Correlation Coefficient Analysis of Data using R Programming: Does It Effectives?

Yagyanath Rimal
Faculty of Science Technology
Pokhara University, Nepal

***Abstract*—This review paper clearly discusses the compression between mathematical regression analysis and R programming output of different research data sets. Its primary purpose is to explain the different regression models using R software whose outputs are sufficiently explanation to reach the conclusion of research variables. Therefore, this paper presents easiest way of data analysis commands and its strengths for data analysis using R programming.**

***Index Terms*— Programming Language; Data analytics; R; Python, Big Data; Regression Equation**

INTRODUCTION

The regression analysis is a statistical process which allows a researcher to estimate the linear relationship among research variables. Data analyzing requires appropriate data analysis tools which predict future design moreover the most important challenge of modern data scientist is to process data, make decisions and forecast the data trend. However out of all languages, R and Python are increasingly using data analysis packages to conclude statistical analysis most. Regression analysis is widely used for prediction and forecast of research based on some pre information values. The linear regression summarizes the association of variables in change with related variables [1] For example sales and advertisement relationship of a company, the sales records is dependent whereas the factors affecting sales are independent relationship which regarded as change with dependent variable or vice versa. Regression analysis would easily help to solve numerically with various relationship. In simple words, regression analysis is used to fit the model relationship between a dependent and independent on research variables. However, the researcher sets its reason to make independent and dependent choices accordingly. In the regression model, the independent variable is labelled as X variable, and the dependent variable is Y variable. The relationship between X and Y can be shown on a graph, with the independent variable X in the horizontal axis, and the dependent variable Y is vertical axis. The aim of the regression model is to determine the straight line relationship that connects X and Y[2]. The straight line connecting any two variables X and Y can be stated algebraically as Y = a + bX where a is called the Y intercept, and b is the slope of the regression line. If the intercept and slope for the line can be determined, then this entirely determines the relationship among research data. The intercept for the line is the point where the line crosses the Y axis. Similarly, the slope of the line b refers to the steepness of the line, whether the line rises or falls sharply. The linear regression model of prediction can apply for single or multiple independent and dependent variables. The dependent model is calculated mathematically with the help of single or multiple independent variable. It is a technique in which the dependent variable is continuous in nature. The relationship between the dependent variable and independent variables is assumed to be linear in nature. The equation of linear regression for multiple is Y=B1+B2X2+B3X3 +…BK+€ where 'y' is the dependent variable to be estimated, and X are the independent variables and ε is the error term. Bi's are the regression coefficients. Although, every regression technique has some assumptions attached to it which we need to meet before running analysis. Once the slope and the intercept have been determined, then the line extends towards infinity in either direction. Therefore, simple linear regression formula Y=βo+β1X1+€ where Y is dependent and X1 is independent variables. B1 is coefficient term which is used to calculate relationship of independent variable, B0 is intercept of two line across when regression line meets at y axis, the € is error terms which regression analysis conclude that for each independent variable the value of y is increased. Therefore, it is concluded that each 1 unit change in one independent variable increased decrease its value of dependent variable if all other coefficient constant.

R programming language is statistical open source programming language which is free and supported by large community where we can do new things is developed by Ross Ihaka and Robert Gentleman. R software is both software and programming language where we can develop many programs. It is also scripting language where we can write many lines of code and produce output in console command. R is available all platform (Linux, Mac, Windows). There are now more than 10,000 R packages available for download [3]. Machils [4] Executive Editor board member of Data & Analytics claimed that there are various large company like Microsoft, the New York Times, Google Maps, Google, Amazon and Facebook are using R programming language for many large servers due to its open source IDE and

Freely distributed under the term of the GNU whose precompiled binaries are distributed form internet site of Comprehensive R Archive Network (CRAN) [5]. R language supports many functions for statistical analyses and graphical visualization in window. Some intermediate results p-value, regression coefficient, residuals can easily produce so that data analysist draw conclusion easily. R has worldwide repository system under http://cran.r-project/org is comprehensive R achieve network. Similar rstudio package is available in https://www.rstudio.com.products/rstudio/download. R stores user defined objects in the workspace the workspace automatically reload the next time starts which supports large interactive feature of different data sets with own help page. The getws () return working directory and setwd ()

command is used to set working directory.

Kopf [6] Explained in New KDnuggets that there were only four dominance languages for Analytics, are R, SAS, Python, and SQL used by 91% of data scientists and decline in popularity of other languages, except for Julia and Scala [6]. Similarly, the author [7] explained that there are R, Python, SAS, MATLAB, SPSS, My SQL and Java were the most dominance languages for data analysis. Increased data availability, more powerful computing, and an emphasis on analytics-driven decision in business has made for data science. Therefore, data scientists should be aware of what are the best solutions for the particular tasks. According to a report from IBM, in 2015 there were 2.35 million openings for data analytics jobs in the US which estimates that number will rise to 2.72 million by 2020. A significant share of people who crunch numbers for a living use Microsoft Excel or other spreadsheet programs like Google Sheets use proprietary statistical software like SAS, Stata, or SPSS. among them R and Python are the two most popular programming languages used by data analysts which are free,R for statistical analysis and Python as a general-purpose programming language which are use machine learning, working with large datasets, for complex data visualizations [6]. Similarly the author Nicolaou [9] estimated that more than 2 million people use R, and a recent poll showed that R is by far the most popular language in data science, used by 61% of respondents (followed by Python, with 39%). Traditionally, banking analysts would pore over Excel files late into the night, but now R is increasingly being used for financial modeling, particularly as a visualization tool Although R works at the core of Google's, Facebook's and banking transition algorithms which hand off the model to be written in Java or Python. Paul Butler famously used R to build a Facebook map of the world back in 2010, proving the rich visualization capabilities of the language. However, some scholars argued that R is a bit passé in industry, because it's slow and clunky with large data sets. Many organizations that leverage his skillset have been able to rediscover and reuse their underutilized data via existing and emerging technologies such as Amazon Web Services, Microsoft Azure, Google Cloud, Microsoft BI Stack, Hadoop, Spark, NoSQL databases, and SharePoint along with related toolsets and technologies [10]. Similarly [11] explained that huge electronic data in modern world should organized through various normalization of data base management so that future analysis and its storage system could only organized in future.

R is cross-platform interoperability is an important feature to have in today's computing world even Microsoft is making its coveted .NET platform available on all platforms after realizing the benefits of technology that runs on all systems [12]. R's software environment is written primarily in C, FORTRAN, and R. Rstudio is very popular IDE used to perform data analysis using R software which primary used for academic research, R is rapidly expanding into enterprise market [13].

You can import data from variety of formats like excel, CSV, and from text files. Data frames is primary data structure which can import files from SPSS or Minitab. Basically, R can handle data from most common sources without problem. However, blogger atlas.it claimed that R is not so great at is data collection from web but lot of work is being carried to address its limitation for data design.

Table No: Two (Mathematically calculation)

s.n	x	y	xy	x2	y2
1	3	11	33	9	121
2	4	12	48	16	144
3	8	9	72	64	81
4	7	3	21	49	9
5	2	5	10	4	25
5	**24**	**40**	**184**	**142**	**380**
	4.8	8	mean		
	?	10			
	10	?	byx	-0.29851	
			bxy	-0.13333	

It is very easy to reshape data frame in R. Tasks like adding new columns, populating missing values etc. can be done with just one line of code. Many new packages like reshape2 allow users to manipulate data frames to fit the criteria set per requirements although R is exploratory analysis. Many models can be written with very few lines of codes. With R, users will be able to build probability distributions and apply statistical methods for machine learning. For advance work in analytics, optimization and analysis, users may have to rely on third party packages. Many popular packages like zoo (to work with time-series), caret (machine learning) represent strength of R. Python is loosely bind programming language with very wide user base. Data Visualization is another strong presentation of data. By default, R allows you to make basic charts and plot graphs which can be saved in variety of formats like jpeg or PDFs. With advance packages like ggvis, lattice and ggplot2 user can extend data visualization capabilities of R program. Like R, data is held in Data Frames which can be used and reused throughout program without hampering performance [14]. Users can apply standard methods of cleaning data or process data to fill out incomplete information just like R. In addition to their differences, there are few common positives about both Python and R which make them so popular among data analysts and statisticians. R is language developed by statisticians for statisticians while python is easier to learn general purpose programming language [15]. The most important algebraic method of measuring correlation is Karl Pearson's Coefficient of correlation or Pearson Ian's coefficient of Correlation. It has widely used application in Statistics. It is denoted by r. Karl Pearson's Coefficient of correlation denoted by r is the degree of correlation between two variables.

REGRESSION

The regression analysis uses various techniques for modeling and analyzing the trends between a dependent

Table No: One

Year	Advertisement	Sales
1	100	500
2	90	450
3	80	340
4	95	480
5	125	?
6	?	600

variables and independent variables. Regression analysis helps us to forecast the prediction of given data. If prediction within the range of given data is known as interpolation whereas if the prediction falls outside is known as extrapolation. Therefore, regression measures how well a given data model fits data. The regression model depends on how the given data behaves well when certain number is made assumption. Before analyzing we must plot scattered plot or scatter graph is a graph using coordinate plane to display value for two variables for a set of data as collection therefore, we can easily predicate from the model. The regression is business mathematics generally use to predicates the future forecast based on the previous records. Suppose XYZ Company had spent money investment on advertisement and gets above sales. On table consideration table no One data sets, what will be the sales if company invest 125 on advertisement in fifth year plan likewise what will be the expenditure to get when sales reaches 600 in 6th year such type of prediction could easily estimate using regression. Suppose in a relational table advertisement is x and sales association are y then what will be the sales when advertisement become 125. Which become y depends on x i.e. (y on x) relationship. Advertisement depends on sales so x depends on y i.e. (x on y) relationship or vice versa. Therefore, there will be two coefficients of regression equation of y on x (i.e. bxy) and x on y (i.e. byx). Which could be written as (y-y-) = byx(x-x-) and (x-x-) = bxy(y-y-) simultaneously, here x- and y- are true mean is easily calculated by summation of x/n and summation of y/n. After putting these values on the above equation, we get two equations in terms of y and x respectively. The value of bxy and byx can be calculated by three ways formula based on given situation the first formula is use if the relationship of two variables are in tabular format and if the standard deviations is given then choice is second formula and covariance is given third priority formula is selected.

1) byx =n€yx-€x€y/n€x2-(€x)2 and bxy =n€xy-€x€y/n€y2-(€y)2

2) bxy=r dx/dy and byx=dy/dx

3) bxy=cov(x,y)/ (dy)2 and byx=cov(x,y)/ (dx)2

Case 1: Calculate the correlation of Table No Two data sets of x and y columns?

Table No: Three (Mathematically calculation)	
byx =n€yx-€x€y/n€x2-(€x)2 =5*184-24*40/5*142-24*24 =(920-960)/(710-576) =-40 /134 =-0.2985	bxy =n€xy-€x€y/n€y2-(€y)2 =5*184-24*40/5*380-40*40 =(920-960)/(1900-1600) =-40/300 =-0.133

Here the mean of x and y variables can be easily calculated then after solving equation byx and bxy can be calculated mathematically with total sum/items (24/5) i.e. 4.8 and (40/5) i.e. 8 corresponding x and y. The Table No Three calculates the bxy and byx respectively then the new value can be predicting easily generally two types of predication will occurs they are; what likely to be the value of x if y equal 10 and what likely to be the value of y if x equal 10. The intermediate values of xy, x2 and y2 are easily calculated when putting values on equations on bxy and byx will easily calculated with Table No Three. Similarly, from Table No Four shows the mathematically solution of bxy and byx regression equation by putting values of bxy and byx respectively and calculated the final model y =-.298x+9.4304 and x=-0.133y+5.864 respectively which describes the variable relationship. When putting x=10 and y=10 on the respective equation are intercepts of line slope have –ve relationship similarly and the constant are added when to explain model. The new values will be easily predicated when we put their values in equation.

Note: In regression equation y on x number associates with variable x represent byx and vice versa.

Case 1: Solution using R Programing

```
oneee <- read_excel("C:/Users/Yagya/Desktop/oneee.xlsx")
> view(oneee)
> mydata=oneee
> str(mydata)
Classes tbl_df, tbl and 'data.frame': 5 obs. of 2 variables:
 $ x: num 3 4 8 7 2
 $ y: num 11 12 9 3 5
> cor(mydata$x,mydata$y)
# x on y correlation
[1] -0.1995019
> mod=lm(mydata)
# lm command gives intercept and slope
> mod
Call:
```

Table No: Four (Mathematically calculation)	
y on x y-y-=byx(x-x-) y-8=-.298(x-4.8) y=-**.298x** +1.4304+8 y=-.298x+9.4304 if x=10 y=-.298*10+9.4304 y= -2.98+9.4304 y=6.450	x on y (x-x-)=bxy(y-y-) x-4.8=-0.133(y-8) x=-.133y+1.064+4.8 x=-**0.133y**+5.864 if y=10 x=-.133*10+5.864 x=-1.33+5.864 x=4.534

```
lm(formula = mydata)
Coefficients:
```

```
(Intercept)          y
    5.8667     -0.1333
>cor(mydata$y,mydata$x)
# y on x correlation of data sets
[1] -0.1995019
> mod=lm(mydata)
> mod
Call:
lm(formula = mydata)
Coefficients:
(Intercept)          y
    5.8667     -0.1333
>oz2=coef(mod)[1]+coef(mod)[2]*10
# prediction when       x=10
> oz2
(Intercept)  4.533333  (i.e x on y)
Similarly
>mydata2=oneee
> str(mydata2)
```

Classes 'tbl_df', 'tbl' and 'data. frame':5 obs. of 2 variables:

```
 $ x: num  3 4 8 7 2
 $ y: num  11 12 9 3 5
> mod2=lm(y~x, data=mydata2)  # y on x correlation
> mod2
Call:
lm(formula = y ~ x, data = mydata2)
Coefficients:
(Intercept)          x
    9.4328     -0.2985
> oz3=coef(mod2)[1]+coef(mod2)[2]*10
> oz3
(Intercept)
  6.447761 (y on x)
```

Here you can calculate linear model of x is dependent and y is independent priority you may reverse its preferences when you think another relation occurs in-between variables. B0 is 5.8667 and b1 slope is -0.1333 when x on y and 9.4328 and -0.2985 is slope when y on x whose values were exactly similar to Table No Three. Similarly, the prediction is calculated when y=10 and x= 10 based on bxy and byx simultaneously.

```
>summary (mod1) # gives more on its minimum, maximum, first quartile and 3 quartile values
Call:
lm (formula = x ~ y, data = q2)
Residuals:
 Min    1Q    Median 3Q    Max
-1.4000 -0.2667  3.3333  1.5333 -3.2000
```

Residual describe the minimum, maximum, median 1quatrialand 3rd quartile values

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.8667    3.2962  1.780   0.173
y           -0.1333    0.3781 -0.353   0.748
```

Here, (5.8667) is intercept (B0) and (-0.1333) is slope B1, So y-=5.8667+-0.1333(X) which can be easily interchange x variable=5.8667+-0.1333(y) variable.

Here p value is greater than 0.05 we cannot reject the Null Hypothesis which implies that there is no such significant relationship between x and y variables when x is dependent and y is independent (y on x).

Note: Correlation function always requires int data type if the data type has alphanumeric characters like 54 Rs, 60 Rs and 70 Rs whose binary position could be changed with command like: ifelse (adult$income==’50 Rs’,0,1) become binary data sets. The default cor(x,y, method=c(“person ”)) is use if sometime data sets has multiple values could easily change into binary using dummy variable: dummyVars(“~.”,data=adults) under caret library.

Case 2: Suppose a company has sales and advertisement mean are 40 and 6 and whose standard deviation are 10 and 1.5 respectively and r=0.9 then a) what likely to be sales if adv. expenditure is 10 thousand b) what likely to be adv. if sales target is 60 cores.

Here suppose sales is (x) independent variable and advertisement is (y) dependency of relational condition then question a become x depends on y whose equation is could be solved with the help of Table No Five. From the given question the mean 6 and 40 are given. The covariance bxy and byx are easily calculated r dx/dy, whose values are again sets to the equation calculate x and y model. Then real requirement values are predicated simultaneously in each equation. The given condition could be mathematically solving on Table No Five.

Table No: Five (Mathematically calculation)

x on y	y on x
(x-x-)=bxy(y-y-)	(y-y-)=byx(x-x-)
x-40=bxy(y-6)	y-6= byx(x-40)
bxy=r*dx/dy	byx=r*dy/dx
=.9*10/1.5	=0.9*1.5/10
=6	=0.135
x-40=6(y-6)	y-6=0.135(x-40)
x=6y+40-36	y=0.135x+6-5.4
if y=10	if x=60
x=6*10+4	y=.135(60)+0.6
=64	=8.7

Case 2: Solution using R Programming

```
> bxy=.9*10/1.5
> bxy
[1] 6
> x=6y+40-36
> x=6*10+6
> x
[1] 66
> x=6*10+4
> x
[1] 64
> byx=.9*1.5/10
> byx
[1] 0.135
>y-6=.135x-.135*40
> y=.135*60+.6
> y
[1] 8.7
```

Table No: Sixa

Age	Total	Players
15-20	200	150
20-25	270	162
25-30	340	170
30-35	360	180
35-40	400	180
40-45	300	120

Case 3: Similarly another example the covariance of (x,y) is 16 and the variance of x and y are 25 and 30 respectively and mean of x- and y- are 20 and 30 then estimate y when x equal to 30. Here from the given relationship y depends on x, which could easily solve mathematically from Table No Six.

Table No: Six (Mathematically calculation)

(x-x-)=byx(y-y-)	=16/25=.64
y-30=byx(x-20)	y-30=.64(x-20)
byx= cov(x,y)/(dx)2	y-30=.64*x-12.8
variance of x=(dx)2	y=.64x+17.2
25=(dx)2	if x=30
5=(dx)2	y=.64(30)+17.2
byx= cov(x,y)/(dx)2	36.4
=16/(5*5)	

Case 3: Solution using R Programming

```
byx=16/25
> byx
[1] 0.64
> y=.64*(30)-.64*(20)+30
> y
[1] 36.4
```

Case 4: To find the correlation between age group data and playing habits of some players from its total players like table no sixa. To solve the relationship of group data having total and its part representation first we should calculate midpoint of group age using just adding two ranges and divide by 2. Likewise the y is calculated proper relation of a part is represented with its total part (150/200)*100 for each respective data items as shown Table No Seven. Then based on x and y data the dx and dy can easily calculated with its mid value of subtraction and divided by fixed constant (x-32.5)/5 and (y-40)/5 here the term subtracted from fixed constant largely reduced the data set, whose dx2 dy2 and summation of dxdy were calculated. After calculation the regression could easily calculated with mathematical formula.

Table No: Seven (Mathematically calculation)

Age	Total	Players	x	y	dx	dy	dx^2	dy^2	dxdy
15-20	200	150	17.5	75	-3	7	9	49	-21
20-25	270	162	22.5	60	-2	4	4	16	-8
25-30	340	170	27.5	50	-1	2	1	4	-2
30-35	360	180	32.5	50	0	2	0	4	0

$$r = \frac{n€xy - €x€y}{\sqrt{n€x2 - (€x)2}\sqrt{n€y2 - (€y)2}}$$

$$r = \frac{n€dxdy - €dx€dy}{\sqrt{n€dx2 - (€dx)2}\sqrt{n€dy2 - (€dy)2}}$$

$$r = \frac{6(-30) - (-3)(16)}{\sqrt{6*19 - (-3)2}\sqrt{6*74 - (16)2}}$$

$$r = \frac{-180 + 46}{\sqrt{114 - (9)}\sqrt{144 - (188)}}$$

$$r = \frac{-132}{\sqrt{105}\sqrt{\ } (188)}$$

$$r = \frac{-132}{\sqrt{105}\sqrt{\ } (188)}$$

$$r = \frac{-132}{140.5}$$

$$r = -.93$$

Case 4: Solution using R Programming

After calculating mid values of Age (15+20)/2and y=150/200*100 of total player and present players respectively.

```
x=(onee$x-32.5)/5
> x
[1] -3 -2 -1  0  1  2
> y=(onee$y-40)/5
> y
[1] 7 4 2 2 1 0
> dx=sum(x)
> dx
[1] -3
> dy=sum(y)
> dy
[1] 16
cor (onee$x,onee$y)
[1] -0.9395077
```

Case 5: How to calculate correlation regression of continuous series of bivariate Table No Eight?

Solution: First calculate mid points of each continuous bivariate data. The dx and dy are calculated by making the term smaller (x-55)/ 10 and (y-62.5)/10 respectively. The cross frequency can be easily calculated by using matrix

Table No: Eight

	Computer				
Math	40-50	50-60	60-70	70-80	Total
50-55	**4**	**7**	**5**	**2**	18
55-60	**6**	**10**	**7**	**4**	27
60-65	**6**	**12**	**10**	**7**	35
65-70	**3**	**8**	**6**	**3**	20
Total	19	37	28	16	100

multiplication (-2)*(-1)*4=8 then fdx is calculated using matrix table 19*(-1) of corresponding matrix value -19. The dx2 and dy2 are calculated by squaring the value of dx and dy respectively similarly, the fdxdy can be calculated by the sum of cross frequency multiplication stored in parenthesis on table data. By putting table no Nine values in formula.

Table No: Nine (Mathematically Calculation)

	Computer									
Math			40-50	50-60	60-70	70-80				
	Mid	x	45	55	65	75				
	Y	dy/dx	-1	0	1	2	f	fdy	fdy^2	fdxdy
50-55	52.5	-2	**4 (8)**	**7 (0)**	**5 (-10)**	**2 (-8)**	18	-36	72	-10
55-60	57.5	-1	**6 (6)**	**10 (0)**	**7 (-7)**	**4 (-8)**	27	-27	27	-9
60-65	62.5	0	**6 (0)**	**12 (0)**	**10 (0)**	**7 (0)**	35	0	0	0
65-70	67.5	1	**3 (-3)**	**8 (0)**	**6 (6)**	**3 (6)**	20	20	20	9
		f	19	37	28	16	**100**	-43	119	-10
		fdx	-19	0	28	32	**41**			
		fdx^2	19	0	28	64	111			
		fdxdy	11	0	-11	-10	-10			

$$r = \frac{n€xy - €x€y}{\sqrt{n€x2} - (€x)2\sqrt{n€y2} - (€y)2}$$

$$r = \frac{n€fdxdy - €fdx * €fdy}{\sqrt{n€fdx2} - (€fdx)2 * \sqrt{n€fdy2} - (€dfy)2}$$

$$r = \frac{100 * (-10) - (-41) * (-43)}{\sqrt{100 * 111} - (41)2 * \sqrt{100 * 119} - (43)2}$$

$$r = \frac{-1000 - 1763}{9729.8}$$

r= -2763/9729.8

r =-0.2839709

Case 5: Solution using R Programming

The mid values of two subjects' mathematics and computer are calculated then their dx and dy are easily calculated (Mathematics) dy= (mid value -62.5)/interval and (Computer) dx= (mid value-55)/ interval of all frequency of bivariate data. With referring dx and dy , the fdx, fdy, fdx2 fdy2 and fdxdy , fdydx and cumulative sum were calculated.

```
>mat=matrix(c(4,7,5,2,  6,10,7,4,  6,12,10,7,  3,8,6,3), nrow=4,byrow=TRUE)
   > mat
       [,1] [,2] [,3] [,4]
   [1,]    4    7    5    2
   [2,]    6   10    7    4
   [3,]    6   12   10    7
   [4,]    3    8    6    3
   > dx=c(-1,0,1,2)
   > dx
   [1] -1 0 1 2
   > dy=c(-2,-1,0,1)
   > dy
   [1] -2 -1  0  1
   f1=c(sum(mat[,1]),sum(mat[,2]),sum(mat[,3]),
   sum(mat[,4]))
   > f1
   [1] 19 37 28 16
   > f2=c(sum(mat[1,]),sum(mat[2,]),sum(mat[3,]),
   sum(mat[4,]))
   > f2
   [1] 18 27 35 20
   dx=c(-1,0,1,2)
   > f1
   [1] 19 37 28 16
   > f1dx=c(f1*dx)
   > f1dx
   [1] -19   0  28  32
   > f1dx2=f1*dx*dx
   > f1dx2
   [1] 19  0 28 64
   mat22=c(8,0,-10,-8,6,0,-7,-8,0,0,0,0,-3,0,6,6)
   > mat22
    [1]  8  0 -10 -8  6  0 -7 -8  0  0  0  0 -3  0  6  6
   > mat2=matrix(mat22,nrow=4,byrow=TRUE)
   > mat2
       [,1] [,2] [,3] [,4]
   [1,]    8    0  -10   -8
   [2,]    6    0   -7   -8
   [3,]    0    0    0    0
   [4,]   -3    0    6    6
   > f1dxdy=c(sum(mat2[,1]),sum(mat2[,2]),sum(
   mat2[,3]),sum(mat2[,4]))
   > f1dxdy
   [1]  11   0 -11 -10
   f2dy=c(f2*dy)
   f2dy2=f2*dy*dy
   cor(f1dx2,f2dy2)
   [1] -0.2672921
   >cor(f1dx2, f2dy2, method = c("pearson", "kendall", "spearman"))
   =-0.2672921  # i. e. ~-0.2839709
```

The value of correlation coefficient (r) will always lie between 1. When r = +1, it means there is a perfect positive correlation between the two values. When r = -1 When the data in the bivariate frequency distribution is large it can be classified into a bivariate frequency table (or correlation table as it is generally called). The correlation coefficient can be computed using the functions cor() or cor.test():

>cor() computes the correlation coefficient

>cor.test() test for association/correlation between paired samples. It returns both the correlation coefficient and the significance level (or p-value) of the correlation.

```
>cor(x, y, method = c("pearson", "kendall", "spearman"))
>cor.test(x, y, method=c("pearson", "kendall", "spearman"))
```

x, y: numeric vectors with the same length and method: correlation method. If your data contain missing values, use the following R code to handle missing values by case-wise deletion. cor(x, y, method = "pearson", use = "complete.obs")

CONCLUSION

The relationship between research variables is considered as correlation, in fact it is a number which can be used to describe the degree of association between them. Correlation lies in between -1 to 1 express its relation among research variables. The multiple correlation and partial correlation are used among three or more variables. Base on the r value there are some correlation occurs when r is –1, we say there is perfect negative correlation, when r is a value between –1 and 0, we say that there is a negative correlation, when r is 0, we say there is no correlation, when r is a value between 0 and 1, we say there is a positive correlation and when r is 1, we say there is a perfect positive correlation. Therefore, the research data could be easily analyzed using R programming rather than mathematically tedious calculation.

REFERENCES

[1]Fiona, G. a. (2018, July).https://www. bmj.com/about-bmj/editorial-staff. Retrieved from https://www.bmj.com/about-bmj/editorial-staff.

[2]Astrid Schneider, D. M. (2010). Linear Regression Analysis. MEDICINE.

[3]Smith, D. (2018, July). http://blog. revolutionanalytics.com/2017/01/cran-10000.html. Retrieved from http://blog.revolution analytics.com/2017/01/cran-10000.html

[4]Machlis, S. (2018). Computerworld. Data & Analytics,.

[5]Paradis, E. (2005). R for Beginers. Universit of Muntipiler, 2.

[6]Kopf, D. (2017). Which Programming Language Should you learn? https://qz.com/1063071/the-great-r-versus-python-for-data-science-debate/.

[7]Piatetsky, G. (2014). Retrieved from https://www.kdnuggets.com/2014/08/four-main-languages-analytics-data-mining-data-science.html

[8]Jeevan, M. (2018). How I chose the right programming language for Data Science. 2.

[9]Nicolaou, A. (2018). The 9 Best Languages For Crunching Data. Retrieved from https://www.fastcompany.com/3030716/the-9-best-languages-for-crunching-data

[10] Cirillo, A. (2017). R Data Mining.

[11]Rimal, Y. (2018). DOES THE DATABASE FUNCTIONAL DEPENDENCY AND ITS NORMALIZATION MAKE UNIFORM DATABASEMANAGEMENT IN FUTURE? Asia Associate Research, 25.

[12]Annanomous. (2018). Learn R Programming. Retrieved from https://www.datamentor.io/

[13]Ihaka, R. a. (1996). Journal of computational and graphical statistics. Retrieved from R: a language for data analysis and graphics. .

[14]Pedregosa, F. V. (2011). Scikit-learn: Machine learning in Python. . Retrieved from The Journal of Machine Learning Research, 12, pp.

[15]Nasridinov, A. a. (2013). Third International Conference on (pp. 564-565). IEEE. Visual Analytics for Big Data Using R. In Cloud and Green Computing (CGC), .

[16] Fan, W. a. (2013). . ACM sIGKDD Explorations Newsletter, 14(2), pp.1-5. Mining big data: current status, and forecast to the future.