Pranav Pillai

IB 500 SMLE HW 2

Q1a) $\hat{\sigma} = 31.5$ on 25 df

b) $\hat{y}_A - \hat{y}_B = \hat{\beta}_{plastic}\, x_{plastic\,A} + \epsilon_A$
$- \hat{\beta}_{plastic}\, x_{plastic} - \epsilon_B$
$= 10\, \hat{\beta}_{plastic} + \epsilon_A - \epsilon_B$

$\hat{E}(\hat{y}_A - \hat{y}_B) = 10\hat{\beta}_{plastic} = 289.2$

$\hat{VAR}(\hat{y}_A - \hat{y}_B) = \hat{VAR}(10\,\hat{\beta}_{plastic}) + \hat{VAR}(\epsilon_A) + \hat{VAR}(\epsilon_B)$
$= 100\,\hat{Var}(\hat{\beta}_{plastic}) + 2s^2$
$= 100(2.82)^2 + 2(31.5)^2 = 2779.74$

95% prediction interval is
$(289.2 - 1.96\sqrt{2779.74},\ 289.2 + 1.96\sqrt{2779.74})$
$(185.8625,\ 392.5375)$

c) $x^* = (1,0,0,0,1,0,0)$
$b^* = 2245.09 + (100)(-37.36)$
$= -1490.91$

Prediction is not reliable because of exploitation. Unlikely to have sample with pure water in data set.

d) B_{paper} 95% CI
$7.64 \pm 1.96(2.31) \rightarrow (3.1124,\ 12.1676)$

e) Shows a positive correlation between predictors paper and garbage. As value for paper increases, so does garbage. If a line of best fit is done, gradient is positive implying positive correlation.

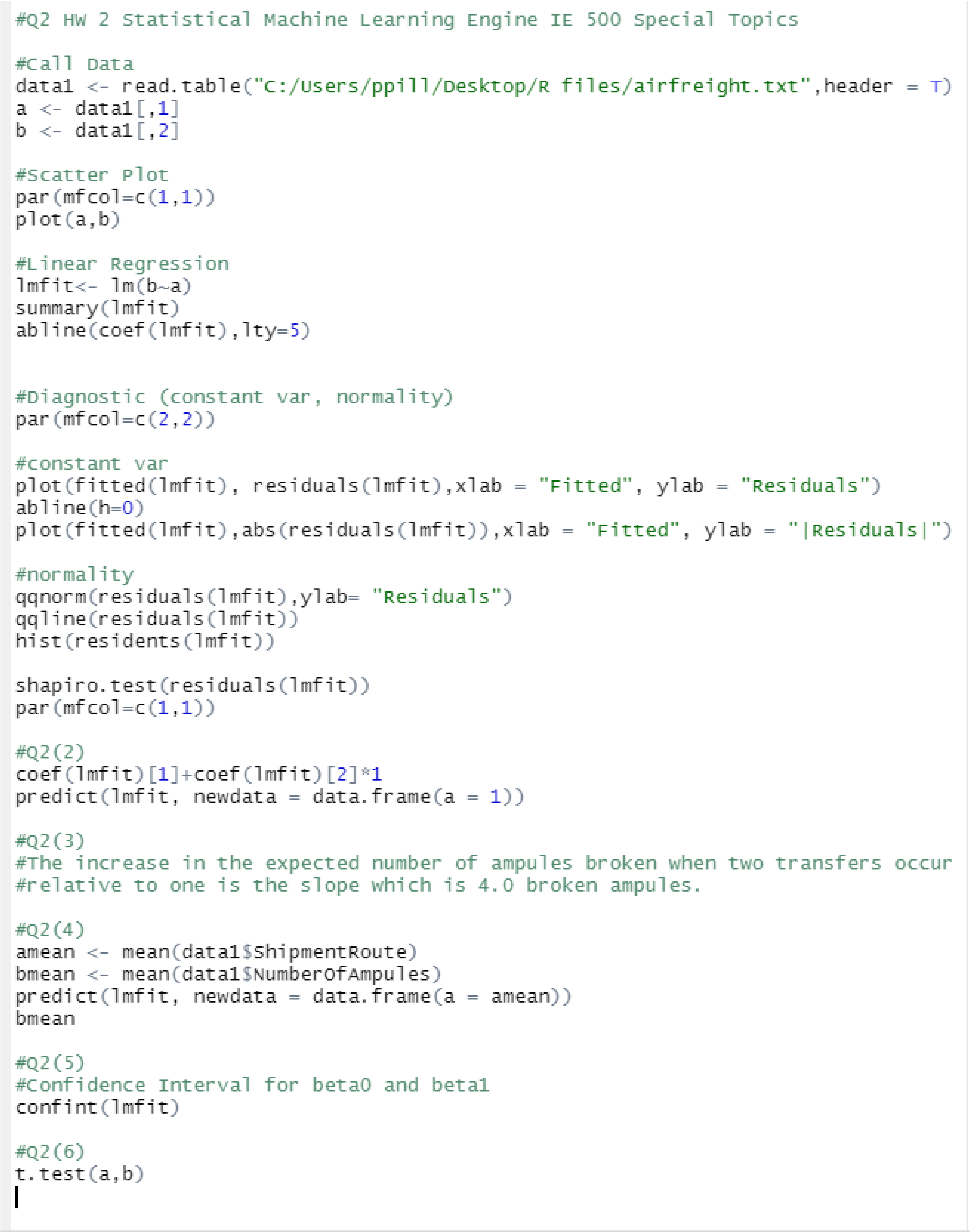
Q1 Part 6

There are 4 assumptions associated with the standard linear model which are:

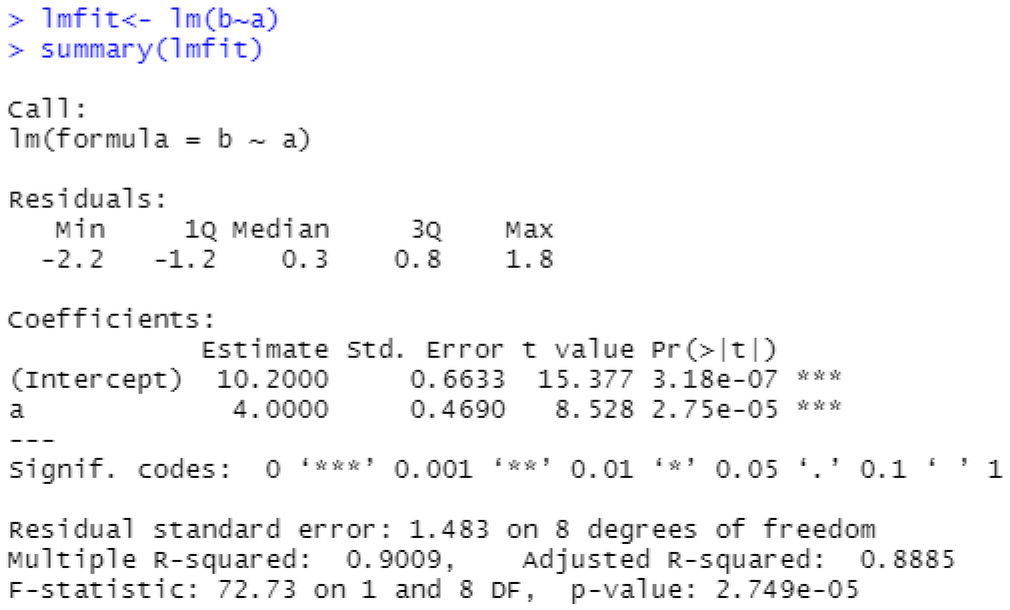
- Linearity and addivity of the relation between the explanatory and response variables
    - The expected value is a straight-line function of each explanatory variable while others are constant
    - Gradient of line doesn't depend on other variable values
    - Effects of the various explanatory variables on the expected values of the response variables are additive.
- Statistical independence of the errors
- Homoscedasticity of the errors (with respect to time, predictions, any explanatory variables)
- Normality of the error distribution

The plot does not violate any assumptions of the standard linear model. For the scatter plot generated for paper and garbage, the line of best fit generated for the points would have a line of a constant gradient – which means it fulfills the linearity of the model. For the fixed and residual plot, there is a fair amount of homoscedasticity as the points are fairly distributed if you look at the middle of the plot. The variables are independent for paper and garbage as they are within the 95% confidence interval as mentioned in the summary.

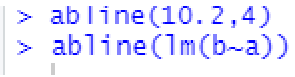
References: *Regression diagnostics: testing the assumptions of linear regression*. Testing the assumptions of linear regression. (n.d.). Retrieved October 20, 2021, from https://people.duke.edu/~rnau/testing.htm.



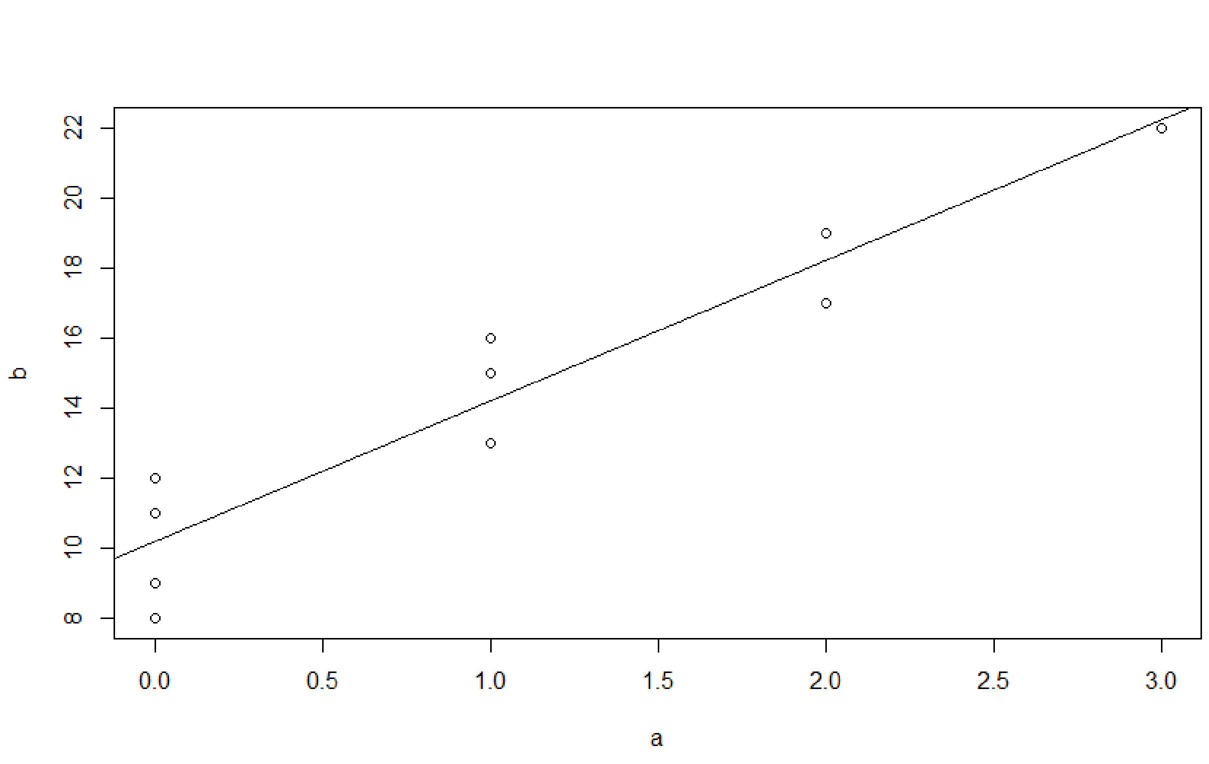
```
#Q2 HW 2 Statistical Machine Learning Engine IE 500 Special Topics

#Call Data
data1 <- read.table("C:/Users/ppill/Desktop/R files/airfreight.txt",header = T)
a <- data1[,1]
b <- data1[,2]

#Scatter Plot
par(mfcol=c(1,1))
plot(a,b)

#Linear Regression
lmfit<- lm(b~a)
summary(lmfit)
abline(coef(lmfit),lty=5)


#Diagnostic (constant var, normality)
par(mfcol=c(2,2))

#constant var
plot(fitted(lmfit), residuals(lmfit),xlab = "Fitted", ylab = "Residuals")
abline(h=0)
plot(fitted(lmfit),abs(residuals(lmfit)),xlab = "Fitted", ylab = "|Residuals|")

#normality
qqnorm(residuals(lmfit),ylab= "Residuals")
qqline(residuals(lmfit))
hist(residents(lmfit))

shapiro.test(residuals(lmfit))
par(mfcol=c(1,1))

#Q2(2)
coef(lmfit)[1]+coef(lmfit)[2]*1
predict(lmfit, newdata = data.frame(a = 1))

#Q2(3)
#The increase in the expected number of ampules broken when two transfers occur
#relative to one is the slope which is 4.0 broken ampules.

#Q2(4)
amean <- mean(data1$ShipmentRoute)
bmean <- mean(data1$NumberOfAmpules)
predict(lmfit, newdata = data.frame(a = amean))
bmean

#Q2(5)
#Confidence Interval for beta0 and beta1
confint(lmfit)

#Q2(6)
t.test(a,b)
```



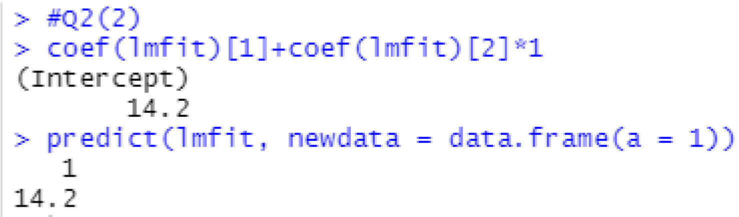
```
> lmfit<- lm(b~a)
> summary(lmfit)

Call:
lm(formula = b ~ a)

Residuals:
   Min     1Q Median     3Q    Max 
  -2.2   -1.2    0.3    0.8    1.8 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  10.2000     0.6633  15.377 3.18e-07 ***
a             4.0000     0.4690   8.528 2.75e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.483 on 8 degrees of freedom
Multiple R-squared:  0.9009,    Adjusted R-squared:  0.8885 
F-statistic: 72.73 on 1 and 8 DF,  p-value: 2.749e-05
```



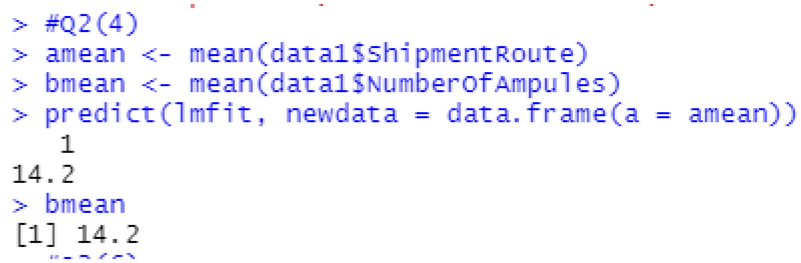
```
> abline(10.2,4)
> abline(lm(b~a))
```



The estimated regression function is $Y = 4.0\ X + 10.2$. Yes this is a good fit, the points are equally distant from the line of best fit.



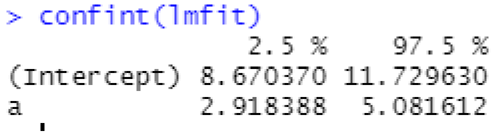
```
> #Q2(2)
> coef(lmfit)[1]+coef(lmfit)[2]*1
(Intercept) 
       14.2 
> predict(lmfit, newdata = data.frame(a = 1))
   1 
14.2 
```



```
> #Q2(4)
> amean <- mean(data1$ShipmentRoute)
> bmean <- mean(data1$NumberOfAmpules)
> predict(lmfit, newdata = data.frame(a = amean))
   1 
14.2 
> bmean
[1] 14.2
```

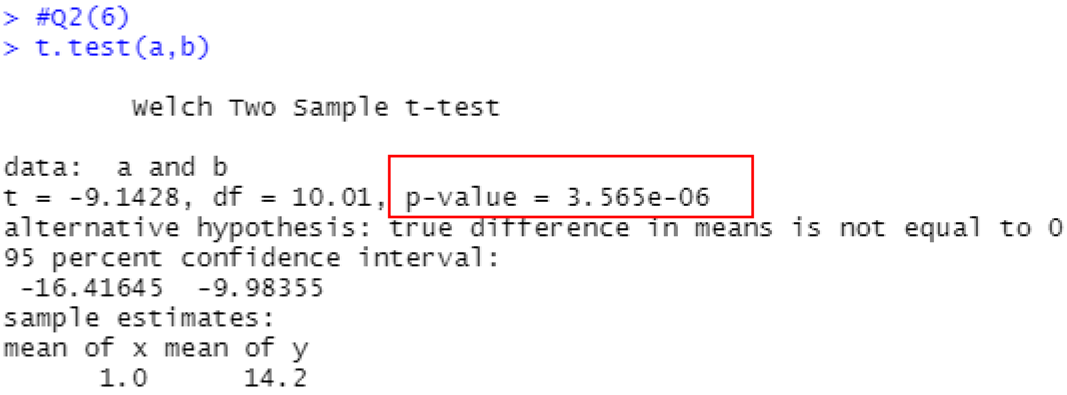
It fits through X bar and Y bar

Q2(5)



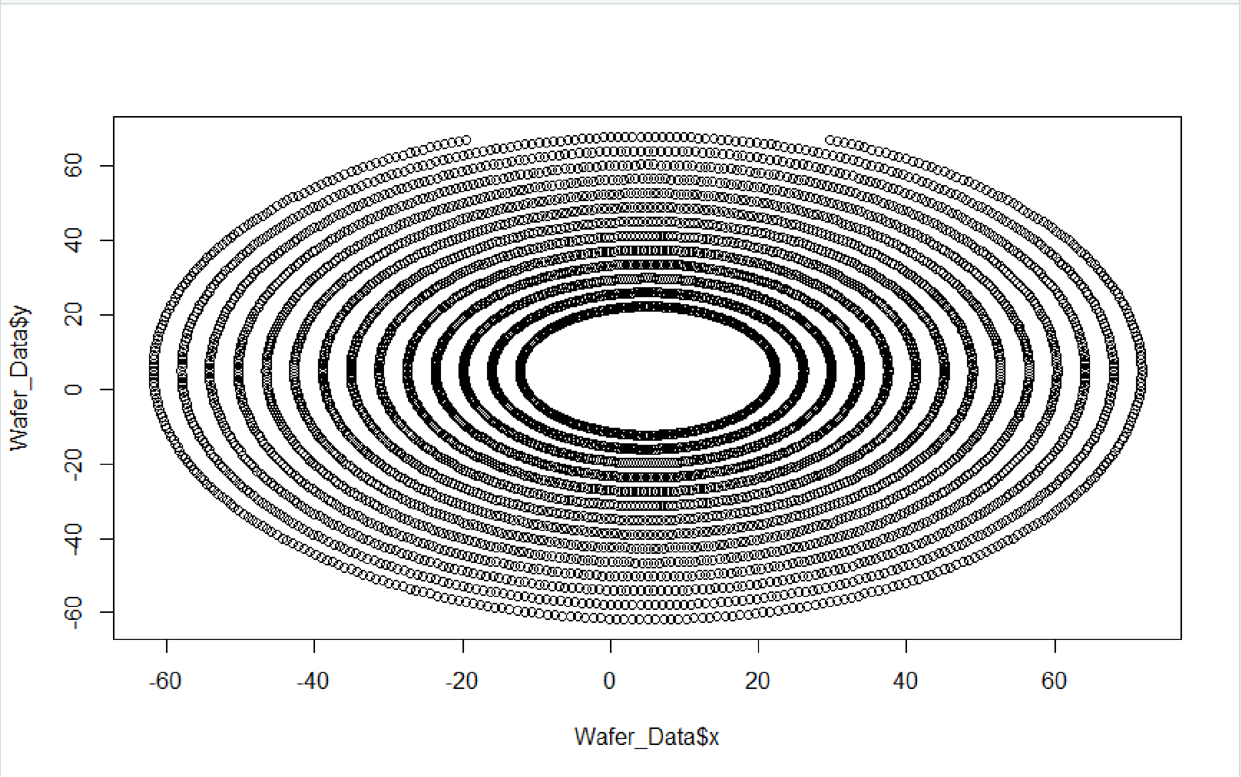
```
> confint(lmfit)
                2.5 %    97.5 %
(Intercept) 8.670370 11.729630
a           2.918388  5.081612
```

The confidence interval for B0 is between 8.6704and 11.7296



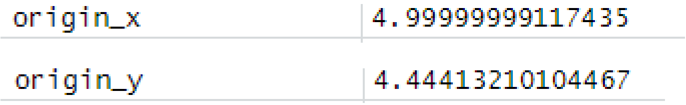
```
> #Q2(6)
> t.test(a,b)

	Welch Two Sample t-test

data:  a and b
t = -9.1428, df = 10.01, p-value = 3.565e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.41645  -9.98355
sample estimates:
mean of x mean of y 
      1.0      14.2 
```

Q3



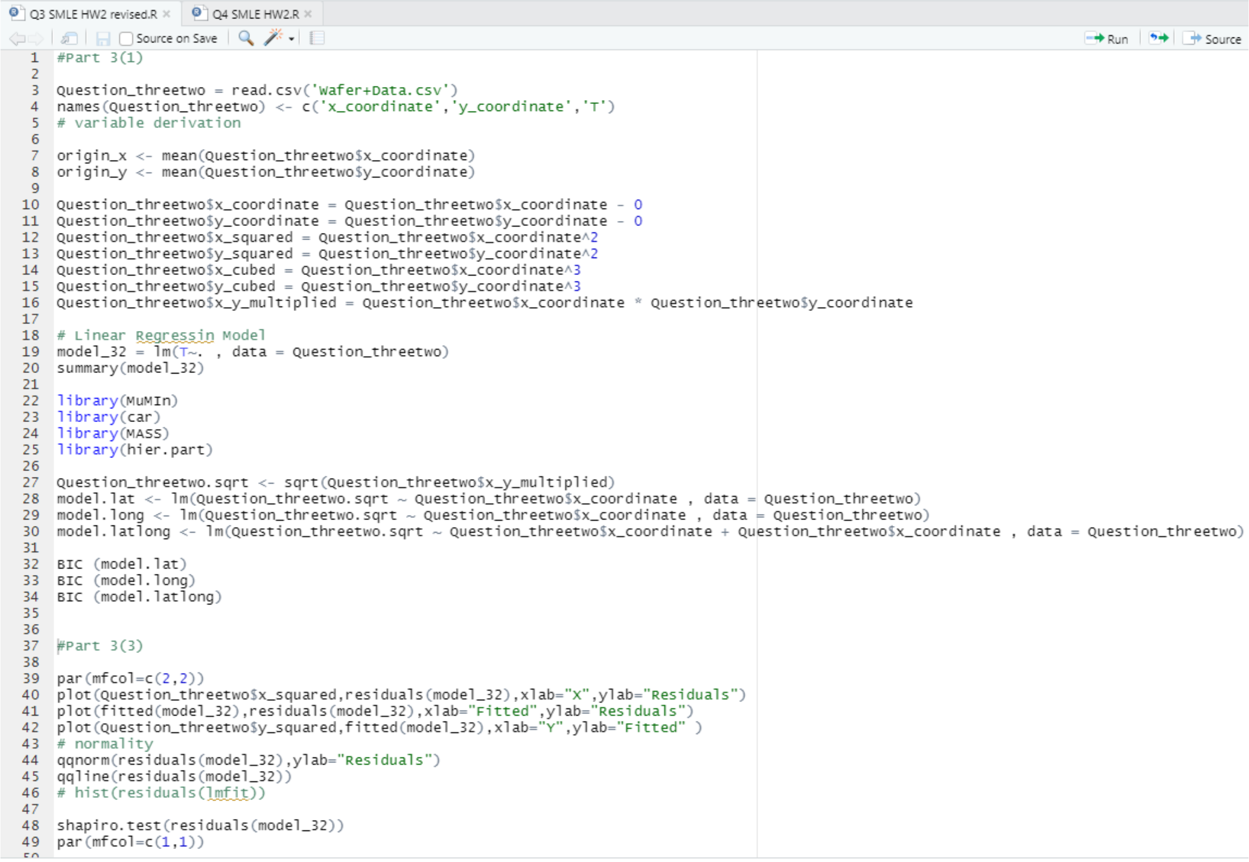
3.1)

The coordinates of circles' origin are at (5,4.44).

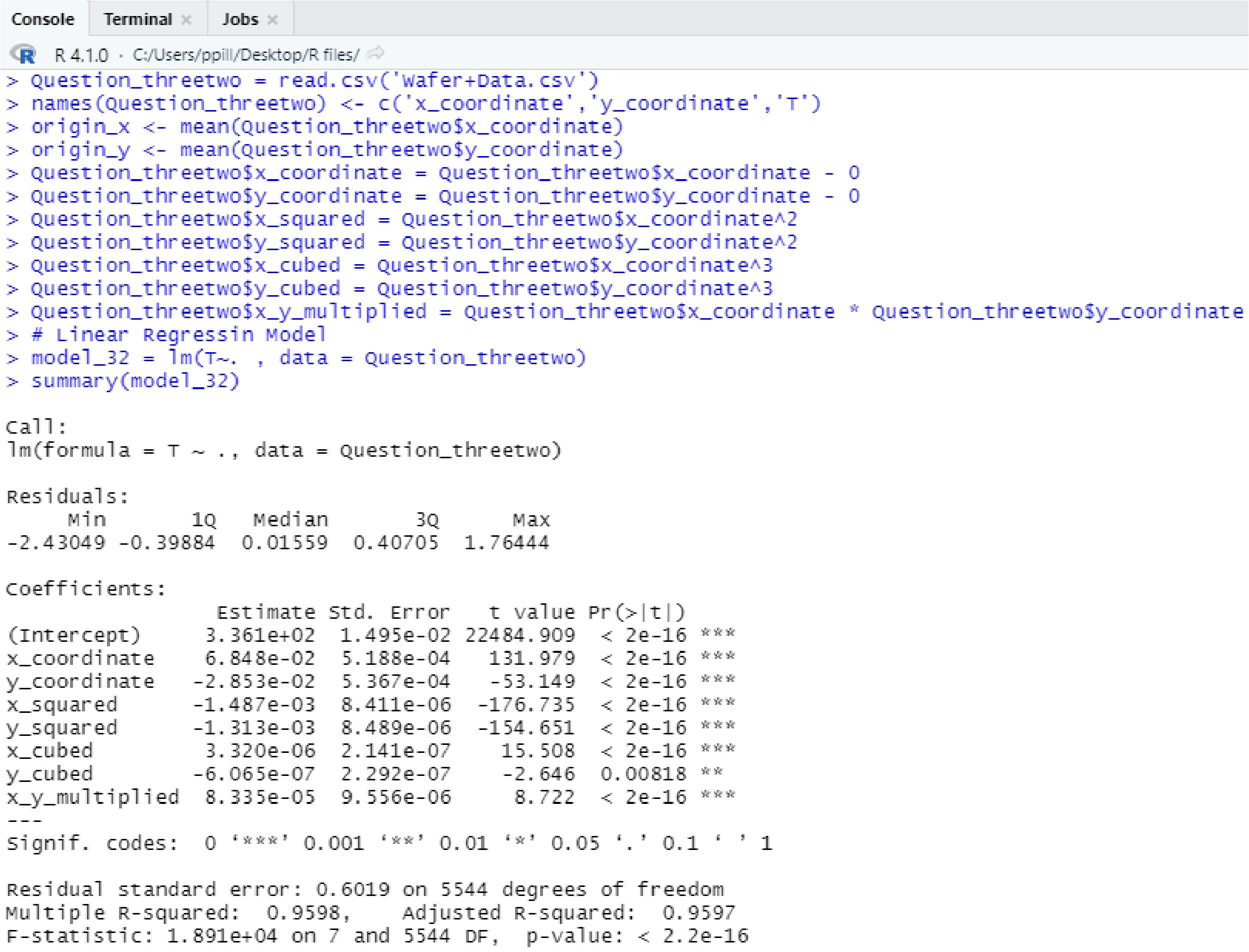


origin_x	4.99999999117435
origin_y	4.44413210104467

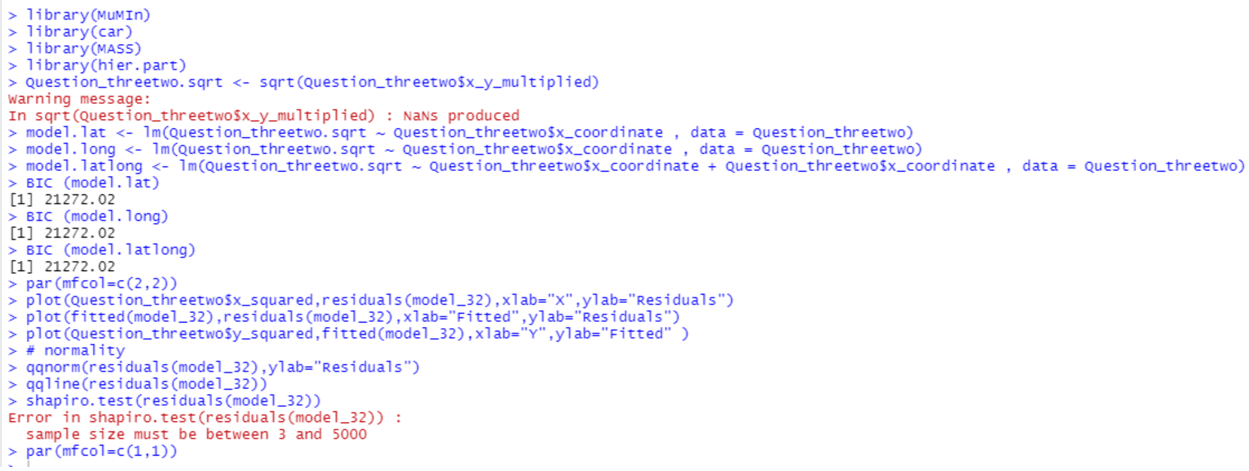
3.2)



```r
#Part 3(1)

Question_threetwo = read.csv('Wafer+Data.csv')
names(Question_threetwo) <- c('x_coordinate','y_coordinate','T')
# variable derivation

origin_x <- mean(Question_threetwo$x_coordinate)
origin_y <- mean(Question_threetwo$y_coordinate)

Question_threetwo$x_coordinate = Question_threetwo$x_coordinate - 0
Question_threetwo$y_coordinate = Question_threetwo$y_coordinate - 0
Question_threetwo$x_squared = Question_threetwo$x_coordinate^2
Question_threetwo$y_squared = Question_threetwo$y_coordinate^2
Question_threetwo$x_cubed = Question_threetwo$x_coordinate^3
Question_threetwo$y_cubed = Question_threetwo$y_coordinate^3
Question_threetwo$x_y_multiplied = Question_threetwo$x_coordinate * Question_threetwo$y_coordinate

# Linear Regressin Model
model_32 = lm(T~.  , data = Question_threetwo)
summary(model_32)

library(MuMIn)
library(car)
library(MASS)
library(hier.part)

Question_threetwo.sqrt <- sqrt(Question_threetwo$x_y_multiplied)
model.lat <- lm(Question_threetwo.sqrt ~ Question_threetwo$x_coordinate , data = Question_threetwo)
model.long <- lm(Question_threetwo.sqrt ~ Question_threetwo$x_coordinate , data = Question_threetwo)
model.latlong <- lm(Question_threetwo.sqrt ~ Question_threetwo$x_coordinate + Question_threetwo$x_coordinate , data = Question_threetwo)

BIC (model.lat)
BIC (model.long)
BIC (model.latlong)


#Part 3(3)

par(mfcol=c(2,2))
plot(Question_threetwo$x_squared,residuals(model_32),xlab="X",ylab="Residuals")
plot(fitted(model_32),residuals(model_32),xlab="Fitted",ylab="Residuals")
plot(Question_threetwo$y_squared,fitted(model_32),xlab="Y",ylab="Fitted" )
# normality
qqnorm(residuals(model_32),ylab="Residuals")
qqline(residuals(model_32))
# hist(residuals(lmfit))

shapiro.test(residuals(model_32))
par(mfcol=c(1,1))
```



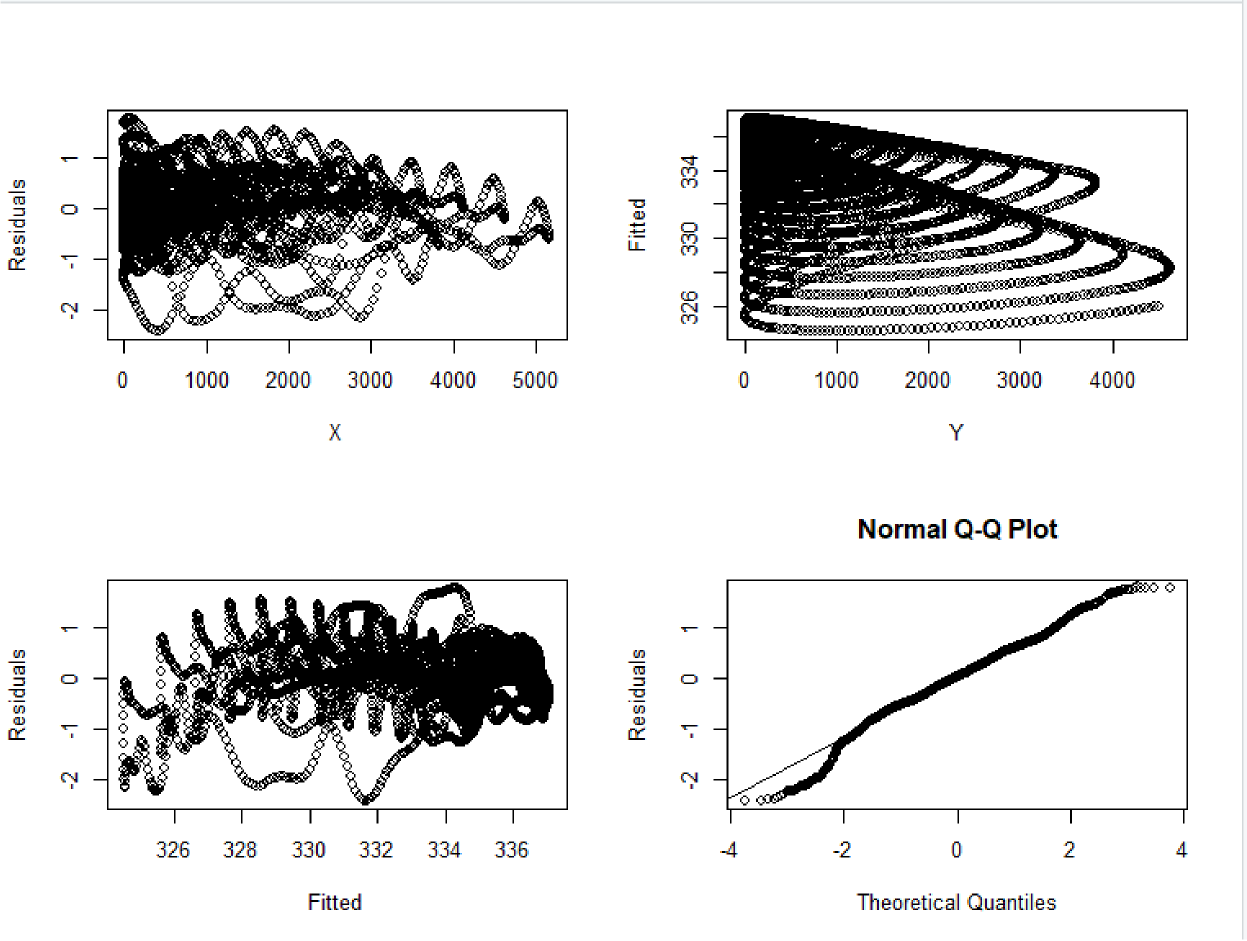
```
> Question_threetwo = read.csv('Wafer+Data.csv')
> names(Question_threetwo) <- c('x_coordinate','y_coordinate','T')
> origin_x <- mean(Question_threetwo$x_coordinate)
> origin_y <- mean(Question_threetwo$y_coordinate)
> Question_threetwo$x_coordinate = Question_threetwo$x_coordinate - 0
> Question_threetwo$y_coordinate = Question_threetwo$y_coordinate - 0
> Question_threetwo$x_squared = Question_threetwo$x_coordinate^2
> Question_threetwo$y_squared = Question_threetwo$y_coordinate^2
> Question_threetwo$x_cubed = Question_threetwo$x_coordinate^3
> Question_threetwo$y_cubed = Question_threetwo$y_coordinate^3
> Question_threetwo$x_y_multiplied = Question_threetwo$x_coordinate * Question_threetwo$y_coordinate
> # Linear Regressin Model
> model_32 = lm(T~. , data = Question_threetwo)
> summary(model_32)

Call:
lm(formula = T ~ ., data = Question_threetwo)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.43049 -0.39884  0.01559  0.40705  1.76444 

Coefficients:
                 Estimate Std. Error   t value Pr(>|t|)    
(Intercept)     3.361e+02  1.495e-02 22484.909  < 2e-16 ***
x_coordinate    6.848e-02  5.188e-04   131.979  < 2e-16 ***
y_coordinate   -2.853e-02  5.367e-04   -53.149  < 2e-16 ***
x_squared      -1.487e-03  8.411e-06  -176.735  < 2e-16 ***
y_squared      -1.313e-03  8.489e-06  -154.651  < 2e-16 ***
x_cubed         3.320e-06  2.141e-07    15.508  < 2e-16 ***
y_cubed        -6.065e-07  2.292e-07    -2.646  0.00818 ** 
x_y_multiplied  8.335e-05  9.556e-06     8.722  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6019 on 5544 degrees of freedom
Multiple R-squared:  0.9598,	Adjusted R-squared:  0.9597 
F-statistic: 1.891e+04 on 7 and 5544 DF,  p-value: < 2.2e-16
```



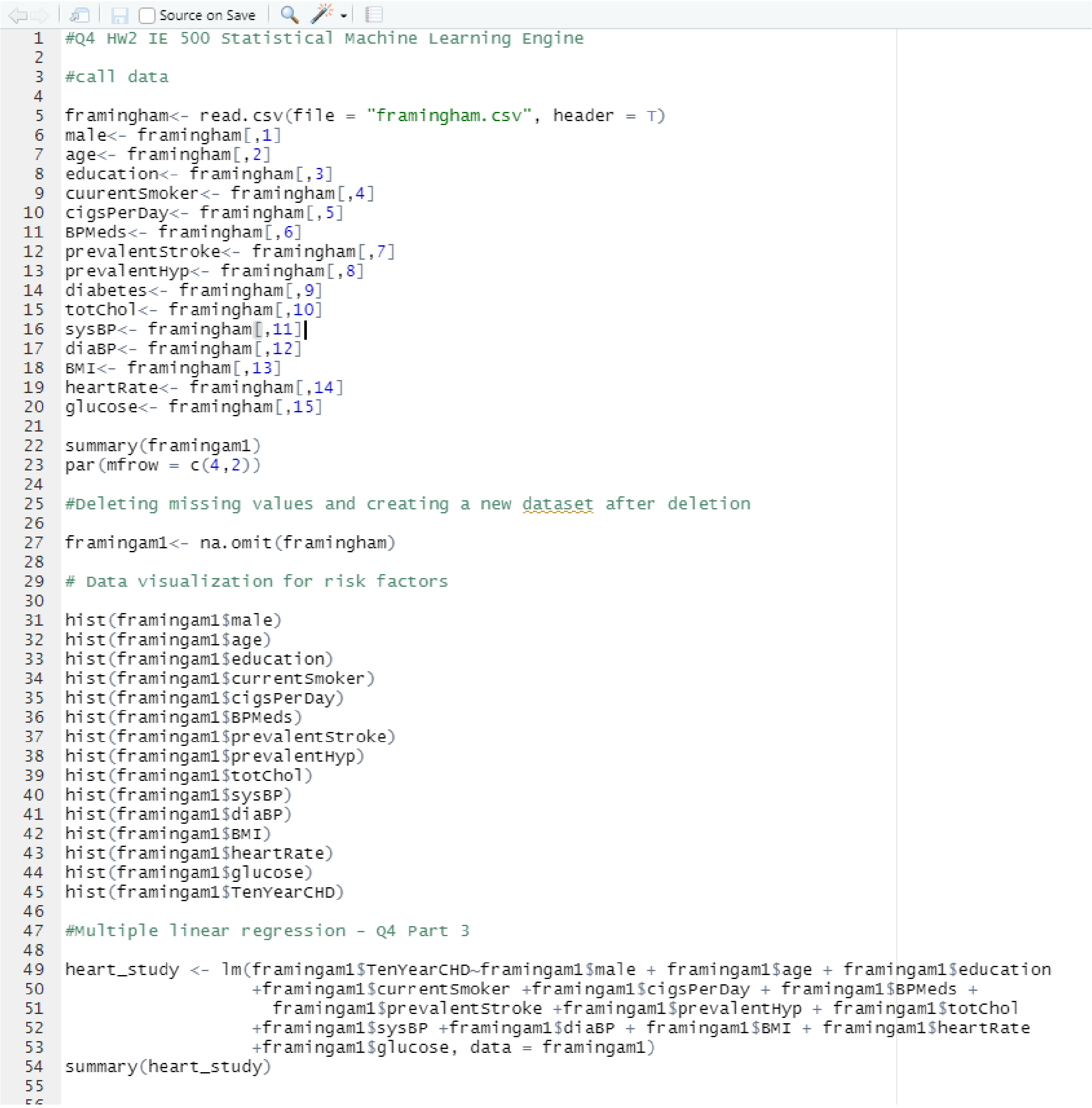
```
> library(MuMIn)
> library(car)
> library(MASS)
> library(hier.part)
> Question_threetwo.sqrt <- sqrt(Question_threetwo$x_y_multiplied)
Warning message:
In sqrt(Question_threetwo$x_y_multiplied) : NaNs produced
> model.lat <- lm(Question_threetwo.sqrt ~ Question_threetwo$x_coordinate , data = Question_threetwo)
> model.long <- lm(Question_threetwo.sqrt ~ Question_threetwo$x_coordinate , data = Question_threetwo)
> model.latlong <- lm(Question_threetwo.sqrt ~ Question_threetwo$x_coordinate + Question_threetwo$x_coordinate , data = Question_threetwo)
> BIC (model.lat)
[1] 21272.02
> BIC (model.long)
[1] 21272.02
> BIC (model.latlong)
[1] 21272.02
> par(mfcol=c(2,2))
> plot(Question_threetwo$x_squared,residuals(model_32),xlab="X",ylab="Residuals")
> plot(fitted(model_32),residuals(model_32),xlab="Fitted",ylab="Residuals")
> plot(Question_threetwo$y_squared,fitted(model_32),xlab="Y",ylab="Fitted" )
> # normality
> qqnorm(residuals(model_32),ylab="Residuals")
> qqline(residuals(model_32))
> shapiro.test(residuals(model_32))
Error in shapiro.test(residuals(model_32)) : 
  sample size must be between 3 and 5000
> par(mfcol=c(1,1))
```

The BIC value is 22272.02

3.3)

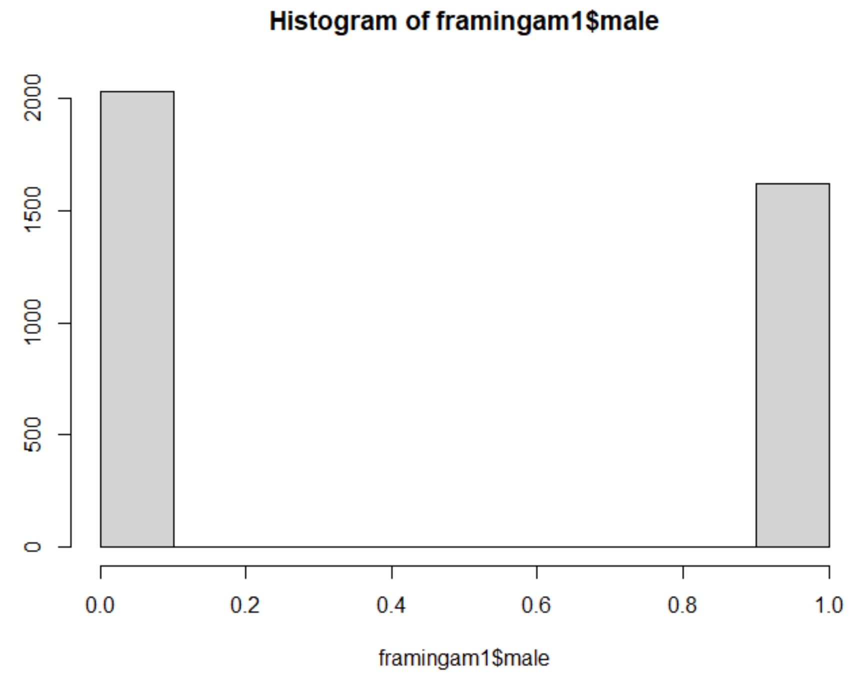

Residuals
1
0
-1
-2
0
1000
2000
3000
4000
5000
X
Fitted
334
330
326
0
1000
2000
3000
4000
Y
Normal Q-Q Plot
Residuals
1
0
-1
-2
326
328
330
332
334
336
Fitted
Residuals
1
0
-1
-2
-4
-2
0
2
4
Theoretical Quantiles

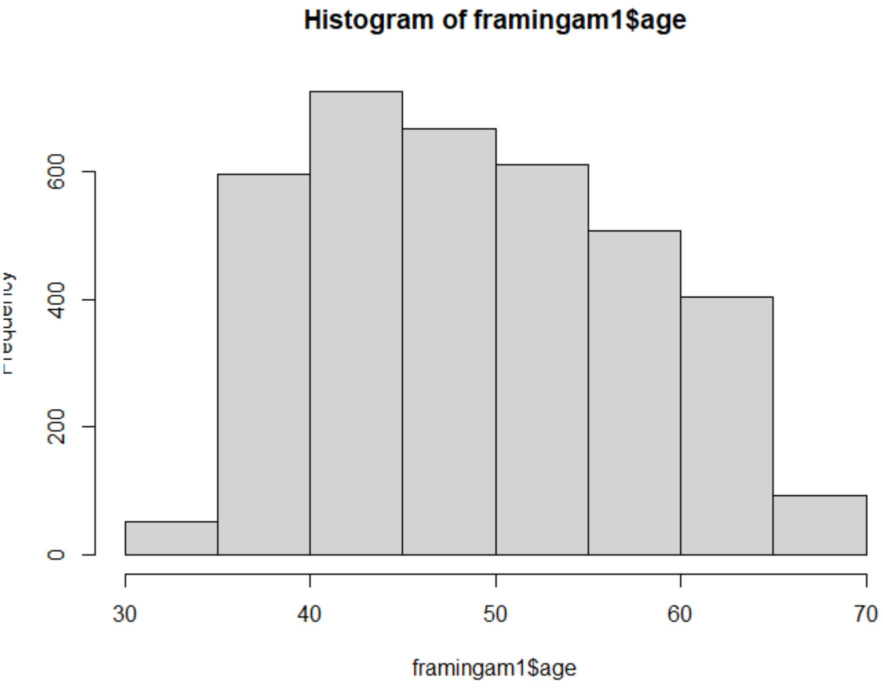
Q4

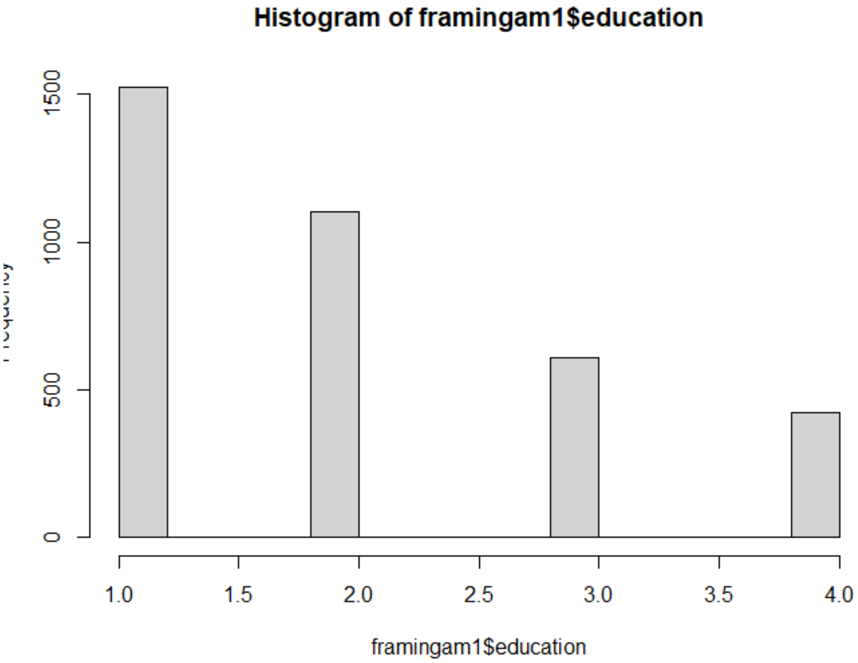


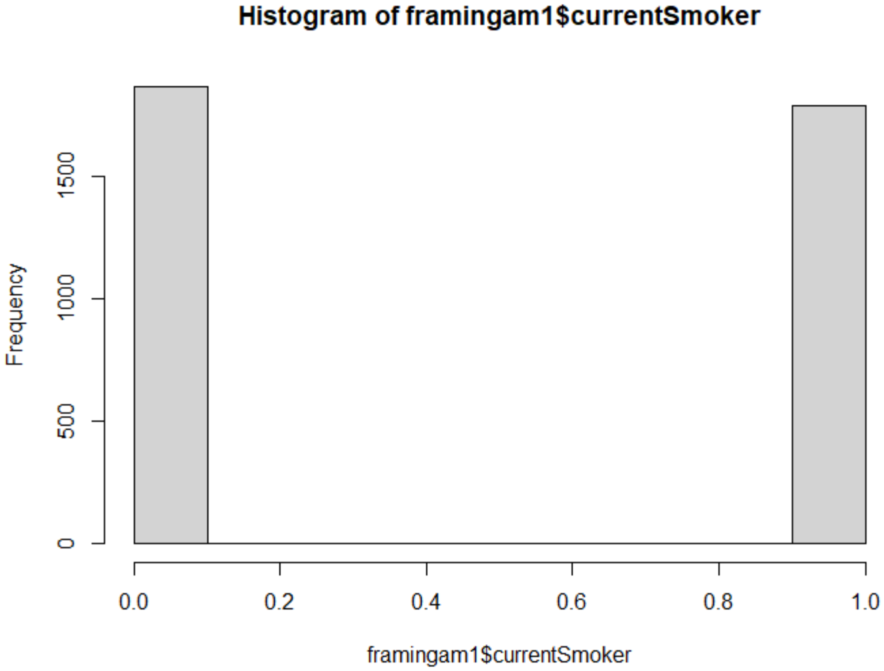
```
#Q4 HW2 IE 500 Statistical Machine Learning Engine

#call data

framingham<- read.csv(file = "framingham.csv", header = T)
male<- framingham[,1]
age<- framingham[,2]
education<- framingham[,3]
cuurentSmoker<- framingham[,4]
cigsPerDay<- framingham[,5]
BPMeds<- framingham[,6]
prevalentStroke<- framingham[,7]
prevalentHyp<- framingham[,8]
diabetes<- framingham[,9]
totChol<- framingham[,10]
sysBP<- framingham[,11]
diaBP<- framingham[,12]
BMI<- framingham[,13]
heartRate<- framingham[,14]
glucose<- framingham[,15]

summary(framingam1)
par(mfrow = c(4,2))

#Deleting missing values and creating a new dataset after deletion

framingam1<- na.omit(framingham)

# Data visualization for risk factors

hist(framingam1$male)
hist(framingam1$age)
hist(framingam1$education)
hist(framingam1$currentSmoker)
hist(framingam1$cigsPerDay)
hist(framingam1$BPMeds)
hist(framingam1$prevalentStroke)
hist(framingam1$prevalentHyp)
hist(framingam1$totChol)
hist(framingam1$sysBP)
hist(framingam1$diaBP)
hist(framingam1$BMI)
hist(framingam1$heartRate)
hist(framingam1$glucose)
hist(framingam1$TenYearCHD)

#Multiple linear regression - Q4 Part 3

heart_study <- lm(framingam1$TenYearCHD~framingam1$male + framingam1$age + framingam1$education
                  +framingam1$currentSmoker +framingam1$cigsPerDay + framingam1$BPMeds +
                    framingam1$prevalentStroke +framingam1$prevalentHyp + framingam1$totChol
                  +framingam1$sysBP +framingam1$diaBP + framingam1$BMI + framingam1$heartRate
                  +framingam1$glucose, data = framingam1)
summary(heart_study)

```

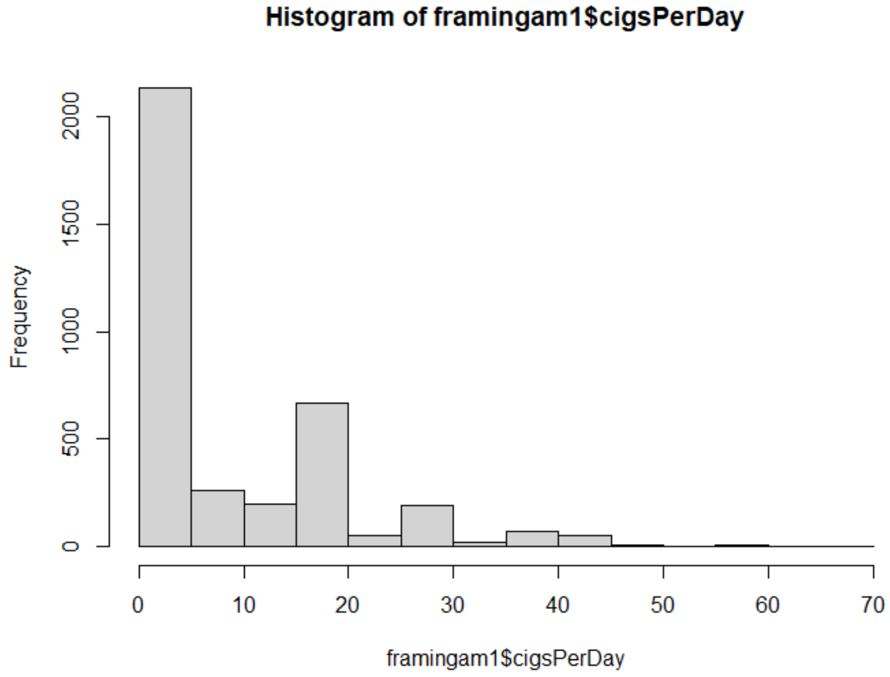
Q4 Part 1

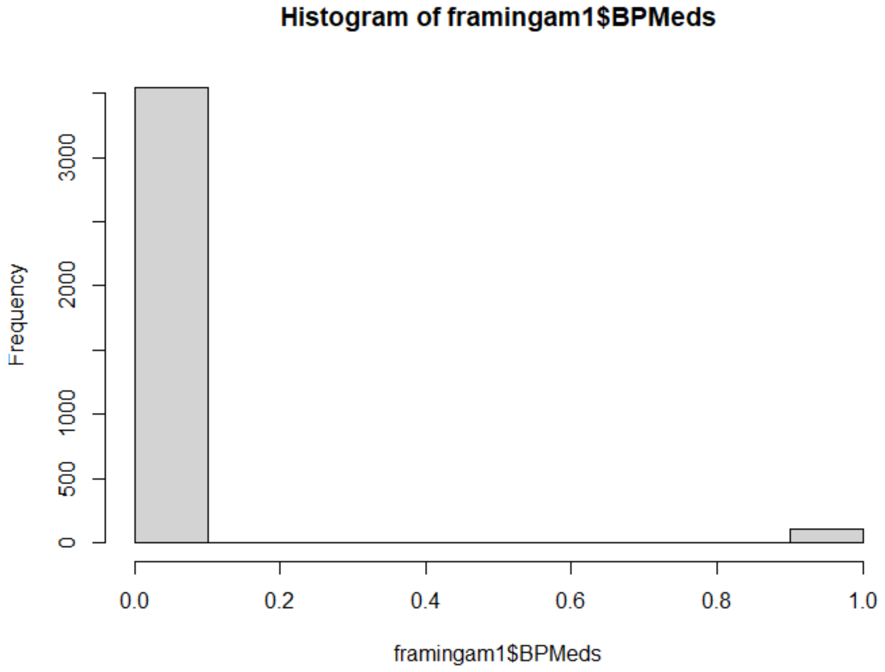

Histogram of framingam1$male
0
500
1000
1500
2000
0.0
0.2
0.4
0.6
0.8
1.0
framingam1$male

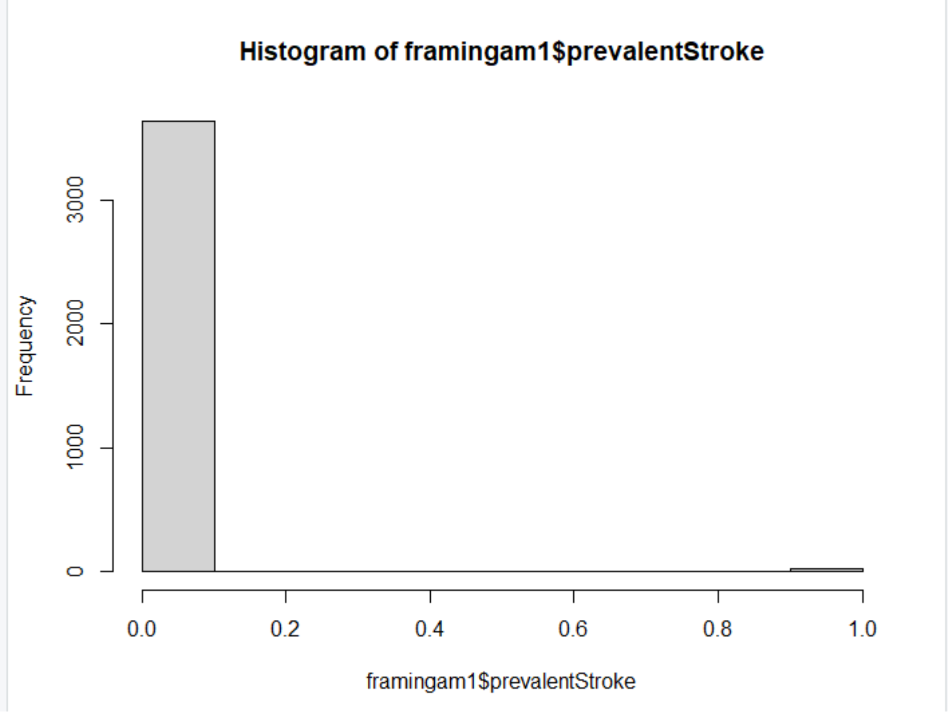

Histogram of framingam1$age
Frequency
0
200
400
600
30
40
50
60
70
framingam1$age

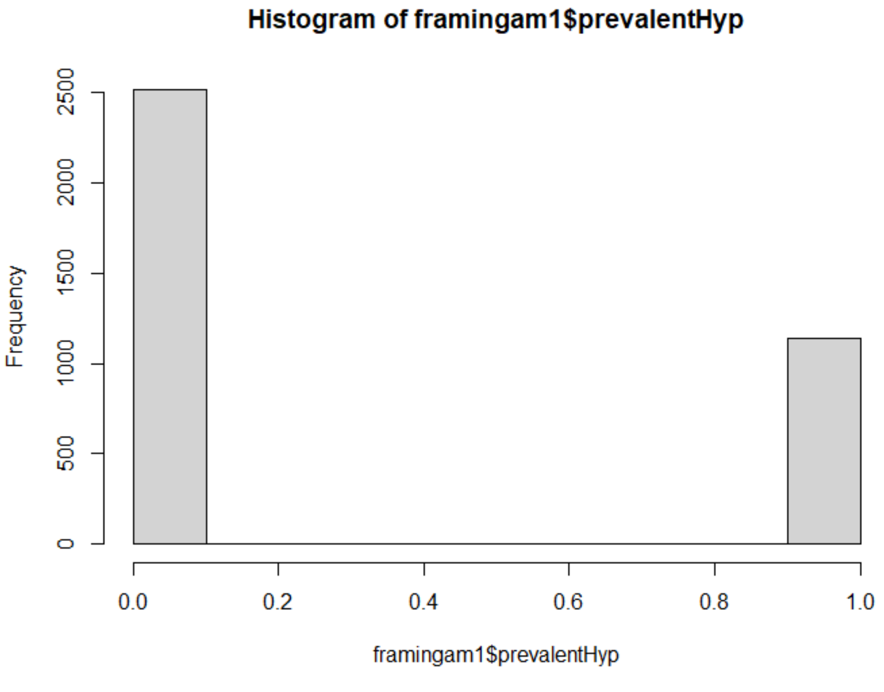

Histogram of framingam1$education
Frequency
0
500
1000
1500
1.0
1.5
2.0
2.5
3.0
3.5
4.0
framingam1$education

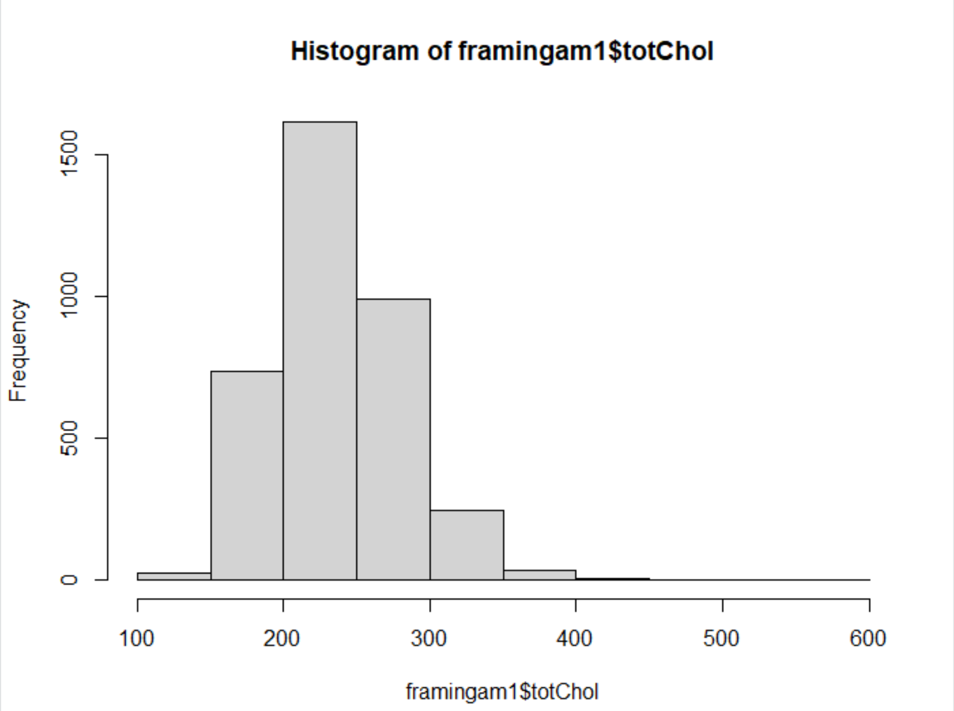

Histogram of framingam1$currentSmoker
Frequency
0
500
1000
1500
0.0
0.2
0.4
0.6
0.8
1.0
framingam1$currentSmoker

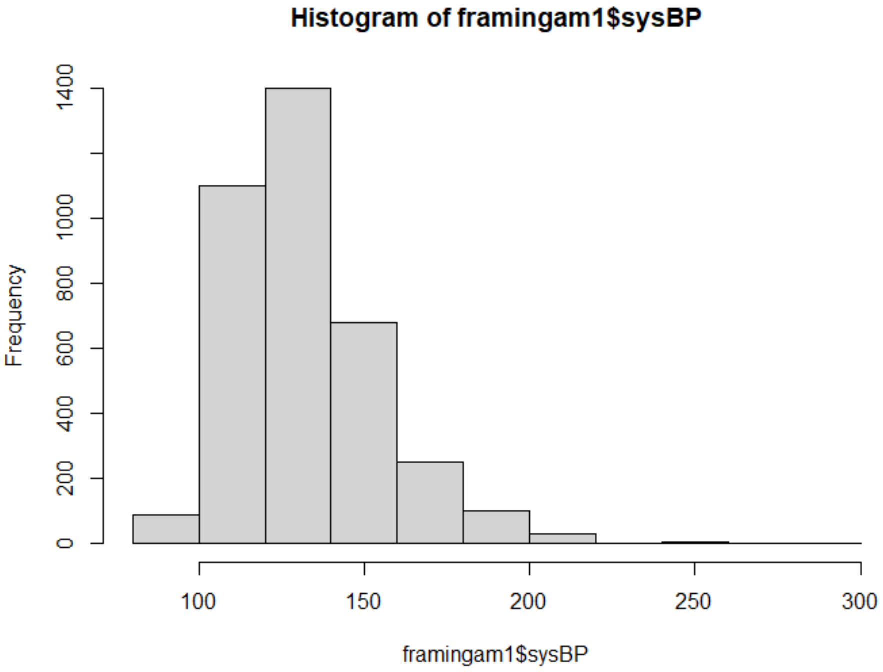

Histogram of framingam1$cigsPerDay
Frequency
0
500
1000
1500
2000
0
10
20
30
40
50
60
70
framingam1$cigsPerDay

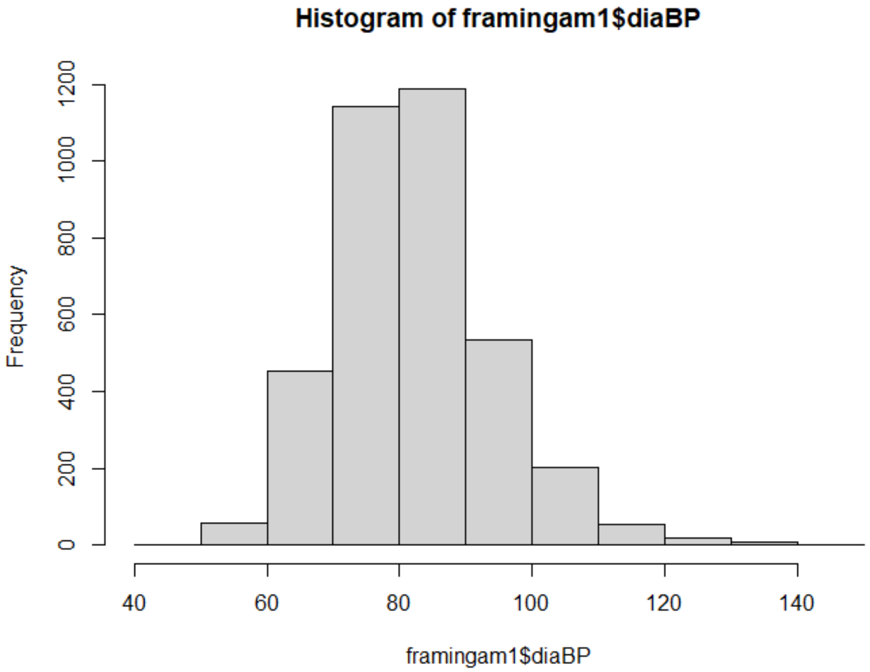

Histogram of framingam1$BPMeds
Frequency
0
500
1000
2000
3000
0.0
0.2
0.4
0.6
0.8
1.0
framingam1$BPMeds

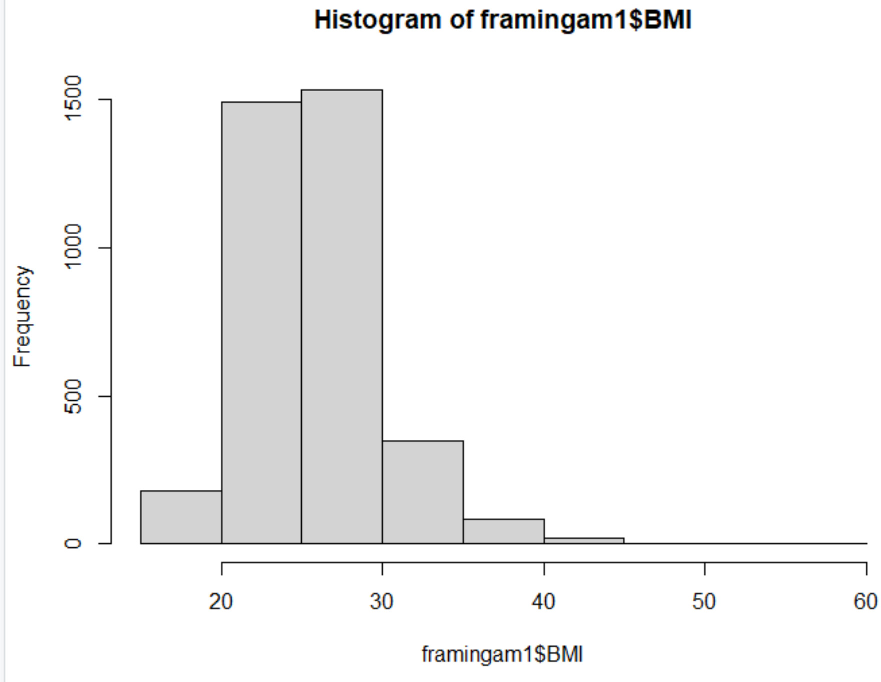

Histogram of framingam1$prevalentStroke
Frequency
0
1000
2000
3000
0.0
0.2
0.4
0.6
0.8
1.0
framingam1$prevalentStroke

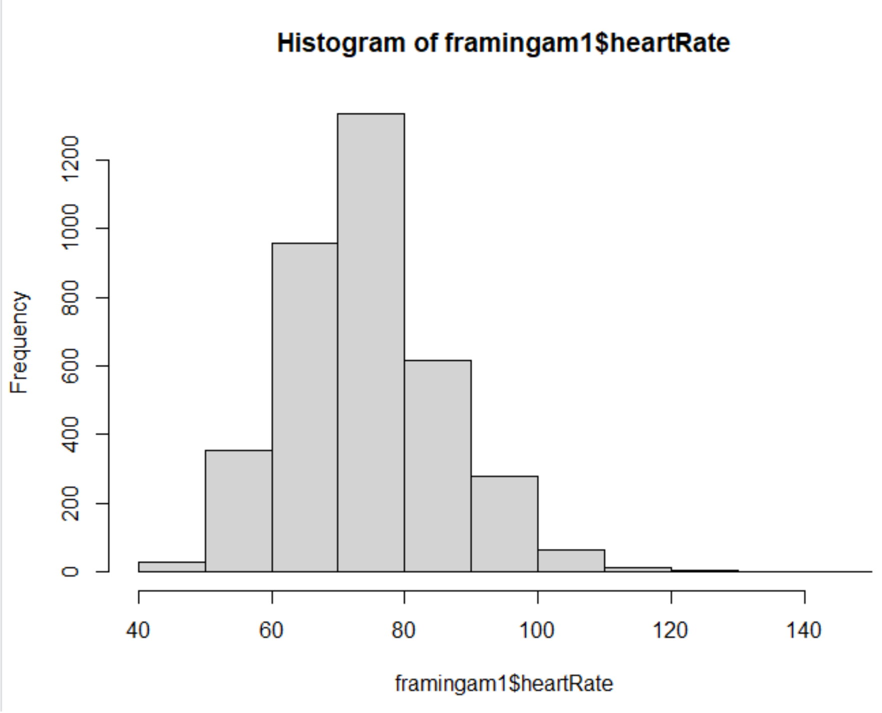

Histogram of framingam1$prevalentHyp
Frequency
0
500
1000
1500
2000
2500
0.0
0.2
0.4
0.6
0.8
1.0
framingam1$prevalentHyp

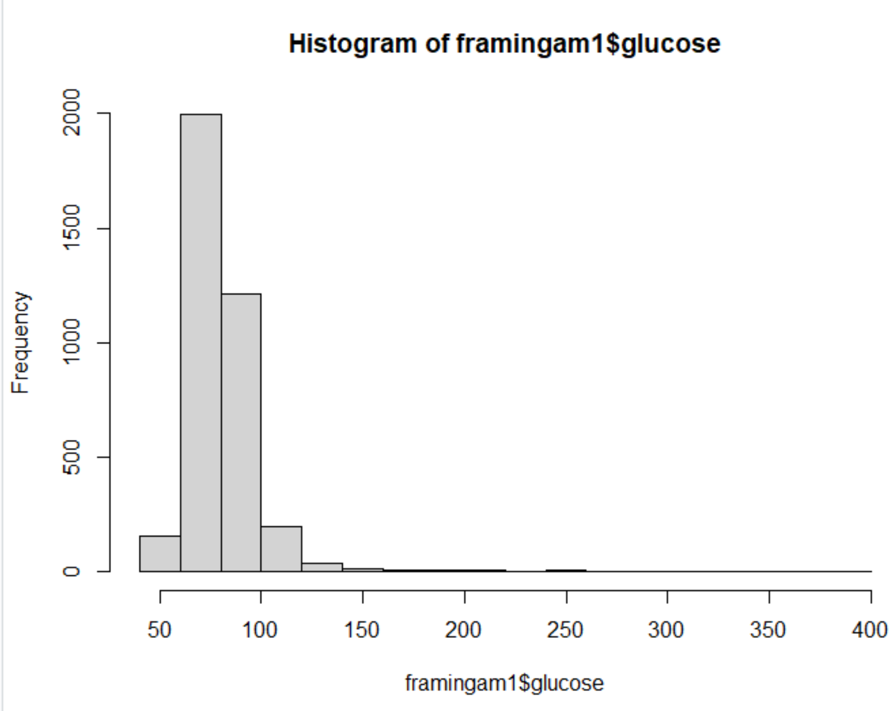

Histogram of framingam1$totChol
Frequency
0
500
1000
1500
100
200
300
400
500
600
framingam1$totChol


Histogram of framingam1$sysBP
Frequency
0
200
400
600
800
1000
1400
100
150
200
250
300
framingam1$sysBP


Histogram of framingam1$diaBP
Frequency
0
200
400
600
800
1000
1200
40
60
80
100
120
140
framingam1$diaBP


Histogram of framingam1$BMI
Frequency
0
500
1000
1500
20
30
40
50
60
framingam1$BMI





Q4 Part 2

The methods to deal with missing data consist of the following:

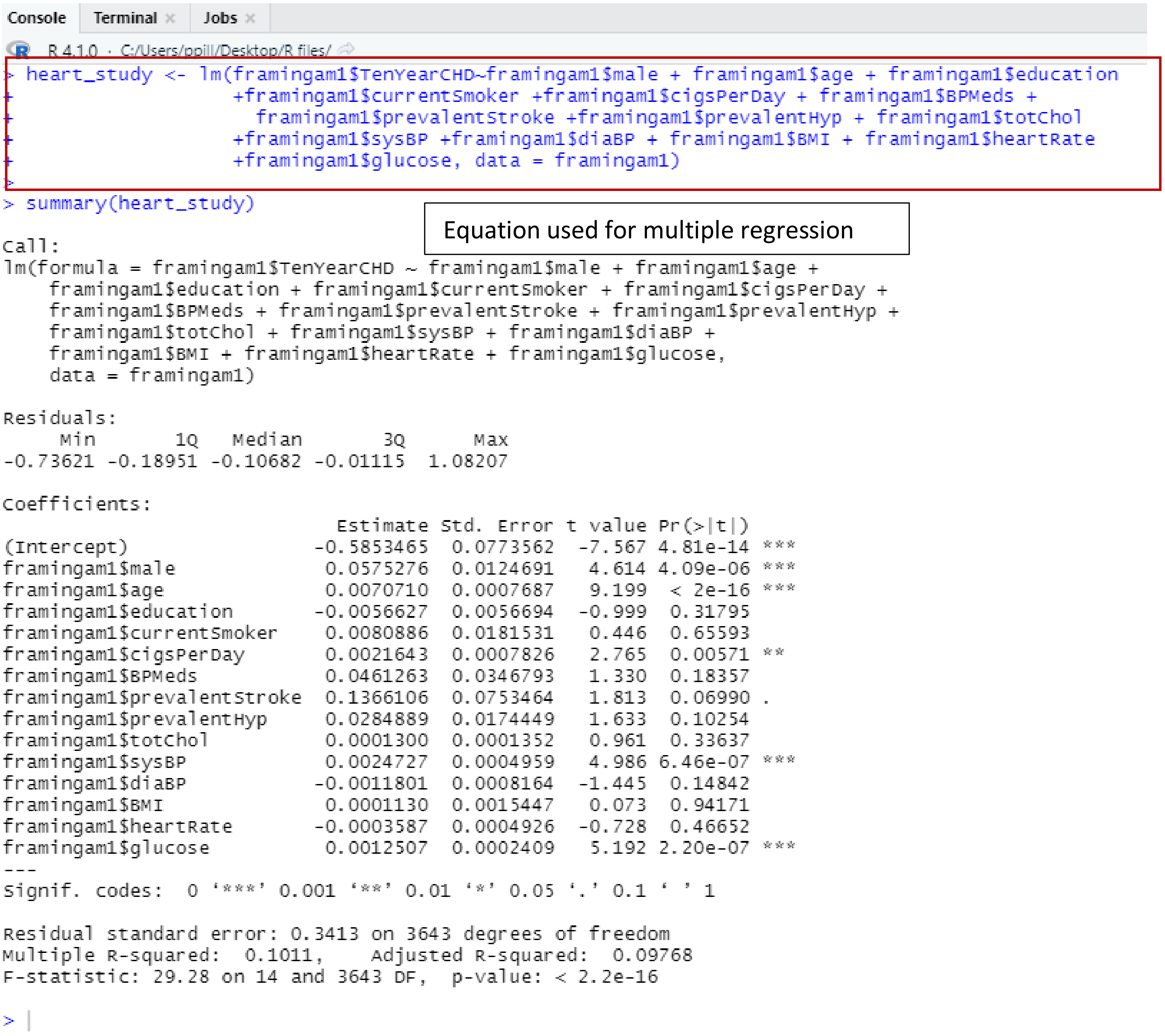
- Listwise deletion: This consists of deleting data with the missing values. If the sample size is quite large, then these missing values can be deleted without it adversely affecting the ability the compute the statistics. One has to ensure that these values are missing in random patterns and that the category of participants are not missing
- Pairwise deletion: This consists of analysis of all cases of the concerned variables present and maximizes data that's available. This increases analysis power but can assume data missing completely at random. Deleting pairwise can lead to different observations leading to different parts of the model which makes it difficult to understand.
- Recovery of values: This involves reaching out to participants and requesting them for the missing values to be filled up. It is noticed that for face to face meetings, having an extra check for the missing values makes a difference towards the study results.
- Educated guessing: It is a form of imputation which is a type of using substitute values in place of missing values. In educated guessing a missing value is predicted.
- Average imputation: It uses the mean response of the participants to complete the missing values.
- Common-point imputation: This method uses the most frequently used value or the common point. It should only be utilized only if there is proper justification and evidence to support the use of this method.
- Regression estimation: This is used to estimate the missing value from other values by using data to create regression equations to predict the missing data.
- Multiple imputation: This consists of the statistical package generating feasible values off the existing data and then takes average of the "simulated datasets" by accounting for random errors in the predictions.

Reference:

Sauro, J. (2015, June 2). *7 ways to handle missing data*. MeasuringU. Retrieved October 20, 2021, from https://measuringu.com/handle-missing-data/.

Swalin, A. (2018, January 31). *How to handle missing data*. Medium. Retrieved October 20, 2021, from https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4.

Q4 Part 3



```
> heart_study <- lm(framingam1$TenYearCHD~framingam1$male + framingam1$age + framingam1$education
+                  +framingam1$currentSmoker +framingam1$cigsPerDay + framingam1$BPMeds +
+                    framingam1$prevalentStroke +framingam1$prevalentHyp + framingam1$totChol
+                  +framingam1$sysBP +framingam1$diaBP + framingam1$BMI + framingam1$heartRate
+                  +framingam1$glucose, data = framingam1)
> 
> summary(heart_study)

Call:
lm(formula = framingam1$TenYearCHD ~ framingam1$male + framingam1$age + 
    framingam1$education + framingam1$currentSmoker + framingam1$cigsPerDay + 
    framingam1$BPMeds + framingam1$prevalentStroke + framingam1$prevalentHyp + 
    framingam1$totChol + framingam1$sysBP + framingam1$diaBP + 
    framingam1$BMI + framingam1$heartRate + framingam1$glucose, 
    data = framingam1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.73621 -0.18951 -0.10682 -0.01115  1.08207 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)               -0.5853465  0.0773562  -7.567 4.81e-14 ***
framingam1$male            0.0575276  0.0124691   4.614 4.09e-06 ***
framingam1$age             0.0070710  0.0007687   9.199  < 2e-16 ***
framingam1$education      -0.0056627  0.0056694  -0.999  0.31795    
framingam1$currentSmoker   0.0080886  0.0181531   0.446  0.65593    
framingam1$cigsPerDay      0.0021643  0.0007826   2.765  0.00571 ** 
framingam1$BPMeds          0.0461263  0.0346793   1.330  0.18357    
framingam1$prevalentStroke 0.1366106  0.0753464   1.813  0.06990 .  
framingam1$prevalentHyp    0.0284889  0.0174449   1.633  0.10254    
framingam1$totChol         0.0001300  0.0001352   0.961  0.33637    
framingam1$sysBP           0.0024727  0.0004959   4.986 6.46e-07 ***
framingam1$diaBP          -0.0011801  0.0008164  -1.445  0.14842    
framingam1$BMI             0.0001130  0.0015447   0.073  0.94171    
framingam1$heartRate      -0.0003587  0.0004926  -0.728  0.46652    
framingam1$glucose         0.0012507  0.0002409   5.192 2.20e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3413 on 3643 degrees of freedom
Multiple R-squared:  0.1011,	Adjusted R-squared:  0.09768 
F-statistic: 29.28 on 14 and 3643 DF,  p-value: < 2.2e-16

> |
```

Equation used for multiple regression

According to the summary of the multiple regression, the significant risk factors are education, current smoker or not , total cholesterol, BMI and heart rate

Q4 Part 4

Using a cutoff of 0.5, we can attempt a prediction accuracy using current smoker or not and BMI as they are the only factors to have a p value more than 0.5.

Q5 Part 2

Fixed Effects and Random effects differences

The fixed effects presume that the independent variable is involved in a constant relationship with the dependent variable throughout all the observations. Random effects presume that the independent variable has a fixed relationship with the dependent variable through all the observations. However, the fixed effects might differ from one observation to another. One of the main differences between fixed and random effects is prediction. A fixed effects model supports inference for only the levels of features utilized for training but a random effects model permits allowing inference about the population from which the sample is drawn. It is possible that there are factors that were not present in the sample. If the effect size related to the variance between samples drawn is huge enough then the conclusion can be that the population displays the effect. Fixed models are suggested when the fixed effect is of primary concern. Random effects model is useful for studies where it is difficult to figure the participants belong to which categories. If a fixed effects model is applied to a random sample, then an inference cannot be made about the data outside the sample dataset for the model. The fixed effects model surmises that the individual effect is correlated to the independent variable. The random effect model permits having predictions on the population data presuming the normal distribution. The random effects model assumes that the individual effects is unrelated with the explanatory variables.

Reference: Kumar , A. (2021, October 2). *Fixed vs random vs mixed effects models - examples*. Data Analytics. Retrieved October 20, 2021, from https://vitalflux.com/fixed-vs-random-vs-mixed-effects-models-examples/.