OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

J.E. CAIRNES SCHOOL OF BUSINESS & ECONOMICS

INDIVIDUAL ASSIGNMENT COVER PAGE

Module Name and Code: Applied Customer Analytics MS5108

Student Name and ID: Dilip Venkatesan Sankar 22225743

In submitting this assignment, I am aware that it is my responsibility to adhere to the submission guidelines. Please tick (double click… or Yes/No) for the following:

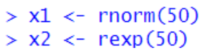
	Yes	No
I am aware of what the NUI Galway plagiarism policy entails.	☒	□
I have named the assignment file (MS Word docx , doc, or jar) to contain my student ID and the module code and assignment number (e.g., 1187404_MS220_A1.docx or 1187404_MS220_A1.jar).	☒	□

Declaration for this Assignment Submission*:*

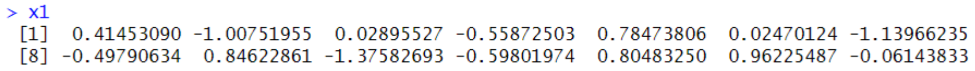
In submitting this work, I confirm that it is entirely my own. I acknowledge that I may be invited to interview if there is any concern in relation to the integrity, and I am aware that any breach will be subject to the University's Procedures for dealing with plagiarism (http://www.universityofgalway.ie/plagiarism).

Question 1:

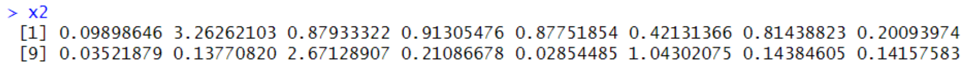
a) The two vectors x1 and x2 are created using Normal Distribution (rnorm) and Exponential Distribution (rexp) respectively,



```
> x1 <- rnorm(50)
> x2 <- rexp(50)
```

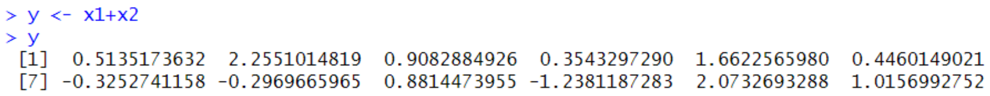


```
> x1
 [1]  0.41453090 -1.00751955  0.02895527 -0.55872503  0.78473806  0.02470124 -1.13966235
 [8] -0.49790634  0.84622861 -1.37582693 -0.59801974  0.80483250  0.96225487 -0.06143833
```



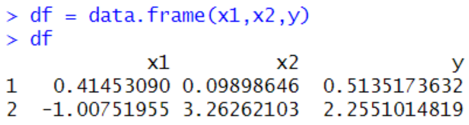
```
> x2
 [1] 0.09898646 3.26262103 0.87933322 0.91305476 0.87751854 0.42131366 0.81438823 0.20093974
 [9] 0.03521879 0.13770820 2.67128907 0.21086678 0.02854485 1.04302075 0.14384605 0.14157583
```

b) Another vector y is created using the linear combination of x1 and x2,

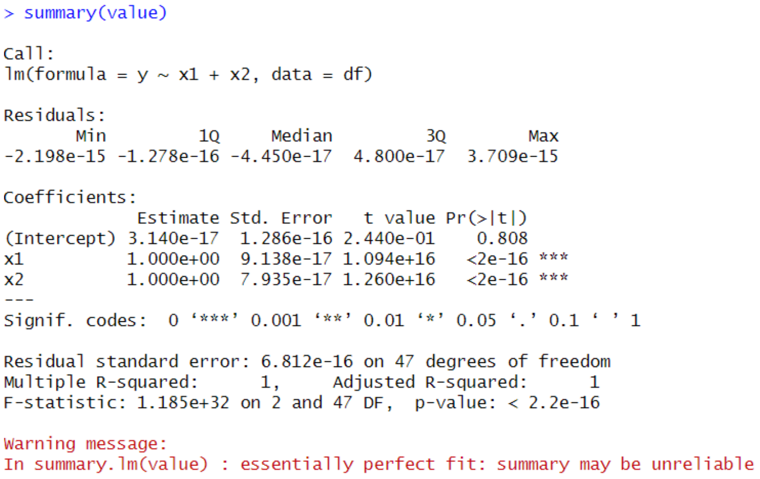


```
> y <- x1+x2
> y
 [1]  0.5135173632  2.2551014819  0.9082884926  0.3543297290  1.6622565980  0.4460149021
 [7] -0.3252741158 -0.2969665965  0.8814473955 -1.2381187283  2.0732693288  1.0156992752
```

c) All three vectors are converted into a data frame and a linear regression model is created lm(),



```
> df = data.frame(x1,x2,y)
> df
           x1         x2             y
1  0.41453090 0.09898646  0.5135173632
2 -1.00751955 3.26262103  2.2551014819
```



```
> summary(value)

Call:
lm(formula = y ~ x1 + x2, data = df)

Residuals:
       Min         1Q     Median         3Q        Max 
-2.198e-15 -1.278e-16 -4.450e-17  4.800e-17  3.709e-15 

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)    
(Intercept) 3.140e-17  1.286e-16 2.440e-01    0.808    
x1          1.000e+00  9.138e-17 1.094e+16   <2e-16 ***
x2          1.000e+00  7.935e-17 1.260e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.812e-16 on 47 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 1.185e+32 on 2 and 47 DF,  p-value: < 2.2e-16

Warning message:
In summary.lm(value) : essentially perfect fit: summary may be unreliable
```

The Residuals which is the distance between the actual data point and the regression line which has the below values,

1. Min = -2.198e-15 and Max = 3.709e-15 almost reflect each other
2. 1st Quartile = -1.278e-16 and 3rd Quartile = 4.800e-17 do not reflect each other
3. Median = -4.450e-17 which is very much less than zero

So, in this model, the residuals are not normally distributed which is also checked using a histogram.

The coefficient shows y intercept = 3.140e-17 which is the point where the regression line crosses the y-axis and the slope(x1 and x2) = 1.000e+00 based on which we can create the linear model equation,

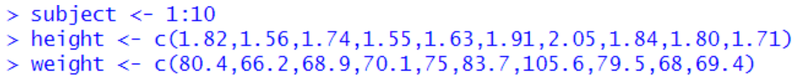
$$y = m.x + b$$

Where y is the Predicted value, m is the slope, x is the independent value (x1+x2) and b is the intercept.

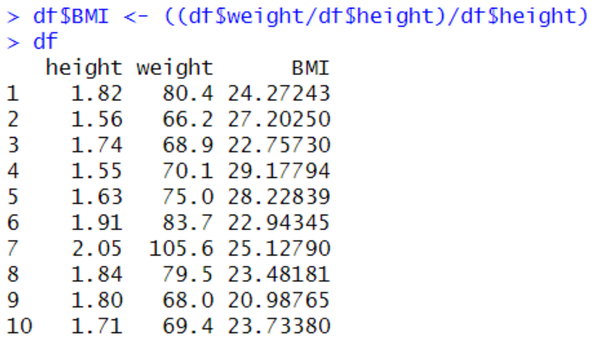
The p-value of intercept(y) = 0.808, slope(x1 and x2) < 2e-16 which shows they are lower than 0.05 supporting the Alternate hypothesis and having a significant difference.
Residual Standard error denotes the average distance that the observed values fall from the regression line = 6.812e-16 with 47 degrees of freedom
Multiple and adjacent R-squared values show how much variance the dependent variable can be accounted for by the independent variable which shows 100% of the variance in y is accounted for by x1 and x2 measures.
F- Statistic = 1.185e+32 and p-value < 2.2e-16 which is lesser than 0.05 which means x1 or x2 is significantly related to y.

Question 2:

a) The dataset is converted into a data frame and the BMI is calculated,

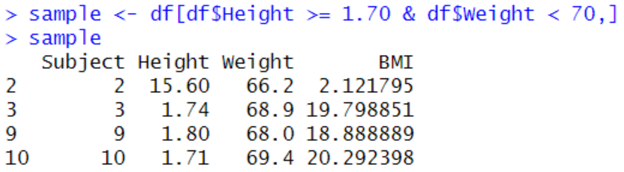


```
> subject <- 1:10
> height <- c(1.82,1.56,1.74,1.55,1.63,1.91,2.05,1.84,1.80,1.71)
> weight <- c(80.4,66.2,68.9,70.1,75,83.7,105.6,79.5,68,69.4)
```



```
> df$BMI <- ((df$weight/df$height)/df$height)
> df
  height weight      BMI
1    1.82   80.4 24.27243
2    1.56   66.2 27.20250
3    1.74   68.9 22.75730
4    1.55   70.1 29.17794
5    1.63   75.0 28.22839
6    1.91   83.7 22.94345
7    2.05  105.6 25.12790
8    1.84   79.5 23.48181
9    1.80   68.0 20.98765
10   1.71   69.4 23.73380
```

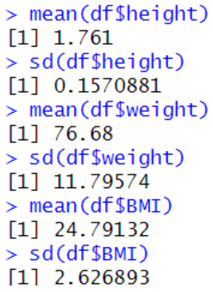
b) An object called a sample is created with the condition height is ≥1.70 and the weight is <70,



```
> sample <- df[df$Height >= 1.70 & df$Weight < 70,]
> sample
   Subject Height Weight       BMI
2        2  15.60   66.2  2.121795
3        3   1.74   68.9 19.798851
9        9   1.80   68.0 18.888889
10      10   1.71   69.4 20.292398
```

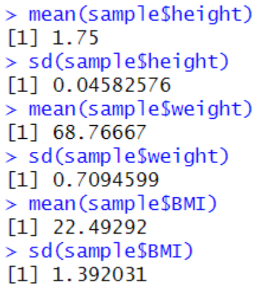
c) The mean and SD of df and the sample for height, weight, and BMI are shown below,

Values of df object



```
> mean(df$height)
[1] 1.761
> sd(df$height)
[1] 0.1570881
> mean(df$weight)
[1] 76.68
> sd(df$weight)
[1] 11.79574
> mean(df$BMI)
[1] 24.79132
> sd(df$BMI)
[1] 2.626893
```

Values of the sample object



```
> mean(sample$height)
[1] 1.75
> sd(sample$height)
[1] 0.04582576
> mean(sample$weight)
[1] 68.76667
> sd(sample$weight)
[1] 0.7094599
> mean(sample$BMI)
[1] 22.49292
> sd(sample$BMI)
[1] 1.392031
```

Based on the above comparison we can incur that the mean and SD for height, weight and BMI in the df object are greater compared to the values in sample object