Correlation and Regression Analysis Using SPSS

Sarad Chandra Kafle
Asst. Professor
Birendra Multiple Campus, Bharatpur

Abstract

The objective of this study is to share knowledge on how to use Correlation and Regression Analysis through Statistical Package for Social Science (SPSS). This study has used secondary data to demonstrate the way of using very popular statistical tool on using correlation and regression analysis for novice researchers. Among various statistical tools, correlation and regression analysis are mostly used tools in many research works., e.g. the field of management, medicine, social science and education. However, not all the researchers may know whether the tools are fit to use, how to carry the analysis and how to interpret the obtained results. The results shows that novice researchers need to know the proper knowledge and skill to analyse the quantitative data. The implications of this study is willing to share the knowledge on correlation and regression analysis and the way of analyzing through very popular software package SPSS.

Keyword: *Statistical tools, Test of Significance, p-value, Hypothesis, Dependent and Independent variables*

1. Introduction

In quantitative study, researcher willing to use very famous statistical tool regression & correlation, however due to lack of sufficient knowledge on regression & correlation analysis their desired havenot fulfilled or even they use the tool, the tool haven't been properly used. To provide clear cut idea on correlation regression, its use way of interpretation of output of analysis, this research article has been prepared. Relation between two or more variables can be studied by using Correlation and Regression. Two variables are said to be related if change in the value of one variable changes the value of other variable. Here the term change implies either increase or decrease in the value of variable. Relationship between variables can be studied by the method of correlation or regression. Such an analysis of relationship can be carried for quantitative or qualitative variable however this paper includes only the analysis of relationship between quantitative variables. Those variables which are measurable and thus have unit are quantitative variables. Study of relationship between two quantitative variables at a time is simple regression or simple correlation and relationship between more than two quantitative variables may be partial correlation or multiple correlation or multiple regression according to the objective/nature of study and variables included in the study (Sthapit, Yadav, Khanal, & Dangol, 2017).

Strength of relationship between two or more variables is studied by using Correlation. Correlation is statistical tool that measures how strong relationship exists between variables. Value of correlation lies in between -1 and +1. Nearer the value of correlation to zero weak is the relationship between the variables, similarly if the value of correlation close to one implies higher (close) relation between variables. Hence correlation is a value which tries to explain degree of association between variables whereas regression tries to explain the relationship between variables using a mathematical function. (Gupta & Kapoor, 2014).

1.1 Correlation Analysis:

The correlation analysis refers the degree of relationship between variables. But it does not explain about which of the variable is cause and which one is the effect. Study of correlation between two variables is called simple and between more than two variables may be partial or multiple.

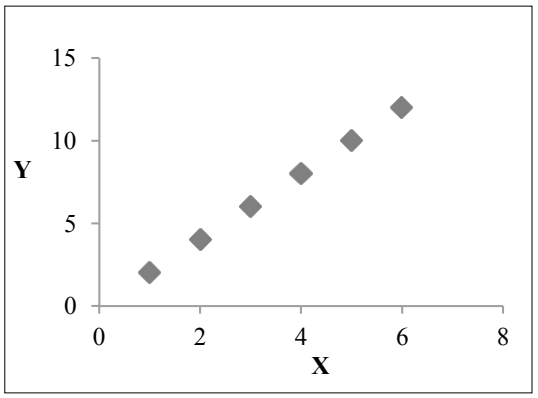
Correlation can be studied by two methods, diagrammatic method and mathematic method. Diagrammatically it is studied with the help of scatter diagram which cannot provide exact value of correlation in all case. Mathematically many methods and formulae are there however Karl Pearson's Method is widely used (Magnello, 2009).

1.2 Diagrammatic method:

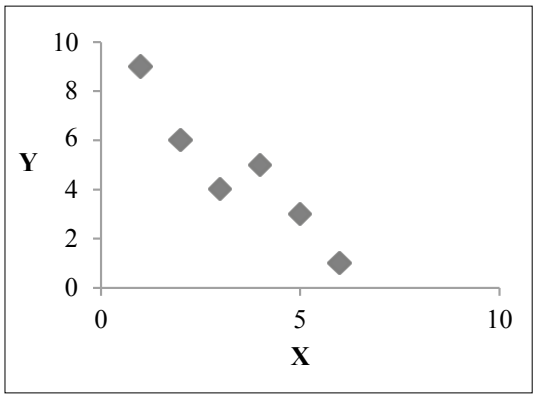
Diagrammatically correlation can be studied by scatter diagram. This is presented in figure-1. To plot a scatter diagram, a dot is provided for each pair of data for X and Y, plotting the value in X axis and That of Y on respective Axis. More closer and the arranged point shows higher correlation between two variables. Analysis of the strength of relationship is based on the trend which is seen in scatter diagram. If increase in the value of one variable makes increase in the value of other variable, (direct relationship), then the correlation is said to be positive whereas if the scatter shows opposite trend to that then the relation is negative. (Shrestha, Khanal, & Kafle, 2014).

Following scatter diagram helps to clearly the different types of correlation between two variables X and Y.

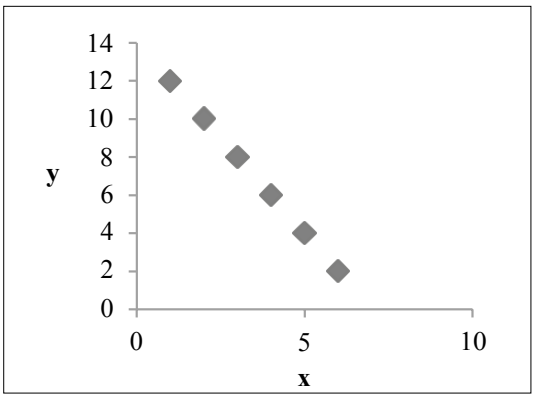
Fig-1: Scatter diagram



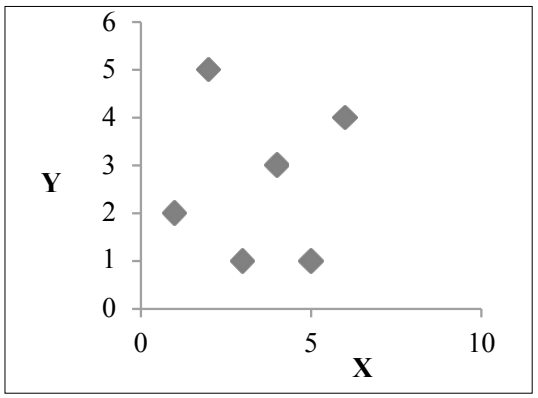
Perfect Positive correlation(r = +1)



Highly negative correlation



perfect negative correlation(r = -1)



No correlation (r =0)

(Fig Source: (Shrestha, Khanal, & Kafle, 2014))

1.3 Karl Pearson's correlation coefficient:

This is mathematical method to study the degree of association between two variables. It is used to study the correlation between two quantitative variables and denoted by r. Formula to calculate Karl Pearson's correlation coefficient is as follow (Sthapit, Yadav, Khanal, & Dangol, 2017) -

$$r = \frac{cov(X, Y)}{\sigma_x \sigma_y}$$

$$\text{or, } r = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

1.4 Spearman's rank correlation:

To study the degree of association between two variables whose values are written in rank, rank correlation is used. For quantitative variables ranks can be provided according to their increasing or decreasing order of magnitude. Rank correlation is denoted by r_s and its formula for calculation is as

$$r_s = 1 - \frac{6\,\Sigma d^2}{n^3 - n}$$; When the ranks are not repeated.

$$= 1 - \frac{6\left[\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \ldots\ldots\ldots\ldots\right]}{n^3 - n}$$; when ranks are repeated (Magnello, 2009)

1.5 Kendal tau:

It also rank correlation and can be used in the case where spearman's rank correlation can be calculated. It is denoted by ī(tau). Formula to calculate Kendal tau is as

$\tau =$ _____ ; when ranks are not repeated
( )

_____ _____
√( ) √( ) = ; when ranks are repeated

(Gupta & Kapoor, 2014)

1.6 Interpretation of Correlation Coefficient:

Correlation calculated using any formula and method stated above can be interpreted as below

If r = 1, the correlation is said to be perfect positive.

If r = -1, the correlation is said to be perfect negative.

If r = 0, the variables X and Y are said to be uncorrelated.

If $0 < r \geq 0.4$, low correlation.

If $0.4 \leq r < 0.7$, moderate correlation.

If $0.7 \leq r < 1$, high correlation.

The value of correlation coefficients nearer to +1 or -1 be interpreted as very high positive or negative correlation and nearing zero is considered as very low (Gupta & Kapoor, 2014).

1.7 Partial correlation:

Correlation between two variables keeping the effect of remaining variable constant is partial correlation. If we are interested to study the relationship between two variables X_1 and X_2 while there exists another variable X_3 then the correlation between X_1 and X_2 keeping the value of X_3 constant is partial correlation between X_1 and X_2 keeping X_3 constant, denoted by $r_{12.3.}$ Value of partial correlation lies in between -1 and +1.

$$r_{12.3} = \frac{\qquad\qquad}{\sqrt{(\qquad)\ (\qquad)}}$$

1. 8 Multiple Correlation:

Correlation between predicted and the actual values of the dependent variable in a linear regression model that includes an intercept. In other words it is the relationship between dependent variable and joint effect of independent variable on dependent variable In statistics, the coefficient of multiple correlation is a measure of how well a given variable can be predicted using a linear function of a set of other variables. If X_1 be dependent variable which is described by X_2 and X_3 then the correlation between actual value of X_1 predicted value of X_1 is denoted by $R_{1.23,}$ in other way, it is the correlation between dependent variable X_1 and joint effect of X_2 and X_3 on X_1 The value of multiple correlation lies in between 0 and 1.

$$R_{1.23} = \sqrt{\qquad\qquad}$$

1.9 Regression analysis:

Regression analysis tries to study the relationship between two or variables with the help of equation, the equation is called regression line. The line is also called line of best fit since it is obtained by the method of least square. Least Square Method is estimation of parameters of regression equation by minimizing the error sum of square of dependent.

Regression analysis established the nature of relationship between two or more variables and then estimates the unknown variable (dependent variable) with the help of known variable (independent variables). In other words there are two types of variables in a regression analysis. The variables, which is used to predict the variable of interest is called the independent or explanatory variable or predictor, and the variable whose value is to be predicted is called the dependent variable or explained variable or regressed. (Montgomery, 1982)

1.10 Simple regression:

If relationship between two (one dependent and other independent) variables is studied at a time then the regression is called simple, whereas the study of more than two variables at a time is multiple.

If Y is a dependent variable and X is an independent variable then regression equation of Y is-

$Y = a + b X$

Where,

a = y intercept = constant = value of Y when X = 0

b = regression coefficient = slope coefficient = change in the value of Y per unit change in the value of X.

1.11 Multiple Regressions:

Let 'y' is the dependent variable and x_1, x_2, x_3 ..... x_k be the 'k' independent variables. Then the multiple regression model is defined as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots\ldots + \beta_k x_k + e$$

Where,

y = dependent variable and x_1, x_2, x_3 …… x_k are independent variables.

β_0 = y-intercept.

β_1 = Slope of y with variable x_1 holding the remaining variables x_2, x_3 …,x_k constant or Regression coefficient of y on x_1 holding the remaining variables x_2, x_3 …… x_k constant. And so on. (Dendukuri & Reinhold, 2005)

Some pre-requisities to carry linear regression model are

- There is linear relationship between quantitative dependent and independent variables
- There is no presence of autocorrelation of residuals.
- The mean of residuals is zero.
- There is equal variance of residuall or presence of homoscedasticity.
- The independent variables are uncorrelated with errors.
- There is absence of multicollinearity. (Zaid, 2015)

1.12 SPSS

SPSS refers to Statistical Package for Social Science. It is statistical software which eases to compile and analyze data. We can compile or entry collected primary data or secondary as same as Microsoft Excel. Its menu bar is helpful to analyze the data thus entered easily. Many statistical analysis can be carried using SPSS (Arkkelin, 2014).

Many researchers have applied the correlation and regression analysis in their thesis, articles and their documents, however; they are not yet confident for the appropriate use of correlation and regression analysis and how to fit these statistical tools in their research works. In some cases, their interpretation may mislead their research studies. Many novice researchers are willing to use correlation and regression analysis but they don't know how to use these tools during their data analysis. The primary objective of this study is to share knowledge on regression and correlation analysis and required conditions to use in their research paper.

2. Method & Materials

This study is based on sampled secondary data of 423 maternity women respondents admitted in Chitwan Medical Sciences(CMS), Bharatpur, Chitwan, Nepal during the period 2017 July to August 2017 for maternity. The data used in this study were accessed via library of Chitwan Medical Sciences. The sample data of infant's ages and weight were entered into computer software (SPSS) and analyzed using regression and correlation. Different published articles were googled through online resources, for example, google, bookboon.com, uef.fi, and http://www.oxfordcollege.edu.np. All the research materials were embedded in correlation and regression analysis. The collected materials were initially observed their abstracts, methods and findings to find the deep knowledge on the research phenomenon.

3. Results & Discussion

To study the association between quantitative variables, correlation analysis can be carried in SPSS. To start this analysis, at first select Analyze then define the variables between which variable researchers wants to determine correlation and then choose Pearson's correlation, Kendal tau or spearman's according to the nature of data. For test of significance tail of the test can be defined. After completing these actions

and clicking on ok button an output window will show result of correlation analysis as in

Table 1. Correlation output table using SPSS

		Age of respondent in month	Height of respondent in cm
Age of respondent in month	Pearson Correlation	1	.853**
	Sig. (2-tailed)		.000
	N	423	423
Height of respondent in cm	Pearson Correlation	.853**	1
	Sig. (2-tailed)	.000	
	N	423	423
**. Correlation is significant at the 0.01 level (2-tailed).			

Table 1 is correlation analysis output table for correlation between age and height of respondents. The correlation coefficient is 0.853 which is high degree of positive correlation between height and weight of the respondents. Also the correlation coefficient is significant as its p-value is 0.00 and is less than significance level($\alpha = 5\ \%$).

To find out how these two variables are related regression analysis is carried. To carry this analysis researcher has selected 'Analyze' then 'Regression' and then 'Linear' successively. Then researcher define dependent and independent and independent variable and then clicking on 'Ok', following output table is obtained as shown in Table-2, Table-3 and Table-4.

Table 2. Model Summary

Model 1	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.853[a]	.728	.727	7.67832

a. Predictors: (Constant), Age of respondent in month

Table 2 shows coefficient of determination (R square) 0.728, which means 72.8% variation in dependent variable (Height) is explained by independent variable (Age).

Table 3. ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	66311.832	1	66311.832	1124.755	.000[b]
	Residual	24820.758	421	58.957		
	Total	91132.590	422			

a. Dependent Variable: Height of respondent in cm

b. Predictors: (Constant), Age of respondent in month

Table 3 tries to test overall goodness of fit of fitted regression model. From above table it can be concluded that the fitted model is significant as P-value of F statistics is 0.00 and it is less than level of significance level($\alpha = 5\%$).

Table 4. Coefficient table

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	59.350	0.679		87.373	0.000
Age of respondent in month	0.811	0.024	0.853	33.537	0.000

a. Dependent Variable: Height of respondent in cm

Coefficient table helps to determine the regression equation, the column Unstandardized Coefficients and its sub column 'B' provides the regression coefficients. First one is constant or y intercept and second one is regression coefficient of height (Y) on age(X). Hence the regression equation using coefficient table is

$$\widehat{Y} = 59.35 + 0.811\ X$$

The regression coefficient of height on age is found to be 0.811 which implies that any child which is one month elder than other child is 0.811 centimeter taller than earlier. Also, the regression coefficient is

significant as p-value (0.00) is less than level of significance level ($\alpha = 5\ \%$).

4. Discussion

The results show that using correlation and regression via SPSS is useful for the novice researchers. The results also highlighted that the using correlation and regression is embedded only in quantitative data. In practical life researcher can find many quantitative variables which are related to each other, their degree of relationship can be measured by correlation and how two or more variables are related can be described by an equation, e.g. an equation is regression equation. Manually, the calculation of regression equation and correlation is very complex for big data ,so it requires software via SPSS which is very easy and faster. The results also highlighted that correlation and regression are two key data analysis tools in quantitative approach because Logistic Regression Model helps in predicting probability of occurrences of y dependent variable to x independent variables, when the dependent variable is dichotomous. Researchers can use dichotomous variables, e.g. health status(sick or not), employment status(employed or unemployed), labour force participation (part or not part of the labour force) and family planning method (which type). The results also summarized that Logistic Regression Analysis is more flexible method because it makes no assumptions about the nature of relationship between independent and dependent variables. The limitations of this study are the secondary data analysis, limited research materials, limited knowledge on statistical tools, limited literature review, limited areas of research knowledge, limited knowledge on correlation and regression analysis. Due to these limitations of this research, the current research cannot give the guarantee for the radiality and validity of data and findings. It is recommended that future research has to focus on rich literature review and primary research on how correlation and regression can be effectively use in data analysis processes of quantitative methods. It is also recommended that a details steps of correlation and regression analysis has to focus in future research study to make helpful for the novice researchers.

Reference

Arkkelin, D. (2014). *Using Spss to Understand Research and Data Analysis.* Valparaiso: Valparaiso University.

Dendukuri, N., & Reinhold, C. (2005). Correlation and Regression. *American journal of Roentgenology*, 3-18.

Draper, N. R., & Smith, H. (2011). *Applied Regression Analysis.* Noida: Wiley India Pvt. Ltd.

Gujarati, D. N., C, P. D., & Gunasekar, S. (2015). *Basic Econometrics.* New Delhi: McGraw Hill Education (india) Pvt. Ltd.

Gupta, S. C., & Kapoor, V. K. (2014). *Fundamentals of Mathematical Statistics.* Mumbai: Sultan Chand and Sons.

Magnello, M. (2009). Karl Pearson and the Establishment of Mathematical Statistics. *MInternational Statistical Review / Revue Internationale De Statistique*, 3-29.

Mehta, B. C., & Kapoor, K. (2005). *Fundamentals of Econometrics.* Mumbai: Himalaya Publishing House.

Montgomery, D. (1982). *Introduction to linear Regression Analysis.* New Delhi: Willy.

Shrestha, M. P., Khanal, P. R., & Kafle, S. C. (2014). *Business Statistics.* Kathmandu: Sabdartha Publication.

Sthapit, A. B., Yadav, R. P., Khanal, S. P., & Dangol, P. M. (2017). *Fundamentals of Statistics.* Kathmandu: Asmita Publication.

Zaid Mohamed Ahmed, (2015). Correlation and Regression Analysis; Statistical Economic and Research and Training Centre for islamic countries.