 1, min) / apply(act_pred, 1, max))
> print(min_max) # show the result
[1] 0.9828594
> mape <- mean(abs((act_pred$predicteds - act_pred$actuals))/act_pred$actuals)
> print(mape) # show the result
[1] 0.01736733
>
```

K -FOLD CROSS Validation:

```
AOCC EXAM.R*
Source on Save | Run | Source
1  min_max <- mean(apply(act_pred, 1, min) / apply(act_pred, 1, max))
2  print(min_max) # show the result
3  mape <- mean(abs((act_pred$predicteds - act_pred$actuals))/act_pred$actuals)
4  print(mape) # show the result
5  kfold <- CVlm(data = newdata2, form.lm = formula(HEMATOCRIT ~ HEMAGLOBIN), m=5,
6                dots = FALSE, seed=123, legend.pos="topleft",
7                main="Cross Validation; k=5",
8                plotit=TRUE, printit=FALSE)
9
```