USEFULNESS OF CORRELATION ANALYSIS

Samithambe Senthilnathan
PhD (Bus./Fin.), MSc (Mgmt.), BSc (Bus. Admin.)
Academic Consultant, International Training Institute
Papua New Guinea

Abstract

The measure of correlation coefficient (r or R) provides information on closeness of two variables. Irrespective of non-linear correlation, this paper mainly considers the linear correlation analysis, as it is most likely applied in social science studies. Explicitly, the purpose of carrying out correlation analysis is almost the same in quantitative analytical studies, thus becoming useful to explore the association between independent and dependent variables. This paper, as an extension, attempts additionally to explain the usefulness of linear correlation coefficient between two variables in the context of identifying the level of multicollinearity and mediating/moderating status of independent variables in a model. This paper also demonstrates how the level of multicollinearity can be explored by using correlation coefficient of two independent variables in a regression model.

Keywords : correlation, multicollinearity, Variance Inflation factor, VIF
JEL code : C12, C18, C20, C30

USEFULNESS OF CORRELATION ANALYSIS

Samithambe Senthilnathan

1. INTRODUCTION

Many studies use correlation analysis to explore the degree association between study variables. Especially in social science research, linear correlation analysis is a tool for representing the closeness of one related variable to another. The linear correlation coefficient (r or R) is such a measure providing information to the extent to which two variables have very close association. Though correlation analysis can be linear and/or non-linear, this paper mainly focuses on the linear correlation analysis, as it is mostly used in social science studies.

The purpose of carrying out correlation analysis is almost the same in every study and mostly, a correlation analysis becomes useful to explore the associative relationship between independent and dependent variables. Thus, as an extension, this paper attempt additionally to explain the usefulness of linear correlation coefficient between two variables in the context of identifying level of multicollinearity and mediating/moderating status of independent variables in a model.

The rest of this paper is therefore organised as: Linear Correlation: Meaning and Coefficient, Use of Correlation Coefficient and Concluding Remarks.

2. LINEAR CORRELATION: MEANING AND COEFFICIENT

Correlation is meant for exploring the degree of relationship between two variables in consideration. Correlation coefficient is the measure to quantify such degree of relationship of the variables. Generally, two correlation coefficients are used in applications, namely: Pearson's Product Moment Correlation Coefficient and Spearman's Rank Correlation Coefficient. This paper primarily considers the applications of Pearson's Simple Linear Correlation in exploring the relationship between variables.[1]

In 1896, correlation coefficient is first formulated and explored by Karl Pearson (Hauke and Kossowski, 2011), with the concepts of correlation by Francis Galton and the relative contribution by Auguste Bravais (Denis, 2001). Hauke and Kossowski (2011) do endorse that the Pearson's

[1] Correlation by measure can be two types: Simple Correlation and Multiple Correlation (Partial and Total). Also, correlation can be linear of non-linear.

Product Moment Correlation Coefficient (**R** or **r**) is a scale to measure the strength of linear association between variables. As it measures the degree of linear association of variables, interval or ratio variables should be in consideration with a condition that the variables considered should fall in normal distribution.

Pearson's mathematical formulation to quantify the degree of relationship (R) between variables, namely, X and Y, can be given as:

$$R = \frac{n(\sum XY) - (\sum X).(\sum Y)}{\sqrt{n(\sum X^2) - (\sum X)^2}\ \sqrt{n(\sum Y^2) - (\sum Y)^2}}$$

where,

n = Number of observations

x = Measures of Variable 1

y = Measures of Variable 2

Σxy = Sum of the product of respective variable measures

Σx= Sum of the measures of Variable 1

Σy = Sum of the measures of Variable 2

Σx^2 = Sum of squared values of the measures of Variable 1

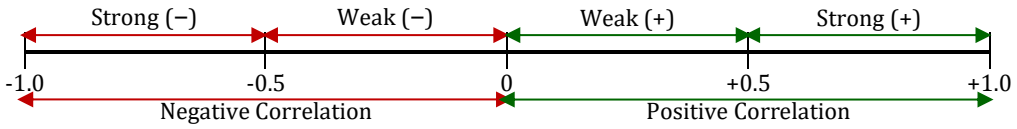
Σy^2 = Sum of squared values of the measures of Variable 2

Based on the direction, the degree of correlative measure can be categorised as Positive, Zero or Negative correlation. Generally, it is rare in application to get exact zero correlation coefficient between variables, and therefore, positive and negative correlations can be the identical categorisation in analyses.

If the trend of a variable is positive and almost similar to another variable, there may be possibility to have positive association of each other and such association can provide positive correlation coefficient; and If the trend of a variable is positive and almost negative to another variable, there may be possibility to have negative association of each other and such association can result in negative correlation coefficient.

Fundamentally, the coefficient of correlation R will range between -1 and +1, i.e., $-1 \leq R \leq +1$. There is no a specific way of interpreting the correlation coefficient. According to Gogtay and Thatte (2017), by measure, the correlation coefficient can be interpreted based on its value as shown in Figure 1.

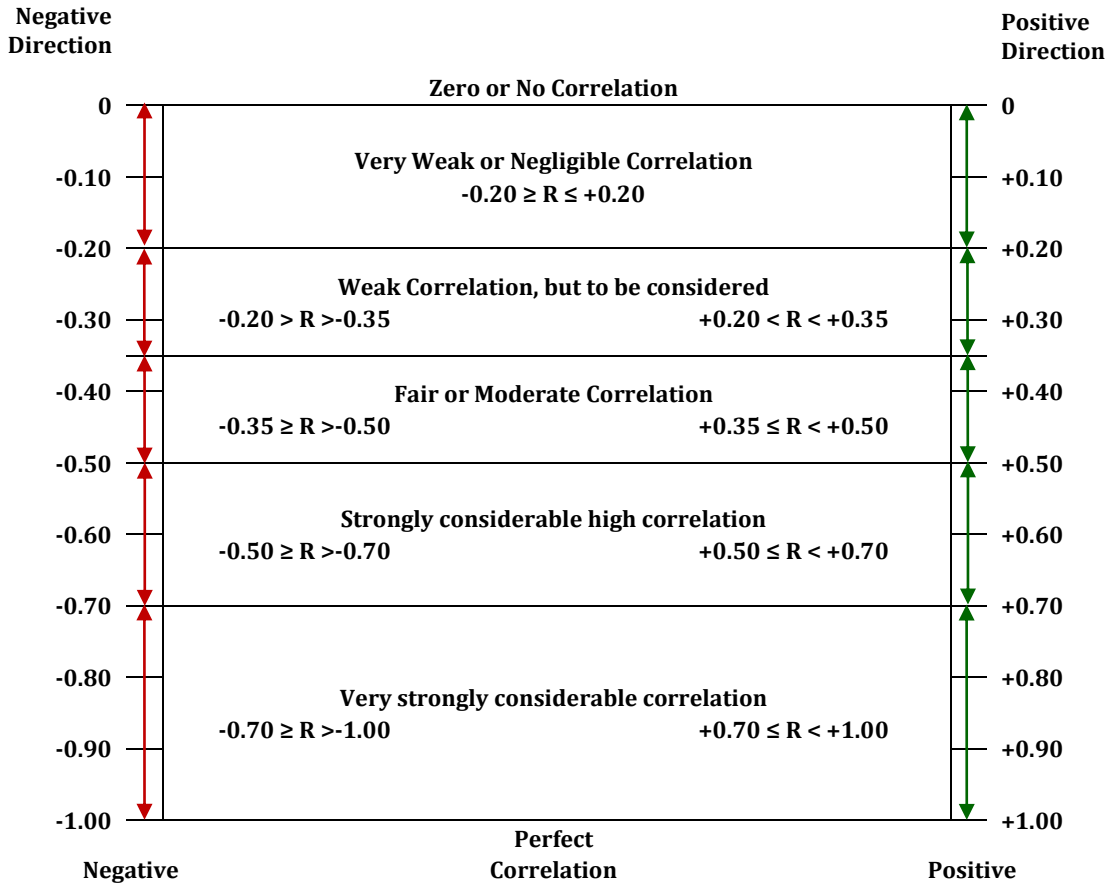
Figure 1: Basic spectrum of interpreting correlation coefficient



(Source: Gogtay and Thatte, 2017, p. 79)

However, in social science studies, when R ≥ 3.5 or R ≤ -3.5 and statistically significant, such a measure can be considered as reasonable correlation, since qualitative measures of social variables are not consistent, but frequently fluctuate. Therefore, this paper suggests the following ideal range as a basic spectrum of interpreting correlation coefficient in social science studies (see Figure 2).

Figure 2: Ideal spectrum of interpreting correlation coefficient in social science studies



As correlation implies associative relationship of two variables, there is still possibility that a study can wrongly conceptualize correlation as a causation effect. Studies should therefore pay attention how to interpret correlation coefficient. Correlation explores the type (positive, negative or none) and degree of association (magnitude of closeness) between two variables. Correlation never provide information on what is the relationship between them. Consider, for example, the activities of a father and a son. Their activities are independent. However, son might have learnt from farther to set his activities, may be positively, negatively or irrelevantly. If the activities of father and son appear to be similar, then there is positive correlation; if the activities of father and son appear to be not similar, but opposite, then there is negative correlation; or if the activities of father and son appear to be not similar and completely different, then there is no possibility for correlation. Notably, the correlation can give only the idea about the direction (positive, negative or none) of their activities, not about their relationship (father and son) and not about their causation effect (son acts because of father).

3. USE OF LINEAR CORRELATION COEFFICIENT

In a common practice of research and analyses, linear correlation coefficient is mostly used to explore the degree of association, level of multicollinearity and reflecting mediating/moderating status between two variables. These can be categorically explained for a simple understanding purpose.

3.1 Linear Relationship between Two Variables and Coefficient of Determination R^2

The degree of association can be between dependent and independent variables, or between two independent variables. If the correlation coefficient is determined for a degree of relationship between dependent and independent variables, their significant relationship can be useful to provide linear regression model utility to predict dependent variable with the independent variable. Correlation becomes significant here, since higher value of correlation coefficient represents better prediction of dependent variable with lowest possible errors.

The significant relationship between the dependent and independent variable can be confirmed with their significant linear correlation coefficient. In this context, the squared value of such a correlation coefficient (known as ***Coefficient of Determination R^2***) is the measure that gives the validity of prediction of dependent variable with the independent variable, i.e., the Coefficient of Determination R^2 provides information that the explained value of dependent variable provides how much accuracy with respect to the independent variable. In other words, how much variation of dependent variable is explained by the variation of dependent variable.

For instance, consider a linear regression model of dependent (y) and independent (x) variables: $y = 0.84x + 0.17$ and this relationship provides correlation coefficient R = 0.85 between the variables.[2] As per the criteria in Figure 2, the degree of association between the variables is very strong (R > 0.7), thus giving coefficient of determination $R^2 = 0.72$.

Though correlation coefficient is well known and meaningful, the coefficient of determination is much more useful to explore the degree of usefulness of linear regression models. As per the example, coefficient of determination implies that variation of independent variable *x* explains 72% variation of dependent variable *y*; and the accuracy of predicting *y* with *x* is aligned in the same context. In other term, independent variable *x* contributes to explain/predict dependent variable *y* with 72% (0.72) accuracy. Further, the unexplained variation (about 28%) of dependent variable depends on other variable(s), thus implying prediction accuracy lacks such 28%.

3.2 Multicollinearity

In a multiple-regression model, when one explaining variable is identical with another independent variable and they produce high strong correlation coefficient, such relationship variables are known to have multicollinearity. Generally, this type of variables consists of similar information to predict the dependent variable, thus causing duplication in of similar variables in the model.

There are basically two types of multicollinearity: (a) data-based multicollinearity and (b) structure-based multicollinearity. When entire data are collected from observations and the data collection methods have no provision for data manipulation, there is significant possibility to have highly possible inter-related variables with high correlation. This is not a scholar's error, since the data collection design has caused such a high correlation. On the other hand, when researcher adds additional explaining variable in the model, there is a possibility of having high correlated variables; and this is known as structure-based (caused by the researcher) multicollinearity (Glen, 2015).

As Glen (2015) indicates, inappropriate use of dummy variables, considering another independent variable in the model that is measured as manipulating measures in the dataset, and accommodating an independent variable that contains almost similar information to explain the dependent variable are the main applications that possibly leads to high level of multicollinearity

[2] Model information is adopted from Smith, et. al. (1986).

between independent variables. Because of these improper applications in studies, occurring multicollinearity is considered as a problem of explaining dependent variable.

Multicollinearity provide tediousness to explore right predicting variable in a study, since partial regression coefficient cannot be estimated accurately; and the standard errors becomes high. This confirms dilution of information contained in the independent variables. Thus, it is important to be away from multicollinearity to occur in analytic regression model; and studies measure variance inflation factor (VIF) to assess level of multicollinearity between independent variables.

VIF is the measure of detecting level of multicollinearity between the predictors. Correlation coefficient is the measure used to determine the VIF for the predictors, i.e.,

$$VIF = \frac{1}{(1 - R^2)}$$

where, R = Correlation coefficient between the predictors

The criteria for interpreting VIF are:

Criteria		**Interpretation on multicollinearity**
VIF = 1	(implies no correlation)	Null
1 < VIF < 5	(correlated)	Low level
5 ≤ VIF < ∞	(Highly correlated)	High level
VIF → ∞	(Perfectly correlated)	Perfect Multicollinearity

In fact, high level of multicollinearity (VIF ≥ 5) becomes possible, approximately when R ≥ 0.9 (for positively correlated predictors) or R ≤ -0.9 (for negatively correlated predictors). Explicitly, VIF depends on the value range of R (≥ 0.9 or ≤ -0.9 for high level of multicollinearity), interpretation on multicollinearity can be possible with the correlation coefficient of the predictors. Therefore, the formula for VIF is a sign transferring process to compromise the plus (+) and minus (-) signs of correlation (see Appendix 1 for details).

3.3 Reflection of Mediating/Moderating Variable

In case of mediating/moderating variable, regression model should have at least three variables, including dependent variable. Assume that a model consists of two independent variables ("A" and "B") and a dependent variable "C". In both cases of identifying mediating/moderating variable, all variables ("A", "B" and "C") should have significant correlation to each other.

If variable "B" act as a moderating variable to transform the information of variable "A" to predict variable "C", the coefficient estimate of variable "B" in the regression model should be significant, while the coefficient estimate of variable "A" is insignificant at the required significant level (normally $p<0.05$). However, the both variables "A" and "B" should have significant correlative relationship with dependent variable "C". On the other hand, if variable "B" act as a mediating variable to transform some of the information of variable "A" to predict variable "C", the coefficient estimates of both variables "A" and "B" should be significant at a required level, normally $p<0.05$ (refer to Senthilnathan, 2017, for a detailed illustration).

4. CONCLUDING REMARKS

A simple correlation analysis represents measures the degree of closeness between two related variables. The correlation coefficient (r or R) as a measure provides information about closeness of the two variables. Irrespective of non-linear correlation, this paper mainly focuses on the linear correlation analysis, as it is mostly used in social science studies.

As the purpose of carrying out correlation analysis is almost the same in quantitative analytical studies, such a correlation analysis becomes useful to explore the association between independent and dependent variables. Thus, as an extension, this paper attempts additionally to explain the usefulness of linear correlation coefficient between two variables in the context of identifying the level of multicollinearity and mediating/moderating status of independent variables in a model. This paper demonstrates how the level of multicollinearity can be explored by using correlation coefficient between two independent variables in a regression model.

REFERENCES

Denis, J. D. (2001), The origins of correlation and regression: Francis Galton or Auguste Bravais and the error theorists?, ***History and Philosophy of Psychology Bulletin***, 13, pp. 36-44.

Gogtay, N. J. and Thatte, U. M. (2017), Principles of correlation Analysis, ***Journal of The Association of Physicians of India***, 65 (March), pp. 78-81.

Glen, S. (2015), Multicollinearity: Definition, Causes, Examples, https://www.statisticshowto.datasciencecentral.com/multicollinearity/, Retrieved: 30-05-2019, 8.56 am, New Zealand.

Hauke, J. and Kossowski, T. (2011), Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data, ***Questiones Geographicae***, 30(2), pp. 87-93 (doi: http://dx.doi.org/10.2478/v10117-011-0021-1).

Senthilnathan, S. (2017), Relationships and Hypotheses in Social Science Research, https://ssrn.com/abstract=3032284 or http://dx.doi.org/10.2139/ssrn.3032284, Retrieved: 09-07-2019, 12.51 pm, New Zealand.

Smith, M. D., Handshoe, R., Handshoe, S., Kwan, O. L., and Demaria, A. N. (1986), Comparative accuracy of two-dimensional echocardiography and Doppler pressure half-time methods in assessing severity of mitral stenosis in patients with and without prior commissurotomy, Circulation, 73(1)-Jan, pp. 100-107.

Appendix 1: VIF – a formula to compromise both plus (+) and minus (-) signs of correlation

Consider VIF formula

$$VIF = \frac{1}{(1 - R^2)}$$

where, R = Correlation coefficient between the predictors

A high level of multicollinearity is demonstrated, when VIF ≥ 5.

Therefore,

$$VIF = \frac{1}{(1 - R^2)} \geq 5$$

$$\frac{1}{5} \geq (1 - R^2) \rightarrow R^2 \geq \left(1 - \frac{1}{5}\right) \rightarrow R^2 \geq 0.80$$

Just for understanding, consider an approximation of $R^2 = 0.80$ as $R^2 = 0.81$ and rewrite the above,

$$R^2 \geq 0.81$$

This implies $R \geq 0.9$ or $R \leq -0.9$.

Above range of R values indicate that ±0.9 is the benchmark to interpret level of multicollinearity between independent variables.