iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** III **Month of publication:** March 2022

DOI: https://doi.org/10.22214/ijraset.2022.40640

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

Explanation of BMI data using Linear Regression Model in R

Sukhvir Singh
AP, Department of Computer Applications, Gulzar Group of Institutes, Khanna, Ludhiana

***Abstract:** This paper describes the regression analysis between different variable like Weight & BMI, Weight & Height, and Height & BMI using Linear Regression Model & data visualization techniques in R Programming from a sample data of 68 students of BCA. The collected data were analyzed for underweight, overweight, obese personalities by using conditional statements. The result of the model will give Residual Standard Error, Multiple R^2, Adjusted R^2, F-statistic and p-value. There is visualization of data using ggplot() and geom() in last steps.*
***Keywords:** BMI, Multiple R^2, Adjusted R^2, F-statistic, p-value, R, ggplot, geom.*

I. INTRODUCTION TO R PROGRAMMING [8]

R is developed by two personalities one is Ross Ihaka and other is Robert Clifford Gentleman. Ross Ihaka, Professor of Statistics at University of Auckland, completed his PhD from University of California in 1985 & Robert Clifford Gentleman, PhD from University of Washington in 1988, founder director of Centre of Computational Biomedicine at Harvard Medical School. The letter "R" of R Programming is taken from the first alphabet of the names of both the programmers. The language was developed in 1993 at A.T. & T Bell Labs USA. The important features of the language are effectiveness, simple to learn, comprise of loops, conditional statements, graphical tools, various testing tools like t-test, F-test, chi square test, easy representation of data using scatter plot, bar plot, box plot and many more. We can easily import CSV (Comma Separated Values) or Excel data files in R and can work on that data. The key part of this paper focuses on data visualization using R.

II. BMI

BMI stands for Body Mass Index. It gives us the information about our weight category as per given in Table 2.1.The mathematical formula for the calculation of BMI is

$$BMI = Weight / (Height)^2$$

(Weight is in Kg and Height in m)

Table 2.1 [7]

BMI (Body Mass Index)	Result
Below 18.5	Underweight
18.5-24.9	Normal Weight
25.0-29.9	Over Weight
30.0-34.9	Obesity Class I
35.0-39.9	Obesity Class II
Above 40	Obesity Class III

III. DATA VISUALIZATION USING R [9]

The platform used in this paper is Jupyter notebook (Anaconda). It's an open source web application allows us to visualize data.
The libraries used for data importing and visualization are as under; read_csv() for comma separated values, read_tsv() for tab separated values, read_delim() for general delimited files, read_table() for tabular files where columns are separated by white space, read_log() for web log files.

- tidyverse
  - ggplot2
  - tibble
  - tidyr
  - readr

- read_csv()
- read_tsv()
- read_delim()
- read_fwf()
- read_table()
- read_log()

The data set used is

- BMI.csv

The command used to read csv file is read.csv & stored in the variable d1. After that using head we can display first 6 rows of data.

```
d1 <- read.csv("BMI.csv")
```

```
head(d1)
```

Age	Gender	Height	Weight
18	Male	175	48
21	Male	173	73
23	Male	170	85
19	Male	181	72
18	Male	163	61
22	Male	160	39

To check dimensions of the data we can use dim (d1).

```
dim(d1)
```

```
68  4
```

The above code represents 68 and 4, means data of 68 persons with 4 parameters (Age, Gender, Height and Weight).
The next step is calculation of BMI and addition of BMI column to the above data.

```
d2 <- mutate(d1,BMI=Weight/(Height/100)^2)
```

```
head(d2)
```

Age	Gender	Height	Weight	BMI
18	Male	175	48	15.67347
21	Male	173	73	24.39106
23	Male	170	85	29.41176
19	Male	181	72	21.97735
18	Male	163	61	22.95909
22	Male	160	39	15.23437

By using mutate we can add column to the existing data. Let the data is stored in the new variable “d2” now.

BMI=Weight/(Height/100)^2

Height is divided by 100 because it's required in meters and in collected data it was in centimeters.
Now using Table 2.1 conditional statements can be applied to display the result column.
We have to apply conditions on BMI column of data "d2". Let's save this in variable T : T <- d2$BMI
Now ifelse condition can be applied to implement the conditions given in Table 2.1

```
Result <- ifelse(T<18.5,"Under_Weight",
       ifelse(T>=18.5 & T<25,"Normal_Weight",
       ifelse(T>=25 & T<30,"Over_Weight",
       ifelse(T>=30 & T<35,"Obesity_Class_I",
       ifelse(T>=35 & T<40,
              "Obesity_Class_II","Obesity_Class_III")))))
```

Let d3 is the new variable to save the updated data, mutate can be used to add new column to the existing data

```
d3 <- mutate(d2,Result)
```

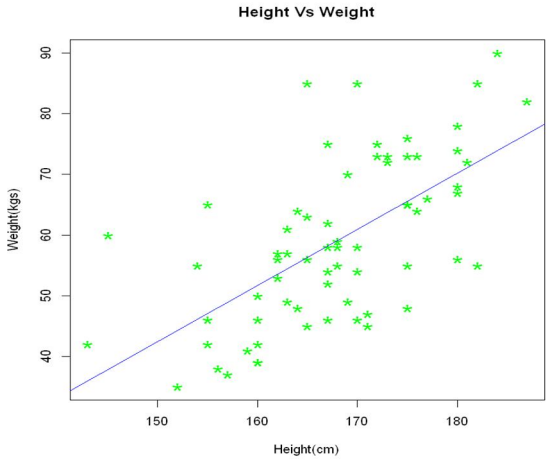
```
head(d3)
```

Age	Gender	Height	Weight	BMI	Result
18	Male	175	48	15.67347	Under_Weight
21	Male	173	73	24.39106	Normal_Weight
23	Male	170	85	29.41176	Over_Weight
19	Male	181	72	21.97735	Normal_Weight
18	Male	163	61	22.95909	Normal_Weight
22	Male	160	39	15.23437	Under_Weight

First Linear Regression
(Height Vs Weight)

```
lm1 <- lm(d3$Weight~d3$Height)
plot(d3$Height,d3$Weight,
     xlab="Height(cm)",
     ylab="Weight(kgs)",
     main="Height Vs Weight",
     pch="*",cex=2.1,col="green")
abline(lm1,col="blue")
```

In the above Linear Regression Model, Height is the explanatory variable (or the independent variable) and Weight is the response variable (or the dependent variable).



Summary [3] [2]

```
summary(lm1)
```

```
Call:
lm(formula = d3$Weight ~ d3$Height)

Residuals:
     Min       1Q   Median       3Q      Max
-17.5558  -7.8618  -0.5558   6.8785  28.6899

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -96.2453    23.5308  -4.090  0.00012 ***
d3$Height     0.9246     0.1399   6.608    8e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.47 on 66 degrees of freedom
Multiple R-squared:  0.3982,    Adjusted R-squared:  0.3891
F-statistic: 43.67 on 1 and 66 DF,  p-value: 8.001e-09
```

Weight = -96.2453 + Height*0.9226

```
#Calculation of Residual Standard Error
#for lm1 (First Model)
k1 = length(lm1$coefficients)-1
SSE1 = sum(lm1$residuals**2)
n1 = length(lm1$residuals)
RSE_lm1 = sqrt(SSE1/(n1-(1+k1)))
```

```
RSE_lm1
```

```
10.4733382337127
```

```
#Calculation of Multiple R-Squared
#for lm1 (First Model)
y1 = d1$Weight
SSyy1 = sum((y1-mean(y1))**2)
SSE1 = sum(lm1$residuals**2)
MRS1 = 1 - SSE1/SSyy1
```

```
MRS1
```

```
0.3982020691696
```

```
#Adjusted R-Squared
#for lm1 (First Model)
n1_1 = length(y1)
ARS1 = 1-(SSE1/SSyy1)*(n1_1-1)/(n1_1-(k1+1))
```

```
ARS1
```

```
0.389083918702473
```

```
# F Statistic
# for lm1 (First Model)
FS1 = ((SSyy1-SSE1)/k1) / (SSE1/(n1_1-(k1+1)))
```

```
FS1
```

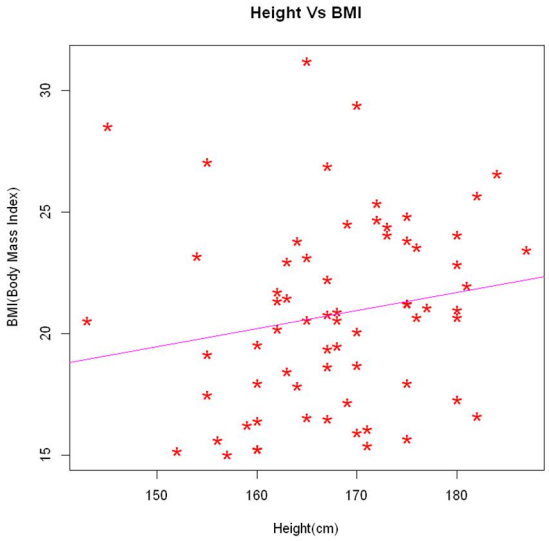
```
43.6713641220549
```

Second Linear Regression
(Height Vs BMI)

```
lm2 <- lm(d3$BMI~d3$Height)
plot(d3$Height,d3$BMI,
     xlab="Height(cm)",
     ylab="BMI(Body Mass Index)",
     main="Height Vs BMI",
     pch="*",cex=2.1,col="red")
abline(lm2,col="magenta")
```

In this model, Height is the explanatory variable (or the independent variable) and BMI (Body Mass Index) is the response variable (or the dependent variable).

- The regression line represents how much and in what direction dependent variable changes with respect to independent variable.
- The line closely approximates all the points.
- The purpose of regression line is make predictions.



Summary [3] [2]

```
summary(lm2)
```

```
Call:
lm(formula = d3$BMI ~ d3$Height)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6444 -2.8305 -0.1254  2.5222 10.6460

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.32208    8.47687   0.982    0.330
d3$Height    0.07426    0.05040   1.473    0.145

Residual standard error: 3.773 on 66 degrees of freedom
Multiple R-squared:  0.03184,   Adjusted R-squared:  0.01718
F-statistic: 2.171 on 1 and 66 DF,  p-value: 0.1454
```

BMI = Intercept + Height*Slope
BMI = 8.32208 + Height*0.07426
(by inserting any desired value of Height we can predict the value of BMI)

```
#Calculation of Residual Standard Error
#for lm2 (Second Model)
k2 = length(lm2$coefficients)-1
SSE2 = sum(lm2$residuals**2)
n2 = length(lm2$residuals)
RSE_lm2 = sqrt(SSE2/(n2-(1+k2)))
RSE_lm2
```

3.7729829693776

```
#Calculation of Multiple R-Squared
#for lm2 (Second Model)
y2 = d3$BMI
SSyy2 = sum((y2-mean(y2))**2)
SSE2 = sum(lm2$residuals**2)
MRS2 = 1 - SSE2/SSyy2
MRS2
```

0.0318449863237873

```
#Adjusted R-Squared
#for lm2 (Second Model)
n2_1 = length(y2)
ARS2 = 1-(SSE2/SSyy2)*(n2_1-1)/(n2_1-(k2+1))
ARS2
```

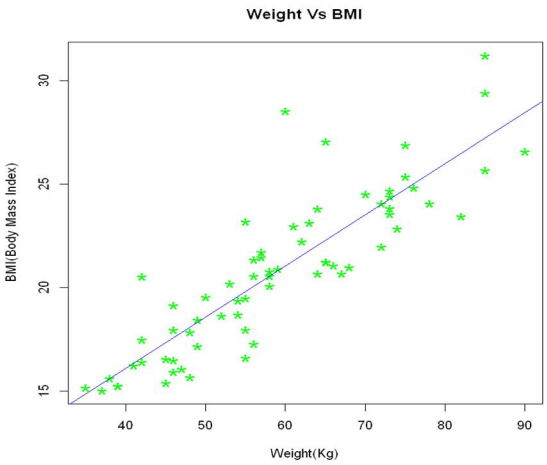
0.0171759709650567

```
# F Statistic
# for lm2 (Second Model)
FS2 = ((SSyy2-SSE2)/k2) / (SSE2/(n2_1-(k2+1)))
FS2
```

2.17090142351199

Third Linear Regression Model
(Weight Vs BMI)

```
lm3 <- lm(d3$BMI~d3$Weight)
plot(d3$Weight,d3$BMI,
     xlab="Weight(Kg)",
     ylab="BMI(Body Mass Index)",
     main="Weight Vs BMI",
     pch="*",cex=2.1,col="green")
abline(lm3,col="blue")
```



Weight is the explanatory variable (or the independent variable) and BMI (Body Mass Index) is the response variable (or the dependent variable).

Summary [3] [2]

```
summary(lm3)
```

```
Call:
lm(formula = d3$BMI ~ d3$Weight)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1928 -1.2081 -0.1813  0.8946  7.5037

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.19345    1.03988   5.956  1.1e-07 ***
d3$Weight    0.24734    0.01719  14.392  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.885 on 66 degrees of freedom
Multiple R-squared:  0.7584,    Adjusted R-squared:  0.7547
F-statistic: 207.1 on 1 and 66 DF,  p-value: < 2.2e-16
```

BMI = 6.19345 + Weight*0.24734
(by inserting any desired value of Weight we can predict the value of BMI)

```
#Calculation of Residual Standard Error
#for lm3 (Third Model)
k3 = length(lm3$coefficients)-1
SSE3 = sum(lm3$residuals**2)
n3 = length(lm3$residuals)
RSE_lm3 = sqrt(SSE3/(n3-(1+k3)))
RSE_lm3
```

1.88492433163864

```
#Calculation of Multiple R-Squared
#for lm3 (Third Model)
y3 = d3$BMI
SSyy3 = sum((y3-mean(y3))**2)
SSE3 = sum(lm3$residuals**2)
MRS3 = 1 - SSE3/SSyy3
MRS3
```

0.758363214172244

```
#Adjusted R-Squared
#for lm2 (Second Model)
n3_1 = length(y3)
ARS3 = 1-(SSE3/SSyy3)*(n3_1-1)/(n3_1-(k3+1))
ARS3
```

0.754702050750611

```
# F Statistic
# for lm3 (Third Model)
FS3 = ((SSyy3-SSE3)/k3) / (SSE3/(n3_1-(k3+1)))
FS3
```

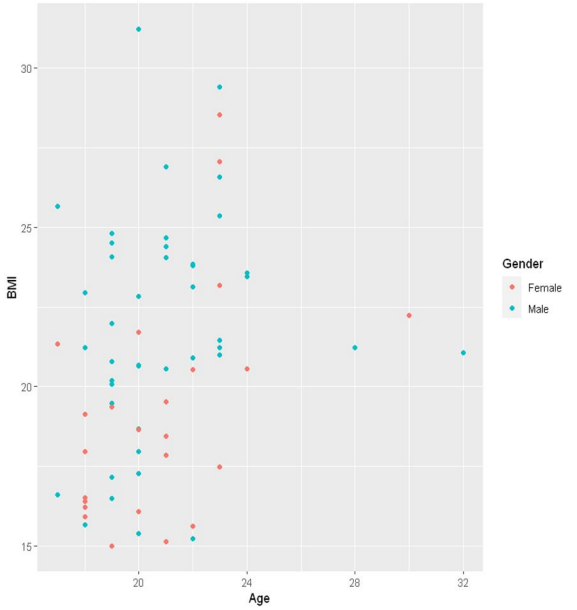
207.137220286676

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

IV. DATA VISUALIZATION USING GGPLOT() & GEOM():

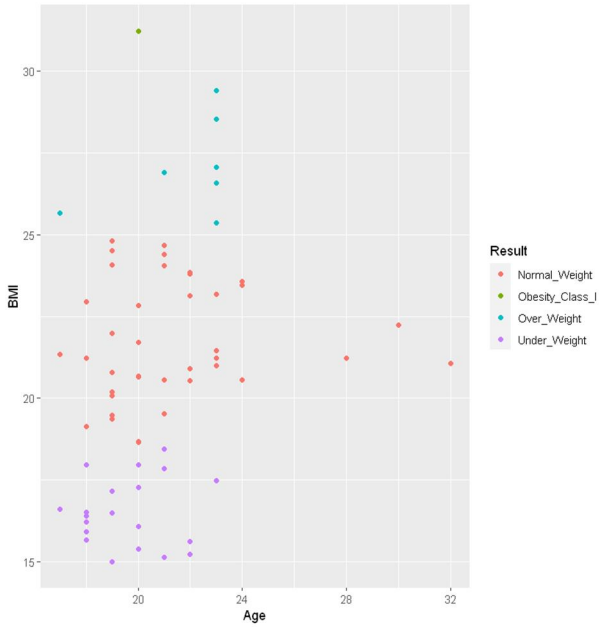
1) Plot 1: Age Vs BMI (Body Mass Index)

```
plot1 <- ggplot(data=d3,aes(x=Age,y=BMI))
plot1+geom_point(aes(color=Gender))
```



2) Plot 2: Age Vs BMI
Keeping in view the Result (Weight factor)

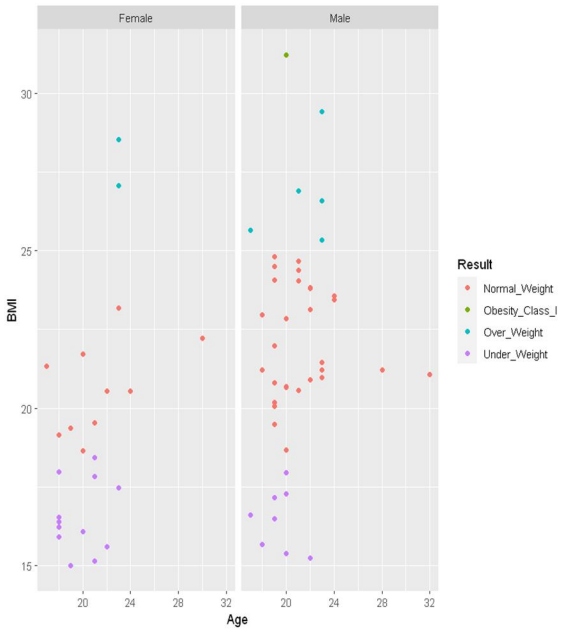
```
plot1+geom_point(aes(color=Result))
```



3) Plot 3: Age Vs BMI

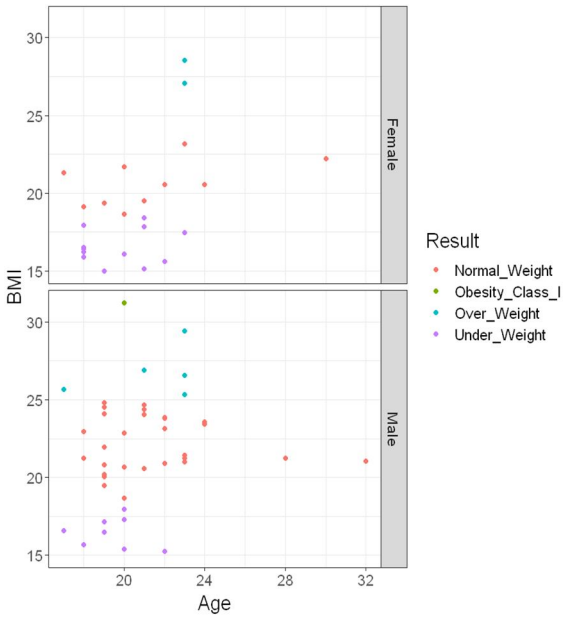
Showing separately the data of Male & Female

```
plot1 + geom_point(aes(color=Result))
      + facet_wrap(facets=vars(Gender))
```



4) Plot 4: Horizontal view of Plot 3

```
plot1 + geom_point(aes(color=Result))
      + facet_grid(rows=vars(Gender))
      + theme_bw()
      + theme(text=element_text(size=16))
```



V. RESULTS

The results of above regression models is [1]

1) Weight = -96.2453 + Height*0.9226
2) BMI = 8.32208 + Height*0.07426
3) BMI = 6.19345 + Weight*0.24734

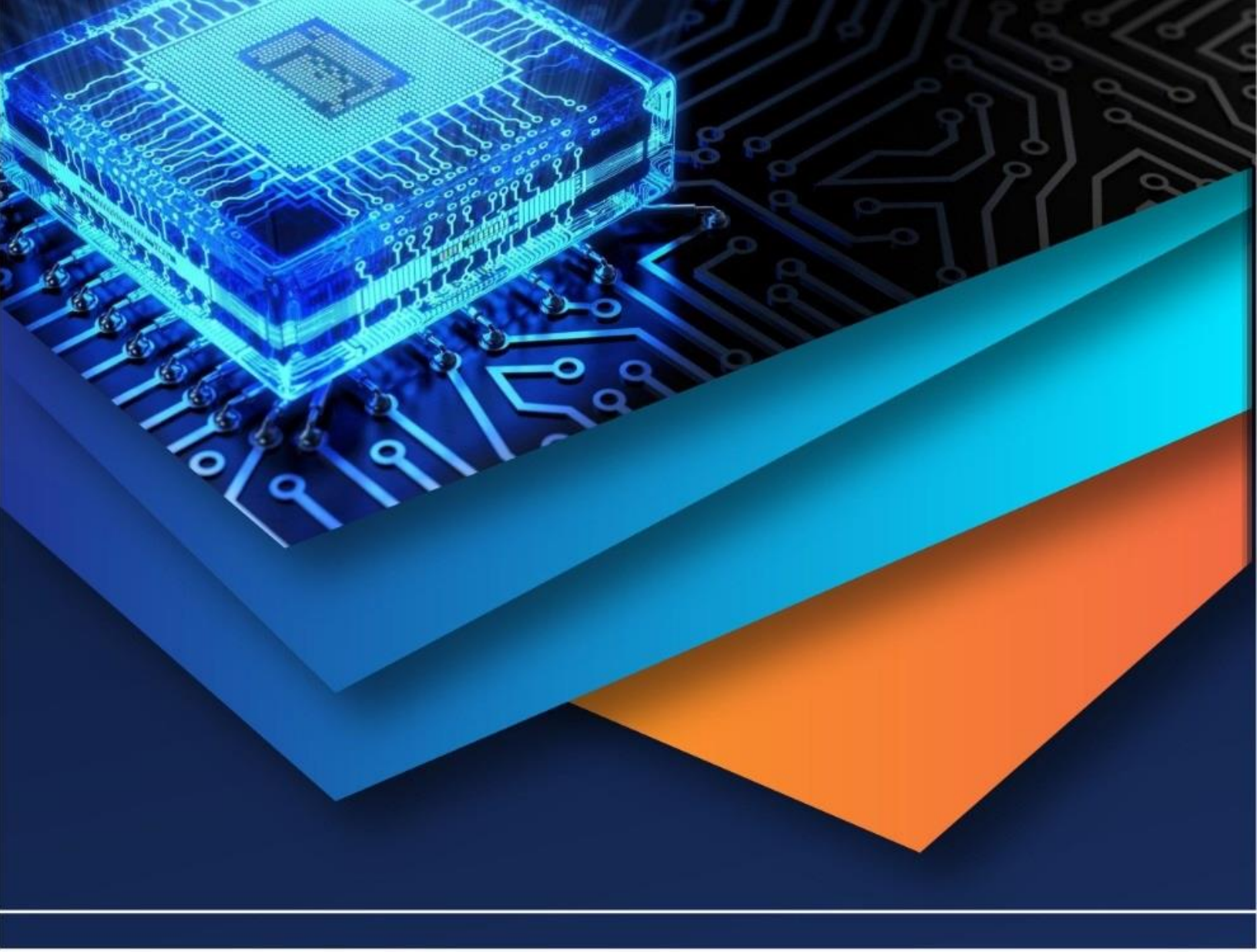
A. Explanation of Summary

1) Call is the feature in R that represents what function & parameters were used to create the model[2]
2) Residuals represents the difference between observed data of the dependent variable (y) and the fitted values($\hat{y}$) $\hat{y} = a + bx$, where a is y intercept, b is slope of the line and x is independent variable [1]
3) In Coefficients four parts are there[2]
- Estimate : gives us intercept and slope regression line
- Std Error : RSE/sq root of sum of squares of x variable
- t value : Estimate/SE
- Pr(>|t|) : Probability of occurrence of t-value

4) Calculation of Residual Standard Error, Multiple R-Squared, Adjusted R-Squared & F-Statistic for each model.
5) In Plot 4
- Count of Females with Normal Weight are less than that of Males
- No female is there in obese category
- Overweight male candidates are more than those of female candidates

REFERENCES

[1] https://www.learnbymarketing.com/tutorials/explaining-the-lm-summary-in-r/
[2] https://www.learnbymarketing.com/tutorials/explaining-the-lm-summary-in-r/
[3] Chan YH. Biostatistics 201: Linear regression analysis. Age (years). Singapore Med J 2004;45:55-61.
[4] Gaddis ML, Gaddis GM. Introduction to biostatistics: Part 6, correlation and regression. Ann Emerg Med 1990;19:1462-8.
[5] Elazar JP. Multiple Regression in Behavioral Research: Explanation and Prediction. 2nd ed. New York: Holt, Rinehart and Winston; 1982.
[6] Schneider A, Hommel G, Blettner M. Linear regression analysis: Part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010;107:776-82.
[7] https://www.ncbi.nlm.nih.gov/books/NBK535456/figure/article-18425.image.f1/
[8] https://www.youtube.com/watch?v=XAnilMY-ILs&list=PLpApktzwiFX9UZk5ZijcDuTa9q9MLgWZD
[9] https://cran.r-project.org/web/packages/readr/readme/README.html



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY