International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 4 (2019) pp. 965-976

Application of Linear Regression Modeling on Continuous and Categorical Data using R Programming Scripts

Anurag Singh Bisht* and Umang Soni,

Netaji Subhas university of Technology New Delhi-110078, India.
** Corresponding author*

Abstract

The Science of converting data into meaningful insights has come to known as data science. Data is being generated rapidly all around the sources. Most of this data is in raw format and is of limited use or almost of no use but interpreting and using insights from this data can definitely do some magic for the growth of organizations as well as for businesses. As part of this paper anthropometric data stuff (mainly human height, weight along with gender, age) have been used to forecast the response variable with the application of linear regression model using the analytical tool R studio.

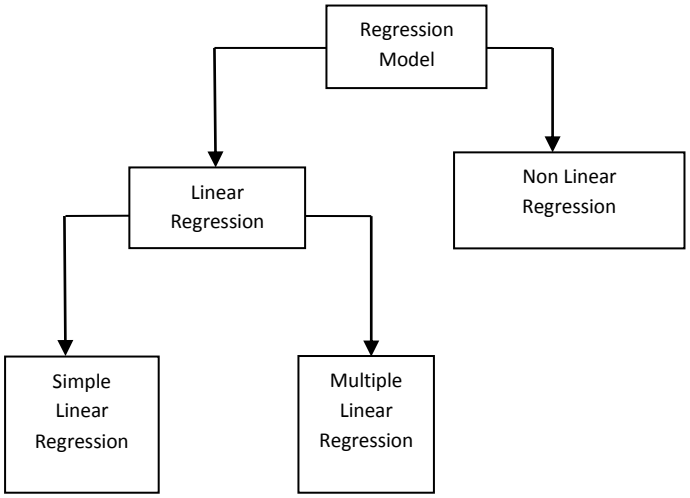
In Today's scenario data science and analytics has become one of the most recent skill and path way that can transform life, business and global economy. Data analytics is the assessment of data with the intent to depict a conclusion from the information. It includes the use of data stuff, technology and statistical tools, methods and modeling. By performing the linear regression in the form of simple linear and multiple linear regression, a well defined trend equation has been established in both (simple and multiple) cases and outcome has been predicted. Apart from this dynamic values of input have been provided in modeling and the outcomes have been received in R programming script.

Keywords: Linear regression, Rstudio, SPSS, excel, correlation, package

1. INTRODUCTION TO REGRESSION MODEL

Linear regression is a statistical measure that attempts to determine the strengths of the relationship between one dependent variable (usually denoted by Y) and a series of the other changing variable (Known as independent variable).

Regression is a technique used to calculate the cause and effect relationship among variables. Thus it is defined in a manner that how strongly the independent and dependent variables are associated with each other. It can be applicable through tools like SPSS and excel [2].



2. TYPES OF REGRESSION ANALYSIS [3]

- **Simple Linear Regression(SLR):** Single explanatory variable that contains only one independent variable X and association between X and Y (dependent variable) is explained by a linear function. In SLR variation in Y are assumed to be related to variation in X.It can be denoted by equation: $Y = c + mX$
- **Multiple Linear Regression(MLR):** Explanatory variable in which two or more independent variables i.e. X's, here association between X's and Y is explained by a linear function and variation in Y are assumed to be related to variation in X. It can be expressed with equation: $Y = c + m1*X1 + m2*X2 + ... + mp*Xp$
- **Non-linear:** It implies curved relationships, for example exponential relationships.

3. TERMINOLOGY IN LINEAR REGRESSION [3]

- **Dependent variable:** It is the variable that depends on other variables. It can also be called Measured variable or Explained variable or Response variable.
- **Independent variable:** Variable(s) that need to predict the response or dependent variable called as Explanatory variable(s) or Manipulated variable(s) or

Controlled variable(s) or Predictor variable(s). It is defined as the known variable.

- **Coefficients:** The estimate of magnitude of impact of changes in the predictor(s) on the predicted variable. It can be denoted by slope of line. It represents that how much each unit change of predictor variable X changes the response variable Y.
- **Intercept:** It is denoted the position where regression line cuts the Y axis.

As part of this paper regression is applied in the R language, which is suitable for ease in data analysis. Although it can be applied on other available editors of R programming (i.e. Rstudio, Eclipse StatET, Rcommander, Rattle, Tinn-R).But in this paper we have used the Rstudio, which is a code editor and development environment with nice features that make code development in R easy and simple. As Rstudio is available free of charge for Linux, Windows and other operating system so it is a good option to use the linear regression with R [2].

4. PROBLEM STATEMENT AND OBJECTIVE

Main objective of this paper is first to predict the dependent variable (Weight) by considering only one independent variable (Height) by implementing simple linear regression model. Later on other independent variables i.e. Age, Gender (along with Height) have been incorporated to predict the Weight by multiple linear regression model. Here Age and Height falls under continuous data and Gender is categorical data.

For analytics point of view data has been taken from India, NFHS-3, 2005–06 shown in appendix1.Analytics has been performed using R Studio which includes the scripts in R programming language. R studio has facility to import data in different format. Here data (from appendix1) is imported in CSV (Comma delimited) format. In Rstudio for simple linear regression SLR.CSV (Male height in meter as independent variable and male weight in kilogram as dependent variable) and for multiple linear regression MLR.CSV (all data of appendix1) is used.

In R studio, R script & data window is used to run R commands and console window is used to view the output [4].

In R to run a command a bundle of code is essential to install and load, this bundle is known as package. Packages are installed in a repository so that they can be installed. Few popular repositories for R package are CRAN, Github. Example of few packages are “dplyr”,”ggplot2.Package in R studio can be installed and loaded as below[5].

install.packages("dplyr",dependencies=T) {To install the package which is a one time activity}

library(dplyr) {To load the package for further use, need to load in every new session of R}

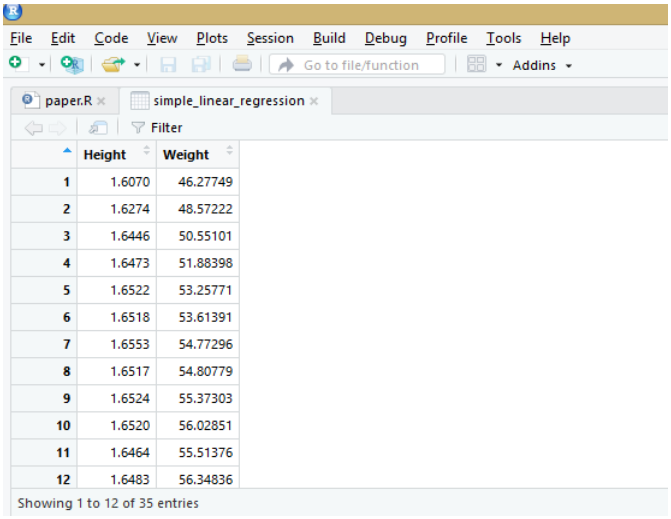
5. ANALYSIS OF SIMPLE LINEAR REGRESSION:

Analysis of SLR.csv data through simple linear regression has been performed between Height (independent variable) and Weight (dependent variable).

Following script need to run on R studio.

simple_linear_regression<-read.csv("SLR.csv") {Import the data file into self created object simple_linear_regression with the help of function read.csv()}

View(simple_linear_regression) {View the created object, it will provide the below output}



	Height	Weight
1	1.6070	46.27749
2	1.6274	48.57222
3	1.6446	50.55101
4	1.6473	51.88398
5	1.6522	53.25771
6	1.6518	53.61391
7	1.6553	54.77296
8	1.6517	54.80779
9	1.6524	55.37303
10	1.6520	56.02851
11	1.6464	55.51376
12	1.6483	56.34836

Showing 1 to 12 of 35 entries

summary(simple_linear_regression) {To see the statistical summary of SLR.csv, which provides mean median and other statistical details of dependent and independent variables}

```
> summary(simple_linear_regression)
     Height          Weight
 Min.   :1.607   Min.   :46.28
 1st Q. :1.643   1st Q. :55.44
 Median :1.647   Median :57.19
 Mean   :1.645   Mean   :56.74
 3rd Q. :1.651   3rd Q.  :59.24
 Maximum:1.655   Maximum :60.27
```

str(simple_linear_regression) {To examine the structure of SLR data}

```
> str(simple_linear_regression)
'data.frame':  35 obs. of  2 variables:
 $ Height: num  1.61 1.63 1.64 1.65 1.65 ...
 $ Weight: num  46.3 48.6 50.6 51.9 53.3 ...
```

Coefficient of correlation- It is the measure of strong relation between variables, it’s value lies between -1 to +1. It is also known as Pearson’s coefficient. In R It can be calculated as below.

cor(simple_linear_regression$Weight,simple_linear_regression$Height) {To check the correlation between independent and independent variables and as per the below output it is 46.35166 percent}

```
> cor(simple_linear_regression$Weight,simple_linear_regression$Height)
[1] 0.4635166
```

6. MODELING OF SIMPLE LINEAR REGRESSION

In R programming, function lm() is used for modeling in below manner.

lm(dependent variable~independent variable,data=source of data)

So as per the SLR data below command on R studio.

simple_linear_regression_model<-lm(Weight~Height,data=simple_linear_regression) {Model will be created in object simple_linear_regression_model}

```
> simple_linear_regression_model<-lm(Weight~Height,data=simple_linear_regression)
```

simple_linear_regression_model {To see the model,run the object simple_linear_regression_model, which provides the coefficients for forming the simple linear regression.}

```
> simple_linear_regression_model
Call:
lm(formula = Weight ~ Height, data = simple_linear_regression)
Coefficients:
(Intercept)       Height
     -233.4        176.3
```

Thus from above output we can write down the SLR equation as below.

Weight=-233.4+176.3×Height

summary(simple_linear_regression_model)

```
> summary(simple_linear_regression_model)
```
{To see the statistical summary of SLR model}

```
Call:
lm(formula = Weight ~ Height, data = simple_linear_regression)
Residuals:
    Min      1Q  Median      3Q     Max
-6.1200 -2.3105  0.8077  2.5283  4.1624
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -233.35      96.54  -2.417  0.02133
Height        176.35      58.68   3.005  0.00504

Multiple R-squared=0.2148,Adjusted R-squared=0.1911
P value=0.005042
```

From above summary we can see that p value of model is .005042 which is less than .05(as we are considering the 95 percent of confidence level). So Null hypothesis will reject and there will be an effect of independent variable (Height) on dependent variable (Weight).

Multiple R^2 is also known as coefficient of determination. For this model R^2 between dependent and independent variables (i.e. Weight and Height) is 0.2148. It shows that as per the available data only 21.48% of the influence in dependent variable is due to independent variable and rest of the percentage is due to other factors and two such additional factors (Gender, Age) can be explained as part of multiple linear regression.

simple_linear_regression_model$fitted.values {Predicted values are the fitted value and can be calculated using this script}

```
> simple_linear_regression_model$fitted.values
       1        2        3        4        5        6        7        8        9
50.04035 53.63784 56.67103 57.14717 58.01127 57.94073 58.55795 57.92310 58.04654
      10       11       12       13       14       15       16       17       18
57.97600 56.98845 57.32352 57.83492 56.91792 58.16999 56.63576 58.02891 57.65858
      19       20       21       22       23       24       25       26       27
57.69385 57.48223 55.73638 57.37642 57.39405 56.37124 57.28825 55.59531 57.12953
      28       29       30       31       32       33       34       35
56.65339 56.42414 56.35360 54.76647 56.70630 56.10672 55.70111 55.73638
```

simple_linear_regression_model$residuals {Residuals are the difference of actual value and predicted values. Residuals of the SLR can be calculated using this script.}

```
> simple_linear_regression_model$residuals
          1           2           3           4           5           6           7
-3.76286054 -5.06562401 -6.12001415 -5.26318754 -4.75356049 -4.32682369 -3.78498945
          8           9          10          11          12          13          14
-3.11531060 -2.67350770 -1.94749782 -1.47469158 -0.97515710 -0.85188050  0.27574533
         15          16          17          18          19          20          21
 0.09202284 -0.17516636  0.53163666  0.80766937  1.49476581  1.64754154  0.99261340
         22          23          24          25          26          27          28
 1.92776435  2.54251295  2.46958806  2.82200373  1.15887673  2.86081901  2.59957773
         29          30          31          32          33          34          35
 3.27520771  3.07379423  1.92951300  2.51418658  4.16240388  3.08276116  4.02926748
```

simple_linear_regression_model$rank {In R it provides the total number of dependent and independent variables}

```
> simple_linear_regression_model$rank
[1] 2
```

simple_linear_regression_model$coefficients {Other way to find out the coefficient of the model}

```
> simple_linear_regression_model$coefficients
(Intercept)      Height
  -233.3508    176.3479
```

new<-data.frame(Height=c(1.6518,1.6553,1.6517)) {To predict the Weight dynamically say we have three patient of height 1.6518,1.6553,1.6517 respectively. Object new has been provided with above input}

```
> new<-data.frame(Height=c(1.6518,1.6553,1.6517))
```

predict(simple_linear_regression_model,new){Prediction of respective weights can be calculated with this script}

```
> predict(simple_linear_regression_model,new)
       1        2        3
57.94073 58.55795 57.92310
```

As in SLR it was observed that only 21.48% of the influence in dependent variable is due to independent one, so in Multiple linear regression require more than one independent variables. Here in this data(MLR.CSV) three independent variables Age, Height (both continuous) and Gender(categorical in nature and is of two form only i.e. Male and Female) has been incorporated.

7. ANALYSIS OF MULTIPLE LINEAR REGRESSION

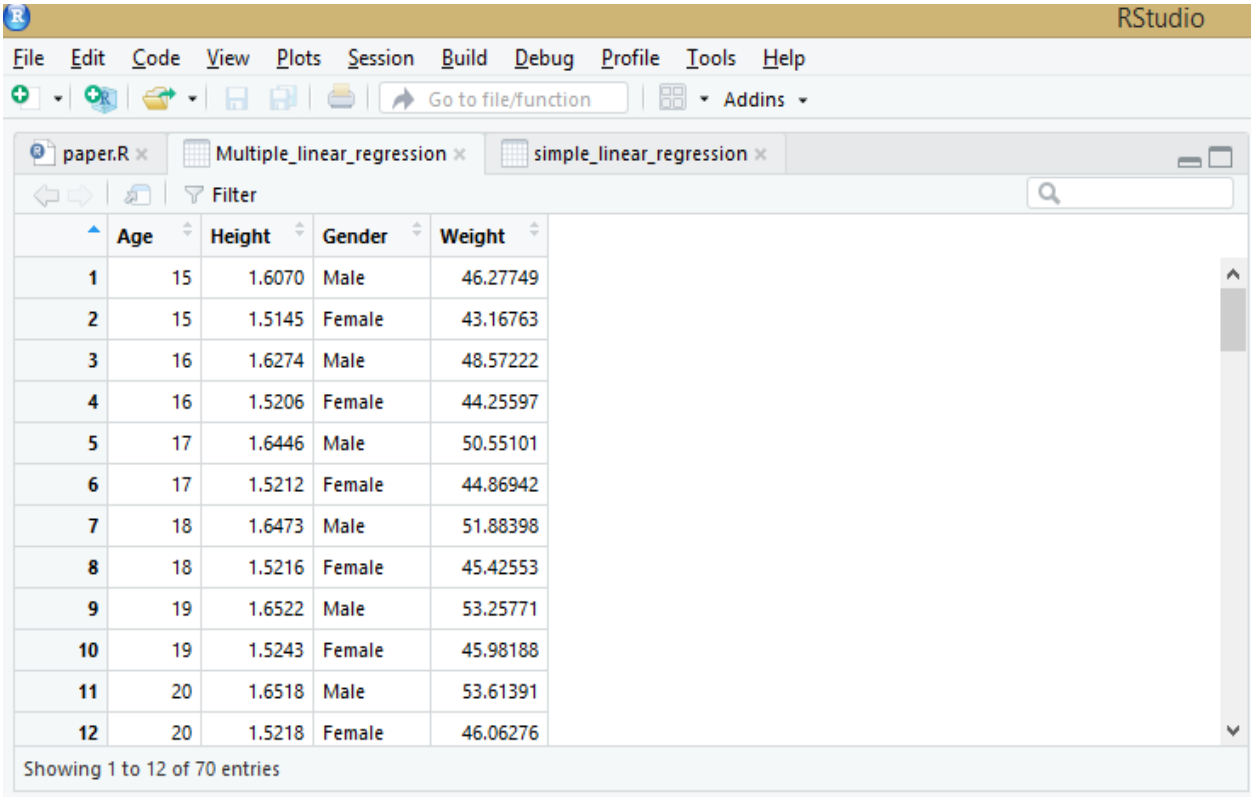
Analysis of MLR.csv data through multiple linear regression has been established between height (independent variable) and weight, Age, Gender (dependent variable).

Following script need to run on R studio.

Multiple_linear_regression<-read.csv("MLR.csv") {Import the data file into self created object Multiple_linear_regression with the help of function read.csv()}

```
> Multiple_linear_regression<-read.csv("MLR.csv")
```

View(Multiple_linear_regression) {View the newly created object, it will provide the below output}



	Age	Height	Gender	Weight
1	15	1.6070	Male	46.27749
2	15	1.5145	Female	43.16763
3	16	1.6274	Male	48.57222
4	16	1.5206	Female	44.25597
5	17	1.6446	Male	50.55101
6	17	1.5212	Female	44.86942
7	18	1.6473	Male	51.88398
8	18	1.5216	Female	45.42553
9	19	1.6522	Male	53.25771
10	19	1.5243	Female	45.98188
11	20	1.6518	Male	53.61391
12	20	1.5218	Female	46.06276

Showing 1 to 12 of 70 entries

summary(Multiple_linear_regression) {To see the statistical summary of MLR data, which provides mean median and other statistical details of dependent and independent variables }

```
> summary(Multiple_linear_regression)
      Age            Height          Gender       Weight
 Min.   :15.00   Min.   :1.514   Female:35   Min.   :43.17
 1st Q.  :23.25  1st Q.  :1.522    Male  :35  1st Q.  :49.33
 Median :32.00   Median :1.566               Median :52.63
 Mean   :32.00   Mean   :1.583               Mean   :53.12
 3rd Q.  :40.75   3rd Q.:1.647                3rd Q.  :57.14
 Maximum :49.00   Maximum:1.655               Maximum:60.27
```

str(Multiple_linear_regression) {To examine the structure of MLR data, which shows Gender has levels as Female and Male}

```
> str(Multiple_linear_regression)
'data.frame': 70 obs. of  4 variables:
 $ Age   : int  15 15 16 16 17 17 18 18 19 19 ...
 $ Height: num  1.61 1.51 1.63 1.52 1.64 ...
 $ Gender: Factor w/ 2 levels "Female","Male": 2 1 2 1 2 1 2 1 2 1 ...
 $ Weight: num  46.3 43.2 48.6 44.3 50.6 ...
```

Multiple_linear_regression$Gender<-factor(Multiple_linear_regression$Gender,labels = c(0,1)) {As independent variable Gender is categorical in nature so for modeling point of view factorization is essential, which means a mathematical factor need to assigned to two components of independent variable Gender. Here Factor() function in R is used to convert the components Male and Female, into factors 0 and 1. In R,assignment of factor 0 is allotted to Female and 1 to Male, based on alphabetic order.)

```
> Multiple_linear_regression$Gender<-factor(Multiple_linear_regression$Gender,labels = c(0,1))
```

str(Multiple_linear_regression) {To examine the structure of SLR data, which shows Gender has levels converted to 0 and 1}

```
> str(Multiple_linear_regression)
'data.frame':     70 obs. of  4 variables:
 $ Age   : int  15 15 16 16 17 17 18 18 19 19 ...
 $ Height: num  1.61 1.51 1.63 1.52 1.64 ...
 $ Gender: Factor w/ 2 levels "0","1": 2 1 2 1 2 1 2 1 2 1 ...
 $ Weight: num  46.3 43.2 48.6 44.3 50.6 ...
```

Before modeling there must be a correlation between independent variable and dependent variable. In MLR two independent variable are continuous (Height and Age) so their correlation with dependent variable (Weight) can be calculated as per the below function in R. ***cor(Multiple_linear_regression$Age,Multiple_linear_regression$Weight)***

```
> cor(Multiple_linear_regression$Age,Multiple_linear_regression$Weight)
[1] 0.5776672
```

cor(Multiple_linear_regression$Height,Multiple_linear_regression$Weight)

```
> cor(Multiple_linear_regression$Height,Multiple_linear_regression$Weight)
[1] 0.7754721
```

Independent variable Gender is categorical in nature hence correlation cannot be calculated as variable must be numeric so it will give the error as per the below.

cor(Multiple_linear_regression$Gender,Multiple_linear_regression$Weight)

```
> cor(Multiple_linear_regression$Gender,Multiple_linear_regression$Weight)
Error in cor(Multiple_linear_regression$Gender, Multiple_linear_regression$Weight) :
  'x' must be numeric
```

8. MODELING OF MULTIPLE LINEAR REGRESSION

It requires that there must be little or no multi-collinearity, which occurs when there are high correlations between two or more predicted variables. For good model multi-collinearity should be less.In R it can be checked using variance inflation factor(VIF).For VIF package “usdm” should be installed and loaded.

install.packages("usdm",dependencies=T){To install the package}

library(usdm) {To load the package}

```
> library(usdm)
Loading required package: sp
Loading required package: raster
Warning messages:
1: package 'usdm' was built under R version 3.5.2
2: package 'sp' was built under R version 3.5.2
3: package 'raster' was built under R version 3.5.2
```

VIF of the independent variable can be calculated using vifstep()

vifstep(Multiple_linear_regression[,-c(3,4)], th = 4) {as per the MLR.CSV data, column 3 and 4 are Gender(categorical data) and Weight(dependent variable) so they have been excluded, hence multi-collinearity check is performed on Height and Age.Result shows that there is no issue of multi-collineaarity}

```
> vifstep(Multiple_linear_regression[,-c(3,4)], th = 4)
No variable from the 2 input variables has collinearity problem.
The linear correlation coefficients ranges between:
min correlation ( Height ~ Age ):  -0.01396793
max correlation ( Height ~ Age ):  -0.01396793
---------- VIFs of the remained variables --------
  Variables      VIF
1       Age 1.000195
2    Height 1.000195
```

Multiple_linear_regression_model<-lm(Weight~Height+Gender+Age,data=Multiple_linear_regression) {More than one independent variables can be included by "+" symbol and Model will be created in object Multiple_linear_regression_model}

```
> Multiple_linear_regression_model<-
lm(Weight~Height+Gender+Age,data=Multiple_linear_regression)
```

Multiple_linear_regression_model<-lm(Weight~.,data=Multiple_linear_regression){Other way to include all independent variable is by the use of "."

```
> Multiple_linear_regression_model<-lm(Weight~.,data=Multiple_linear_regression)
```

Multiple_linear_regression_model {To see the model,run the object Multiple_linear_regression_model, which provides the coefficients used for forming the multiple linear regression.}

```
> Multiple_linear_regression_model
Call:
lm(formula = Weight ~ ., data = Multiple_linear_regression)
Coefficients:
(Intercept)          Age       Height       Gender
  -245.7354       0.2883     188.0287     -16.0566
```

Thus from above output we can write down the MLR equation as below.

Weight=-245.7354+.2883×Age+188.0287×Height-16.0566×Gender

summary(Multiple_linear_regression_model) {To see the statistical summary of MLR model}

```
> summary(Multiple_linear_regression_model)
Call:
lm(formula = Weight ~ ., data = Multiple_linear_regression)
Residuals:
     Min       1Q   Median       3Q      Max
-1.79042 -0.35711  0.07132  0.44933  1.58257
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.457e+02  2.027e+01 -12.125  < 2e-16
Age          2.883e-01  8.591e-03  33.560  < 2e-16
Height       1.880e+02  1.330e+01  14.139  < 2e-16
Gender1     -1.606e+01  1.657e+00  -9.691 2.62e-14

Multiple R squared=0.9784,Adjusted R squared=0.9775
P value < 2.2e-16
```

As per the above summary we can see that p value of model is less than 2.2×10^{-16} which is less than .05(as we are considering the 95 percent of confidence level). So Null hypothesis will reject and there will be an effect of independent variables on dependent variable.

Multiple R-squared (coefficient of determination) is .9784 (in SLR it was 0.2148). This means that as per the available data 97.84 % of the variance in Weight is due to independent variables.

The adjusted R^2 is a modified version of R^2 that has been adjusted for the number of predictors in the model. As the number of independent variables increases then R^2 value increases so adjusted R^2 increases only if the new term improves the model more than would be expected by chance. It is always lower than the R^2 and result shows that its value is 97.75 percent.

Multiple_linear_regression_model$fitted.values {Predicted values are the fitted value and can be calculated using this script}

```
> Multiple_linear_regression_model$fitted.values
       1        2        3        4        5        6        7        8        9
44.69492 43.35885 48.81902 44.79414 52.34143 45.19528 53.13743 45.55881 54.34709
      10       11       12       13       14       15       16       17       18
46.35481 54.56020 46.17305 55.50662 47.06306 55.11803 47.14455 55.53797 47.41407
      19       20       21       22       23       24       25       26       27
55.75108 47.62718 54.98644 47.76507 55.63201 48.20382 56.46562 48.83059 55.77619
      28       29       30       31       32       33       34       35       36
48.57362 57.39951 48.46708 56.05198 49.01864 57.82572 49.64541 57.71918 50.06535
      37       38       39       40       41       42       43       44       45
58.04511 50.35367 58.10779 50.32234 56.53463 50.14059 58.57161 50.78616 58.87874
      46       47       48       49       50       51       52       53       54
51.24371 58.07649 51.32519 59.34256 51.12464 57.82580 51.73261 59.74997 52.07733
      55       56       57       58       59       60       61       62       63
59.53061 52.23403 59.57449 52.59756 59.78760 52.88588 58.38366 52.61012 60.74030
      64       65       66       67       68       69       70
52.84203 60.38932 52.77309 60.24517 53.24944 60.57110 52.97367
```

Multiple_linear_regression_model$residuals {Residuals are the difference of actual value and predicted values. Residuals of the MLR can be calculated using this script.}

```
> Multiple_linear_regression_model$residuals
          1           2           3           4           5           6           7
 1.58257074 -0.19122078 -0.24680043 -0.53816791 -1.79041967 -0.32585999 -1.25345049
          8           9          10          11          12          13          14
-0.13327946 -1.08937864 -0.37292937 -0.94628875 -0.11029508 -0.73365658 -0.45753976
         15          16          17          18          19          20          21
-0.31024608 -0.46690109 -0.16493898 -0.34780904  0.27742398  0.04097835  0.52732349
         22          23          24          25          26          27          28
-0.05423736  0.71634519  0.06773423  0.51742821 -0.11922423  1.41747499  0.34663488
         29          30          31          32          33          34          35
 0.86249953  0.71117026  0.40861410 -0.05075255  0.73481951  0.13462392  0.74706267
         36          37          38          39          40          41          42
 0.08564610  1.14350371  0.07607256  1.02197716  0.45850532  0.19436976  0.26570726
         43          44          45          46          47          48          49
 0.73256997  0.67258633  1.05783124  0.22955285  0.76433578  0.16633405  0.76769202
         50          51          52          53          54          55          56
 0.42180594 -1.07161849 -0.37134911  0.24038154  0.57658242 -0.27764081  0.09407185
         57          58          59          60          61          62          63
 0.12485498  0.01245854 -0.36020543  0.27914614 -1.68767864 -0.46084003 -1.51981317
         64          65          66          67          68          69          70
-0.34496462 -0.12020037  0.07490094 -1.46129648 -0.69964427 -0.80544557  0.33050271
```

Multiple_linear_regression_model$coefficients {Other way to find out the coefficients of the model}

```
> Multiple_linear_regression_model$coefficients
 (Intercept)          Age       Height      Gender1
-245.7354444    0.2883192  188.0287246  -16.0565894
```

Multiple_linear_regression_model$rank {In R it provides the total number of dependent and independent variables}

```
> Multiple_linear_regression_model$rank
[1] 4
```

new<-data.frame(Age=c(35,25,45),Height=c(1.23,1.6,1.637),Gender=as.factor(c(0,1,1))) {To predict the Weight dynamically say we have three players of Age 35,25,45 with height as 1.23,1.6,1.637 meter and Gender 0,1,1 (i.e. female, Male,Male) respectively. Object new has been provided with above input}

```
> new<-
data.frame(Age=c(35,25,45),Height=c(1.53,1.62,1.652),Gender=as.factor(c(0,1,1)))
```

predict(Multiple_linear_regression_model, new) {Prediction of respective weights can be calculated with this

script}

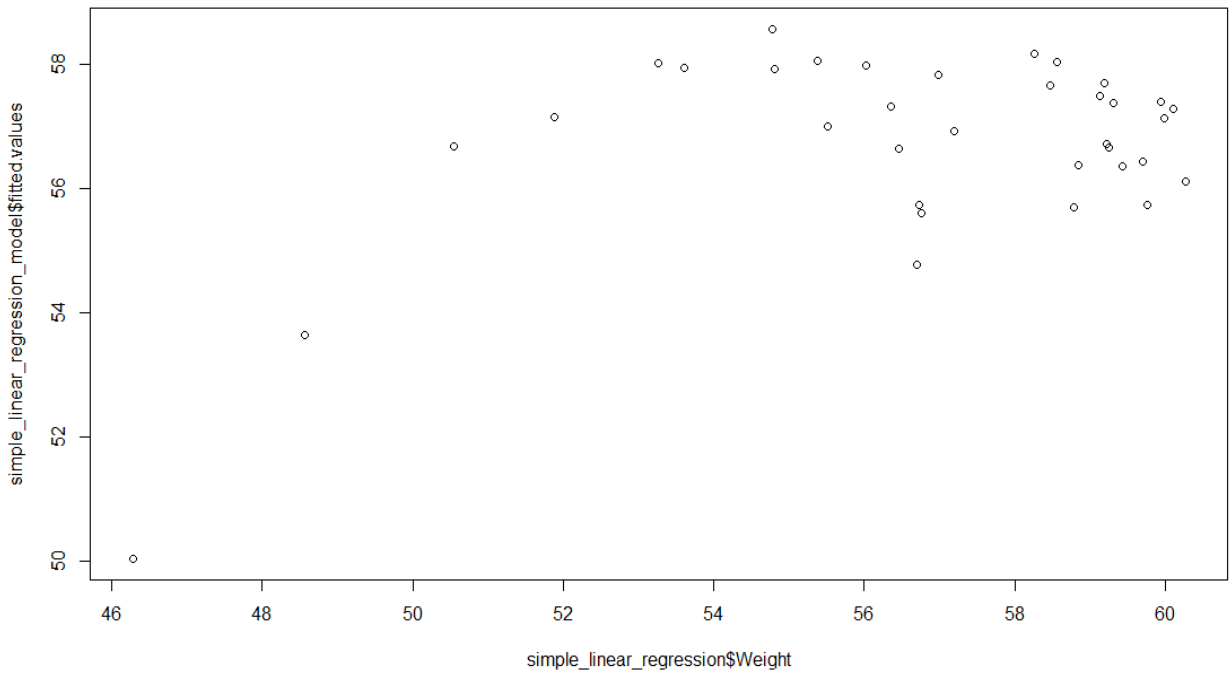
```
> predict(Multiple_linear_regression_model, new)
    1     2     3
52.03968 50.02248 61.80579
```

9. VALIDATION OF REGRESSION MODELS

Fitted value (predicted value) in association with actual value is used to validate a model**.** In a good model, the differences between actual and fitted value should be very less and close to each other **.**A good model has higher number of overlap**.**

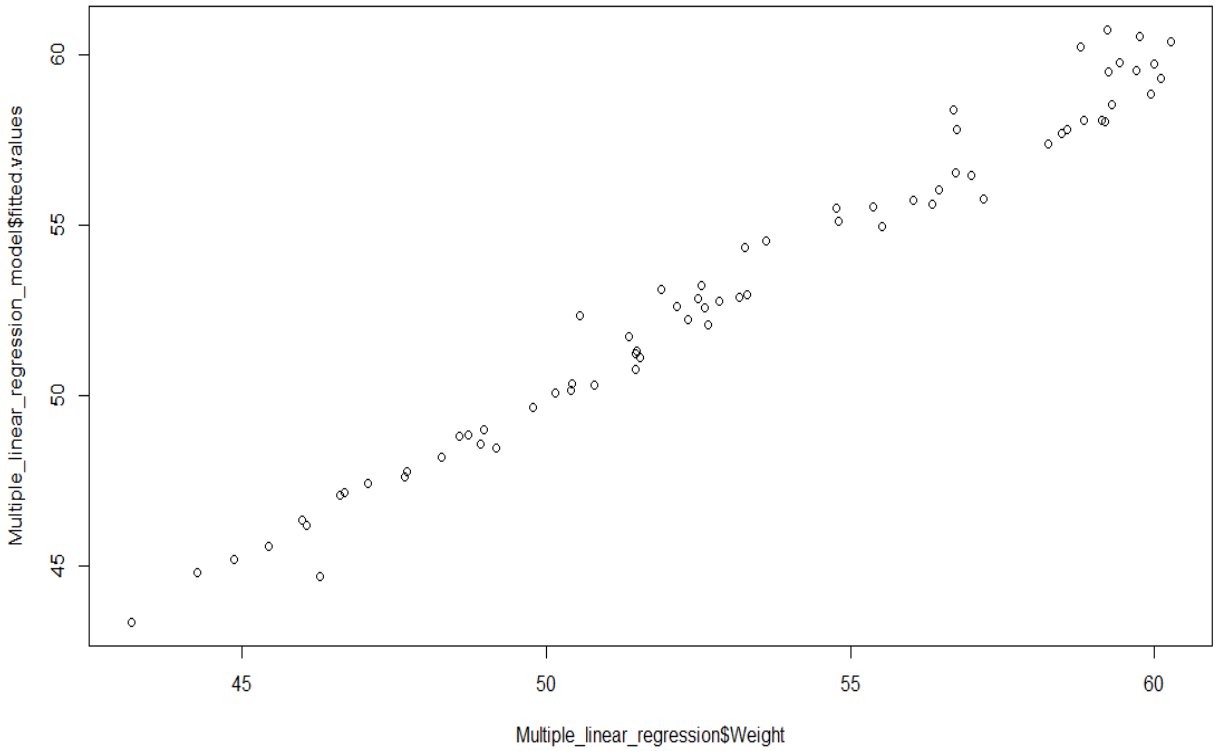
In SLR simple_linear_regression_model$fitted.values and simple_linear_regression$Weight gives the predicted and actual value respectively.Plot bteween them in R can be created as below

plot(simple_linear_regression$Weight,simple_linear_regression_model$fitted.values)



Like above in MLR Multiple_linear_regression_model$fitted.values and Multiple_linear_regression$Weight gives the predicted and actual values and script in R for their relation is as below.

plot(Multiple_linear_regression$Weight,Multiple_linear_regression_model$fitted.values)



As per our model in linear regression the actual weight and predicted weight are overlapping significantly which shows a strong validation of results (More overlapping in MLR, hence better result in MLR as compare to SLR)

10. CONCLUSION

As part of this paper in first analysis (simple linear regression) through Rstudio, weight was treated as the dependent and height as independent variable. As per the result obtained from analysis, R square{which tells that how much percentage variation in dependent variable(weight) can be explain by independent variable(height)}is 21.48 percent, which was low hence as a modification other factors were considered in multiple linear regression and controlled the dependent variable by adding two more independent variables gender and age. This lead to increment in R square value up to 97.84 percent which is statistically significant.

Similar results can be obtained from other statistical tools like SPSS, excel, hence Rstudio is surely not the replacement of other tools but is one of the alternative to perform the significant analysis in effective and efficient manner.

REFERENCES

[1] Kumari K, Yadav S. Linear regression analysis study. J Pract Cardiovasc Sci 2018;4:33-6

[2] R for dummies by Andrie de vries and Joris Meys

[3] Statistics for management by Richard I Levin & David S Rubin

[4] https://www.rstudio.com

[5] R for data science by Hadley Wickham & Garrett Grolemund

APPENDIX 1

Data contains distribution of mean heights and weights of males and females by age.

Age	Height	Gender	Weight
15	1.607	Male	46.27748608
15	1.5145	Female	43.16762691
16	1.6274	Male	48.57222014
16	1.5206	Female	44.25597425
17	1.6446	Male	50.5510142
17	1.5212	Female	44.86941864
18	1.6473	Male	51.88398018
18	1.5216	Female	45.42552991
19	1.6522	Male	53.25771203
19	1.5243	Female	45.9818768
20	1.6518	Male	53.61390967
20	1.5218	Female	46.06275852
21	1.6553	Male	54.77296162
21	1.525	Female	46.605525
22	1.6517	Male	54.80778796
22	1.5239	Female	46.67765132
23	1.6524	Male	55.37303441
23	1.5238	Female	47.06625974
24	1.652	Male	56.02850512
24	1.5234	Female	47.66815488
25	1.6464	Male	55.51376302
25	1.5226	Female	47.71083544
26	1.6483	Male	56.34835854
26	1.5234	Female	48.27154925
27	1.6512	Male	56.9830441
27	1.5252	Female	48.71136174
28	1.646	Male	57.19366076
28	1.5223	Female	48.92025679
29	1.6531	Male	58.26200849
29	1.5202	Female	49.17825109
30	1.6444	Male	56.4605924
30	1.5216	Female	48.96788774
31	1.6523	Male	58.56054397
31	1.5234	Female	49.78003516
32	1.6502	Male	58.46624606
32	1.5241	Female	50.15099669
33	1.6504	Male	59.18861208
33	1.5241	Female	50.42974239
34	1.6492	Male	59.12977031

Age	Height	Gender	Weight
34	1.5224	Female	50.78084556
35	1.6393	Male	56.72899778
35	1.5199	Female	50.40629494
36	1.6486	Male	59.30418437
36	1.5218	Female	51.45874783
37	1.6487	Male	59.93656776
37	1.5227	Female	51.47325944
38	1.6429	Male	58.84082494
38	1.5216	Female	51.49152829
39	1.6481	Male	60.11024979
39	1.519	Female	51.54644474
40	1.6385	Male	56.75418277
40	1.5207	Female	51.36125776
41	1.6472	Male	59.99035194
41	1.521	Female	52.65391716
42	1.6445	Male	59.25297128
42	1.5203	Female	52.32810572
43	1.6432	Male	59.69934897
43	1.5207	Female	52.61002315
44	1.6428	Male	59.42739632
44	1.5207	Female	53.16502999
45	1.6338	Male	56.69598383
45	1.5177	Female	52.14927689
46	1.6448	Male	59.22048451
46	1.5174	Female	52.49706293
47	1.6414	Male	60.26911889
47	1.5155	Female	52.84799315
48	1.6391	Male	58.78387596
48	1.5165	Female	52.54979591
49	1.6393	Male	59.76565186
49	1.5135	Female	53.30417596