

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>

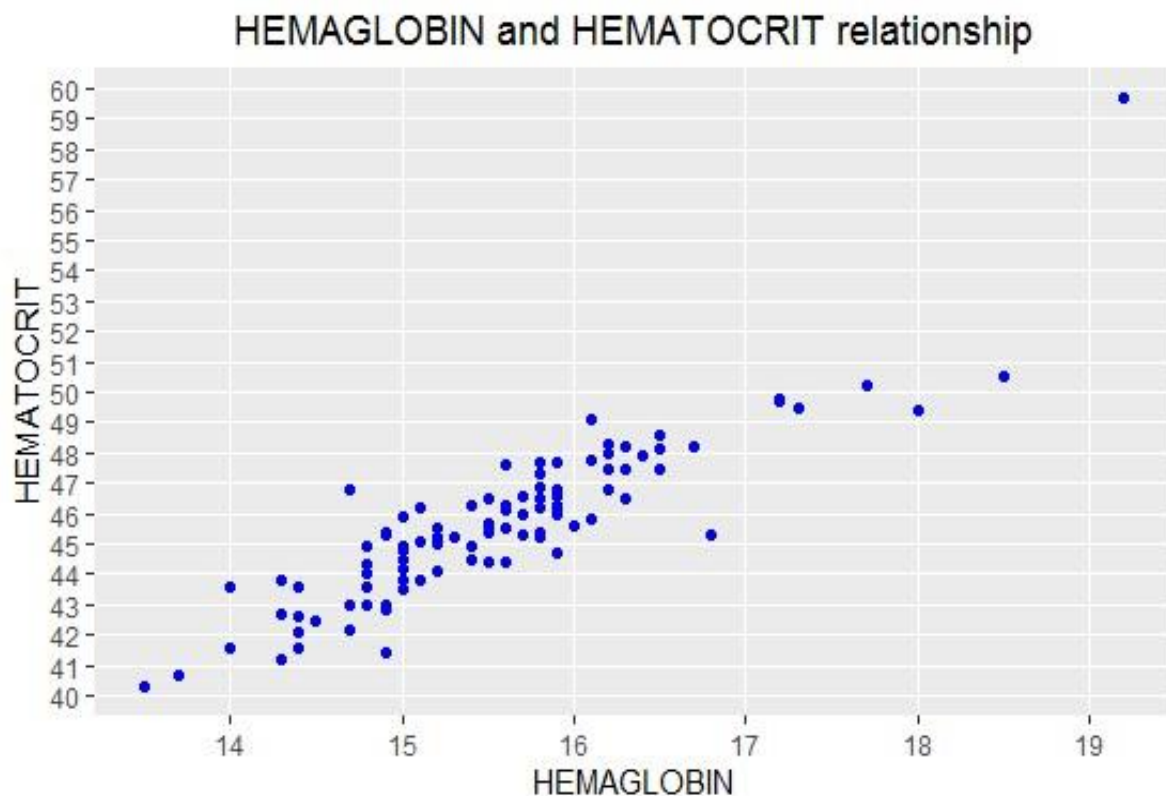
```
AOCC EXAM.R x
Source on Save
1 library(DAAG)
2 head(ais, n=10)
3 library(e1071)
4 library(plyr)
5 library(ggplot2)
6 ais2 <- subset(ais, sex=="m")
7 ais3 = ais2[,c(3,4)]
8 newdata <- rename(ais3, c("hg"="HEMAGLOBIN", "hc"="HEMATOCRIT"))
9 str(newdata)
10 summary(newdata)
```

```
Source
Console Terminal x Jobs x
~/
> library(DAAG)
> head(ais, n=10)
  rcc wcc hc hg ferr bmi ssf pcBfat lbm ht wt sex sport
1 3.96 7.5 37.5 12.3 60 20.56 109.1 19.75 63.32 195.9 78.9 f B_Ball
2 4.41 8.3 38.2 12.7 68 20.67 102.8 21.30 58.55 189.7 74.4 f B_Ball
3 4.14 5.0 36.4 11.6 21 21.86 104.6 19.88 55.36 177.8 69.1 f B_Ball
4 4.11 5.3 37.3 12.6 69 21.88 126.4 23.66 57.18 185.0 74.9 f B_Ball
5 4.45 6.8 41.5 14.0 29 18.96 80.3 17.64 53.20 184.6 64.6 f B_Ball
6 4.10 4.4 37.4 12.5 42 21.04 75.2 15.58 53.77 174.0 63.7 f B_Ball
7 4.31 5.3 39.6 12.8 73 21.69 87.2 19.99 60.17 186.2 75.2 f B_Ball
8 4.42 5.7 39.9 13.2 44 20.62 97.9 22.43 48.33 173.8 62.3 f B_Ball
9 4.30 8.9 41.1 13.5 41 22.64 75.1 17.95 54.57 171.4 66.5 f B_Ball
10 4.51 4.4 41.6 12.7 44 19.44 65.1 15.07 53.42 179.9 62.9 f B_Ball
> library(e1071)
> library(plyr)
> library(ggplot2)
> ais2 <- subset(ais, sex=="m")
> ais3 = ais2[,c(3,4)]
> newdata <- rename(ais3, c("hg"="HEMAGLOBIN", "hc"="HEMATOCRIT"))
> str(newdata)
'data.frame': 102 obs. of 2 variables:
 $ HEMATOCRIT: num 46.8 45.2 46.6 44.9 46.1 45.1 47.5 45.5 48.6 44.9 ...
 $ HEMAGLOBIN: num 15.9 15.2 15.9 15 15.6 15.2 16.3 15.2 16.5 15.4 ...
> summary(newdata)
  HEMATOCRIT  HEMAGLOBIN 
Min.   :40.30  Min.   :13.50 
1st Qu.:44.23  1st Qu.:14.93 
Median :45.50  Median :15.50 
Mean   :45.65  Mean   :15.55 
3rd Qu.:46.80  3rd Qu.:15.90 
Max.   :59.70  Max.   :19.20 
>
```

K. J. Somaiya College of Engineering, Mumbai – 77
(Autonomous College Affiliated to University of Mumbai)

Graphical analysis:

```
AOCC EXAM.R* x
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
```

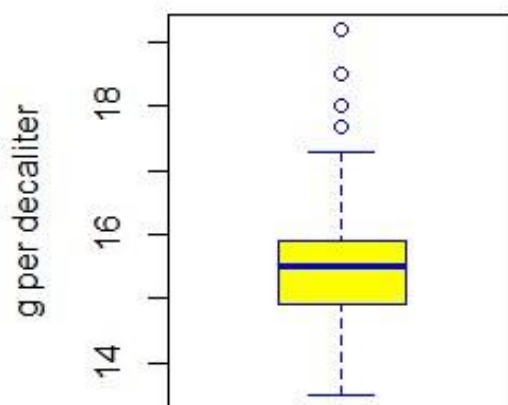


K. J. Somaiya College of Engineering, Mumbai – 77
(Autonomous College Affiliated to University of Mumbai)

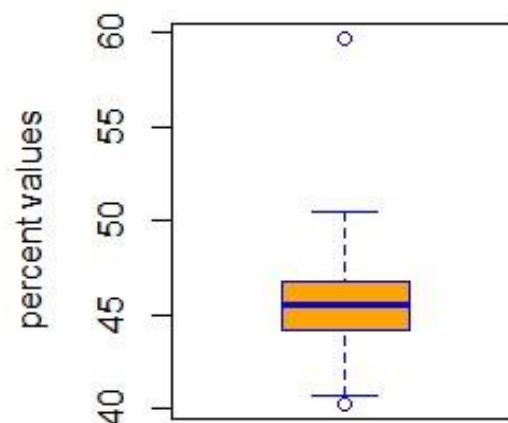
Removable of Outlier:

```
AOCC EXAM.R* x
Source on Save Run Source
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 par(mfrow=c(1, 2)) # it divides graph area in two parts
19 boxplot(newdata$HEMAGLOBIN, col = "yellow", border="blue",
20         main = "HEMAGLOBIN boxplot",
21         ylab = "g per decaliter")
22
23 boxplot(newdata$HEMATOCRIT, col = "orange", border="blue",
24         main = "HEMATOCRIT boxplot",
25         ylab = "percent values")
26
```

HEMAGLOBIN boxplot



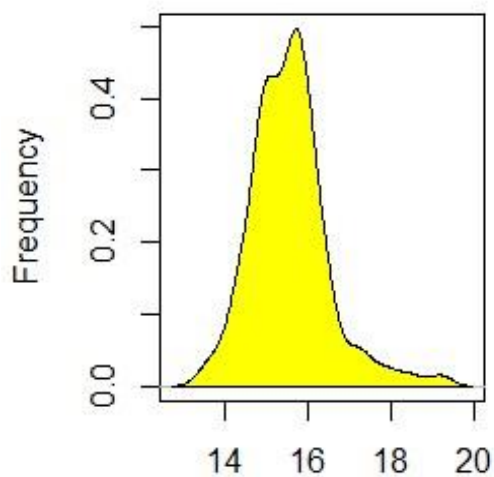
HEMATOCRIT boxplot



K. J. Somaiya College of Engineering, Mumbai – 77
(Autonomous College Affiliated to University of Mumbai)

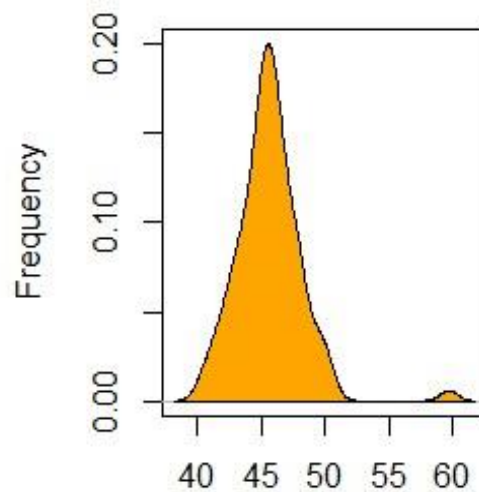
```
AOCC EXAM.R* x
Source on Save
Run Source
1 boxplot(newdata$HEMAGLOBIN, col = "yellow", border="blue",
2         main = "HEMAGLOBIN boxplot",
3         ylab = "g per decaliter")
4 boxplot(newdata$HEMATOCRIT, col = "orange", border="blue",
5         main = "HEMATOCRIT boxplot",
6         ylab = "percent 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 |
```

Density: HEMAGLOBIN



N = 102 Bandwidth = 0.2597
Skewness: 0.96

Density: HEMATOCRIT

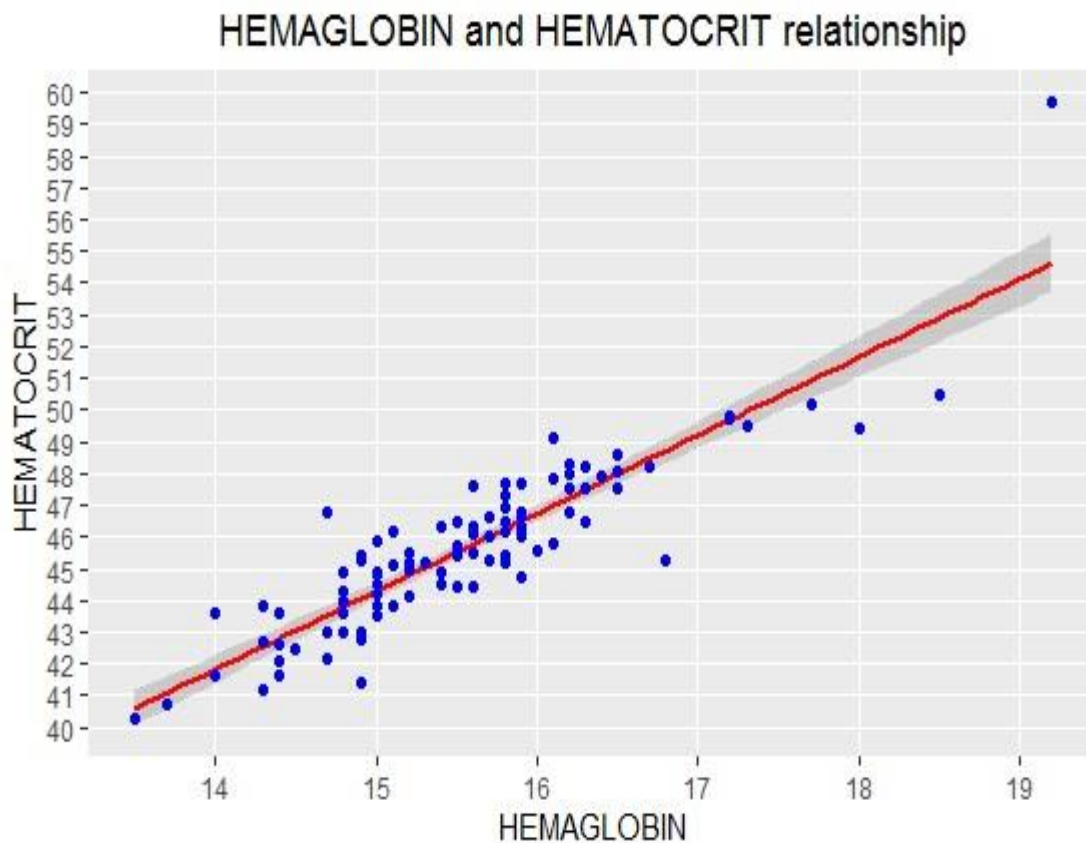


N = 102 Bandwidth = 0.6858
Skewness: 1.47

K. J. Somaiya College of Engineering, Mumbai – 77
(Autonomous College Affiliated to University of Mumbai)

Graphical analysis after pre-processing:

```
AOCC EXAM.R* x
1 plot(density(newdata$HEMAGLOBIN), main="Density: HEMAGLOBIN", ylab="Frequency",
2       sub=paste("Skewness:", round(e1071::skewness(newdata$HEMAGLOBIN), 2)))
3 polygon(density(newdata$HEMAGLOBIN), col="yellow")
4
5 plot(density(newdata$HEMATOCRIT), main="Density: HEMATOCRIT", ylab="Frequency",
6       sub=paste("Skewness:", round(e1071::skewness(newdata$HEMATOCRIT), 2)))
7 polygon(density(newdata$HEMATOCRIT), col="orange")
8 qplot(HEMAGLOBIN, HEMATOCRIT, data = newdata,
9       main = "HEMAGLOBIN and HEMATOCRIT relationship") +
10 theme(plot.title = element_text(hjust = 0.5)) +
11 stat_smooth(method="lm", col="red", size=1) +
12 geom_point(colour = "blue", size = 1.5) +
13 scale_y_continuous(breaks = c(30:65), minor_breaks = NULL) +
14 scale_x_continuous(breaks = c(10:25), minor_breaks = NULL)
15
```



K. J. Somaiya College of Engineering, Mumbai – 77
(Autonomous College Affiliated to University of Mumbai)

Regression Analysis :

```

1  qplot(HEMAGLOBIN, HEMATOCRIT, data = newdata,
2      main = "HEMAGLOBIN and HEMATOCRIT relationship") +
3      theme(plot.title = element_text(hjust = 0.5)) +
4      stat_smooth(method="lm", col="red", size=1) +
5      geom_point(colour = "blue", size = 1.5) +
6      scale_y_continuous(breaks = c(30:65), minor_breaks = NULL) +
7      scale_x_continuous(breaks = c(10:25), minor_breaks = NULL)
8  set.seed(123) # setting seed to reproduce results of random sampling
9  HEMAGLOBIN_CENT = scale(newdata$HEMAGLOBIN, center=TRUE, scale=FALSE) # center the variable
10 # Show the relationship with new variable centered, creating a regression line
11 qplot(HEMAGLOBIN_CENT, HEMATOCRIT, data = newdata,
12     main = "HEMAGLOBIN_CENT and HEMATOCRIT relationship") +
13     theme(plot.title = element_text(hjust = 0.5)) +
14     stat_smooth(method="lm", col="red", size=1) +
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(-2,-1.5,-1,-0.5,0,0.5,1,1.5,2,2.5,3,3.5,4), minor_breaks = NULL)
18 mod1 = lm(HEMATOCRIT ~ HEMAGLOBIN_CENT, data = newdata)
19 summary(mod1)
20

```

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

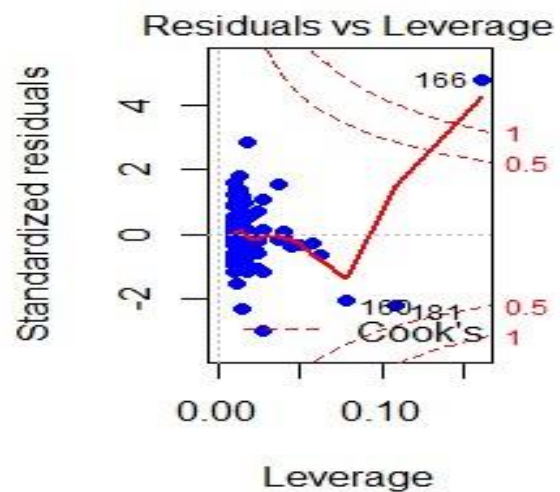
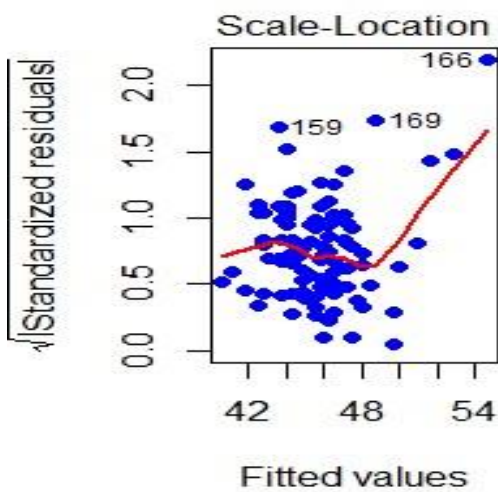
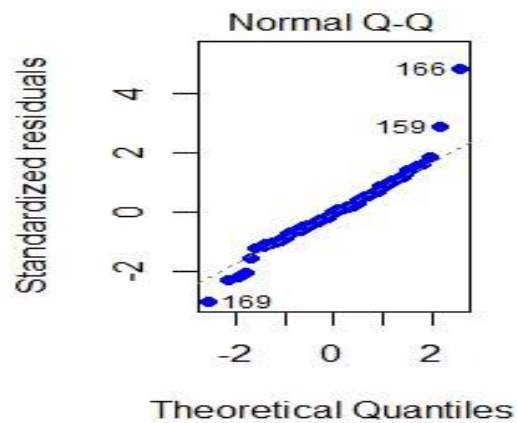
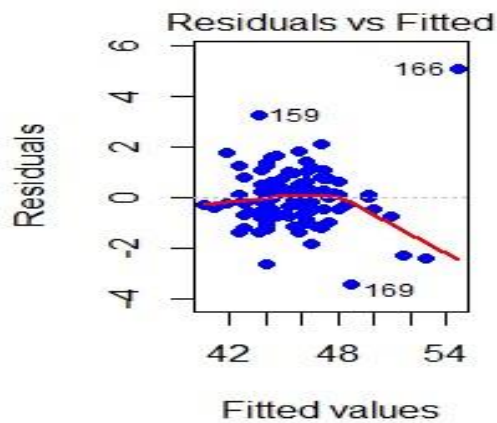
HEMATOCRIT

Criteria to prove analysis is correct

```

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

```



K. J. Somaiya College of Engineering, Mumbai – 77
(Autonomous College Affiliated to University of Mumbai)

Calculation of t statistic and p-statistic value:

```
AOCC EXAM.R* x
Source on Save
Run Source

1 modSummary <- summary(mod1)
2 modCoeff <- modSummary$coefficients
3 beta.estimate <- modCoeff["HEMAGLOBIN_CENT", "Estimate"]
4 std.error <- modCoeff["HEMAGLOBIN_CENT", "Std. Error"]
5 t_value <- beta.estimate/std.error # calculated t statistic
6 print(t_value)
7 f_statistic <- mod1$fstatistic[1] # calculated F statistic
8 f <- summary(mod1)$fstatistic
9 print(f)
10
```

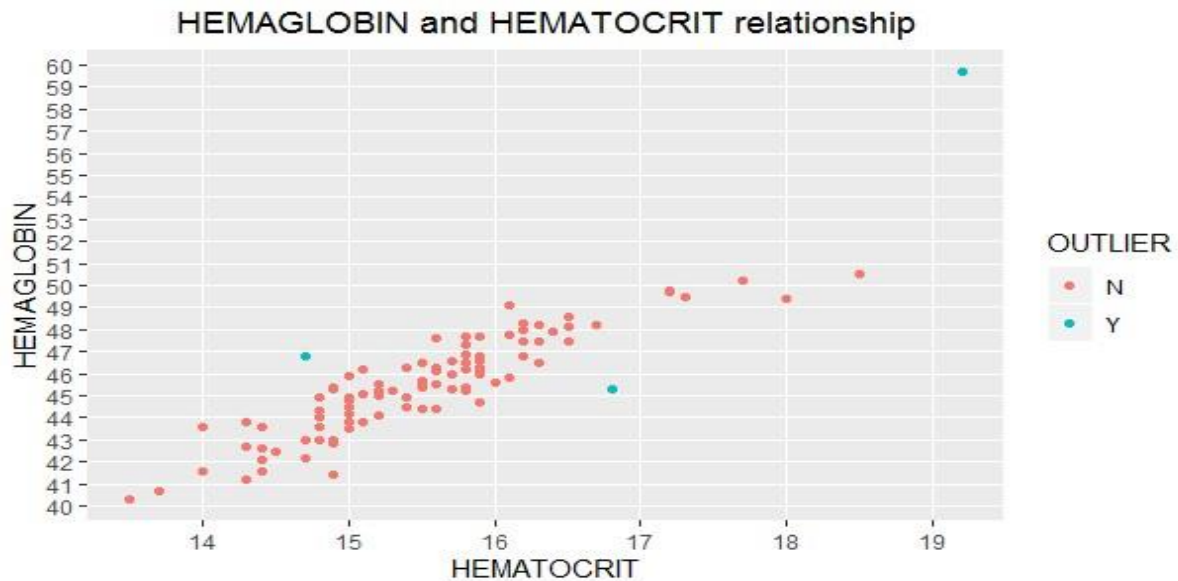
```
Console Terminal x Jobs x
~/
> modSummary <- summary(mod1)
> modCoeff <- modSummary$coefficients
> beta.estimate <- modCoeff["HEMAGLOBIN_CENT", "Estimate"]
> std.error <- modCoeff["HEMAGLOBIN_CENT", "Std. Error"]
> t_value <- beta.estimate/std.error # calculated t statistic
> print(t_value)
[1] 20.0601
> f_statistic <- mod1$fstatistic[1] # calculated F statistic
> f <- summary(mod1)$fstatistic
> print(f)
      value      numdf      dendf
402.4075      1.0000 100.0000
> |
```

Model improvement:

```
AOCC EXAM.R* x
Source on Save
Run Source

1 newdata1 <- setNames(cbind(rownames(newdata), newdata, row.names = NULL),
2                       c("OBS", "HEMAGLOBIN", "HEMATOCRIT"))
3 newdata1$OUTLIER = ifelse(newdata1$OBS %in% c(159,166,169), "Y", "N")
4 # create condition Yes/No if outlier
5
6 qqplot(HEMATOCRIT, HEMAGLOBIN, data = newdata1, colour = OUTLIER,
7        main = "HEMAGLOBIN and HEMATOCRIT relationship") +
8  theme(plot.title = element_text(hjust = 0.5)) +
9  scale_y_continuous(breaks = c(30:65), minor_breaks = NULL) +
10 scale_x_continuous(breaks = c(10:25), minor_breaks = NULL)
11 newdata2 <- subset(newdata1, OBS != 159 & OBS != 166 & OBS != 169,
12                   select=c(HEMAGLOBIN, HEMATOCRIT))
13 HEMAGLOBIN_CENT = scale(newdata2$HEMAGLOBIN, center=TRUE, scale=FALSE)
14 mod2 = lm(HEMATOCRIT ~ HEMAGLOBIN_CENT, data = newdata2)
15 summary(mod2)
16 AIC(mod1)
17 AIC(mod2)
18 BIC(mod1)
19 BIC(mod2)|
```


K. J. Somaiya College of Engineering, Mumbai – 77
(Autonomous College Affiliated to University of Mumbai)



Blue points represent the three outliers identified. 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.

K. J. Somaiya College of Engineering, Mumbai – 77
(Autonomous College Affiliated to University of Mumbai)

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, predicted=predict))
10 # actuals_predicted
11 cor(act_pred) # correlation_accuracy
12 head(act_pred, n=10)
```

```
Console Terminal x Jobs x
~/
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, predicted=predict))
> # actuals_predicted
> cor(act_pred) # correlation_accuracy
              actuals predicted
actuals  1.0000000  0.9156139
predicted 0.9156139  1.0000000
> head(act_pred, n=10)
      actuals predicted
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* x
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$predicted - act_pred$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$predicted - act_pred$actuals))/act_pred$actuals)
> print(mape) # show the result
[1] 0.01736733
> |
```

K -FOLD CROSS Validation:

```
AOCC EXAM.R* x
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$predicted - act_pred$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 |
```

