Principles of Correlation and Regression Analysis

Shivam Pandey

Department of Biostatistics, All India Institute of Medical Sciences, New Delhi, India

Abstract

In statistics, the correlation analysis quantifies the strength of the association between two quantitative variables. In the present article, we discuss how to establish a relationship or an association between two quantitative variables. While the correlation provides a quantitative way of measuring the degree or strength of a relation between two variables, regression analysis describes this relationship. Correlation and agreement are also distinguished, and pitfalls of correlation are discussed.

Keywords: Agreement, correlation, regression

Introduction

In statistics, the correlation analysis quantifies the strength of the association between two quantitative variables.[1] In the present article, we discuss how to establish a relationship or an association between the two quantitative variables. While the correlation provides a quantitative way of measuring the degree or strength of a relation between two variables, regression analysis describes this relationship. The correlation and regression analysis are therefore related concepts. The present article provides the concept behind some commonly used correlation coefficients and when they should be used, some pitfalls of correlation analysis and how correlation can be linked to regression analysis.

The Scatter Plot

When exploring the relationship between two numerical variables, the first and essential step is to graphically depict the relationship on a scatter plot or scatter diagram or scattergram. This is simply a bivariate plot of one variable against the other. Before plotting, one or both variables may be logarithmically transformed to obtain a more normal distribution. On a scatter diagram, it is customary to plot the independent variable on the X-axis and the dependent variable on the Y-axis. However, "independent" and "dependent" distinction can be confusing at times. For instance, if we are exploring the relationship between age and stature in children, it is reasonable to assume that age is the "independent" variable on which the height depends. Hence customarily, we will plot age on the X-axis and height on the Y-axis. However, if we are exploring the relationship between serum potassium and venous plasma glucose levels, which variable do we treat as the "dependent" variable? In such cases, it usually does not matter which variable is attributed to a particular axis of the scatter diagram. Usually, the dependent variable is plotted on the Y-axis, and the independent variable is plotted on the X-axis.

The Correlation Coefficient

To quantify the strength of the relationship between two variables shown to have a linear relationship on the scatter plot, we calculate the correlation coefficient. The coefficient takes values only between −1 and +1, with the numerical magnitude depicting the strength of the relationship, and the sign indicating its direction. Thus, the sign accompanying a correlation coefficient is not a + or − sign in the arithmetic sense. Rather the plus sign denotes a direct relationship, whereas minus denotes an inverse relationship. If both variables x and y are normally distributed, we calculate Pearson's product moment

Access this article online

Quick Response Code:

Website:
www.j-pcs.org

DOI:
10.4103/jpcs.jpcs_2_20

Address for correspondence: Dr. Shivam Pandey,
Department of Biostatistics, All India Institute of Medical Sciences,
New Delhi, India.
E-mail: shivam.pandey2809@gmail.com

Submitted: 09-Jan-2020 **Revised:** 23-Jan-2020
Accepted: 21-Feb-2020 **Published:** 17-Apr-2020

For reprints contact: reprints@medknow.com

How to cite this article: Pandey S. Principles of correlation and regression analysis. J Pract Cardiovasc Sci 2020;6:7-11.

correlation coefficient r or Pearson's correlation coefficient r, or simply *r* (after Karl Pearson). It is calculated as the covariance of the two variables divided by the product of their standard deviations (SDs). A value of r close to +1 indicates a strong direct linear relationship (i.e., one variable increases with the other). A value close to −1 indicates a strong inverse linear relationship (i.e., one variable decreases as the other increases). A value close to 0 indicates a random scatter of the values; alternatively, there could be a nonlinear relationship between the variables. The scatter plot is used in checking the assumption of a linear relationship and it is meaningless to calculate a correlation coefficient without such a relation between the two variables. In between the state of "no correlation at all" ($r = 0$) and "perfect correlation" ($r = 1$), interim values of the correlation coefficient are interpreted by convention. Thus, values >0.7 may be regarded as "strong" correlation; values between 0.5 and 0.7 may be interpreted as "good" correlation; between 0.3 and 0.5 may be treated as "fair" or "moderate" correlation; and any value <0.30 would be a poor correlation.

If one or both variables in a correlation analysis are not normally distributed, a rank correlation coefficient that depends on the rank order of the values rather than the actual observed values can be calculated. The examples include Spearman's rho (ρ) (after Charles Edward Spearman) and Kendall's tau (τ) (after Maurice George Kendall) statistics. Essentially, Spearman's rank correlation coefficient rho, which is the more frequently used nonparametric correlation, is simply Pearson's product-moment correlation coefficient calculated for the rank values of *x* and *y* rather than their actual values. It is also appropriate to use ρ rather than r when at least one variable is measured on an ordinal scale or when the sample size is small (say $n \leq 10$).[2] Although less often used, Kendall's tau is another nonparametric correlation offered by many statistical packages. Some statisticians recommend that it should be used, rather than Spearman's coefficient, when the data set is small with the possibility of a large number of tied ranks. This means that if we rank all of the scores and many scores have the same rank, Kendall's tau should be used.

Point Biserial and Biserial Correlation

The point-biserial correlation is a special case of the product-moment correlation, in which one variable is continuous, and the other variable is binary. The point-biserial correlation coefficient measures the association between a binary variable X, taking values 0 or 1, and a continuous numerical variable Y. It is assumed that for each value of X, the distribution of Y is normal, with different means but the same variance. It is often abbreviated as rPB (Point Biserial Correlation). The binary variable frequently has categories such as yes or no, present or absent, and success or failure. If the variable *x* is not naturally dichotomous but is artificially dichotomized, we calculate the biserial correlation coefficient rB instead of the point-biserial correlation coefficient. Although not often used, an example where we may apply the point-biserial correlation coefficient would be in cancer studies. How strong is the association between administering the anticancer drug (active drug vs. placebo) and the length of survival after treatment? The value would be interpreted in the same way as Pearson's *r*. Thus, the value would range from −1 to +1, where −1 indicates a perfect inverse association, +1 indicates a perfect direct association, and 0 indicates no association at all.

Factors that Affect a Correlation Analysis

Several factors must be considered when a correlation analysis is planned. Correlation analysis should not be used when the data are the repeated measures of the same variable from the same individual at the same or varied time points. For example, if you have measured pain scores in patients with rheumatoid arthritis at monthly intervals over 6 years in a study, it is inappropriate to find out a correlation coefficient for these data. Second, outliers should be checked for when calculating the correlation coefficients. An outlier is essentially an infrequently occurring value in the data set. It is important to remember that even a single outlier can dramatically alter the correlation coefficient. Third, if there is a nonlinear relationship between the quantitative variables, correlation analysis should not be performed. For example, during the growth phase in adolescence, there would a linear relationship between height and weight, as both increases. However, this relationship ceases once a person enters adulthood. Fourth, if the dataset has two distinct subgroups of individuals whose values for one or both variables differ considerably from each other, a false correlation may be found, when none may exist. An example given by Aggarwal and Ranganathan illustrates this point well.[3] If you were to plot heights (on X-axis) and hemoglobin levels (on Y-axis), of a group of men (n = 20) and women (n = 20), most women may end up in the left lower corner (shorter and lower hemoglobin) and most men in the right upper corner (taller and higher hemoglobin). The analysis would suggest a relationship with a positive "*r*" value between height and hemoglobin levels. Fifth, if one data set forms part of the second data set, for example, weight at age 12 (X-axis) and weight at age 30 (Y-axis), we would expect to find a positive correlation between them because the second quantity is a subset of the first quantity.

Assessing Agreement

In the past, correlation has been used to assess this degree of agreement between the sets of paired measurements. However, correlation quantifies the relationship between numerical variables and has limitations if used for assessing the comparability between the methods. Two sets of measurements would be perfectly correlated if the scatter diagram shows that they all lie on a single straight line, but they are not likely to be in perfect agreement unless this line passes through the origin. It is very likely that two tests designed to measure the same variable would return figures that would be strongly correlated, but that does not automatically mean that the repeat measurements are also in strong agreement. Data which seem to be in poor agreement can produce quite high correlations. In addition, a change in the scale of measurement does not affect the correlation, but it can affect the agreement. There are several other ways to assess the agreement. These include the following:

Bland–Altman Plot

Bland and Altman devised a simple but informative graphical method of comparing the repeat measurements. When the repeat measurements have been taken on a series of subjects or samples, the difference between the pairs of measurements (Y-axis) is plotted against the arithmetic mean of the corresponding measurements (X-axis). The resulting scatter diagram is the Bland–Altman plot (after John M. Bland and Douglas G. Altman, who first proposed it in 1983, and then, popularized it through a Lancet paper in 1986) an example of which is given in Figure 1.[4,5] The repeat measurements could represent the results of two different assay methods or scores from the same subjects by two different raters. Usually, a “bias” line parallel to the X-axis is added. This represents the difference between the means of the two sets of measurements. Lines denoting 95% limits of agreement (mean difference ± 1.96 SD of the difference) can be added on either side of the bias line. Usually, three points need to be considered while interpreting these plots:

The magnitude of the average discrepancy between the methods, which is indicated by the position of the bias line. If the differences within mean ±1.96 SD are not clinically important, the two methods may be used interchangeably.

Whether the scatter around the line of bias is too much, with a number of points falling outside the 95% agreement limits.

Whether the difference between the methods tends to get larger or smaller as the values increase. If it does, it indicates the existence of a proportional bias which means that the methods do not agree equally through the range of measurements.

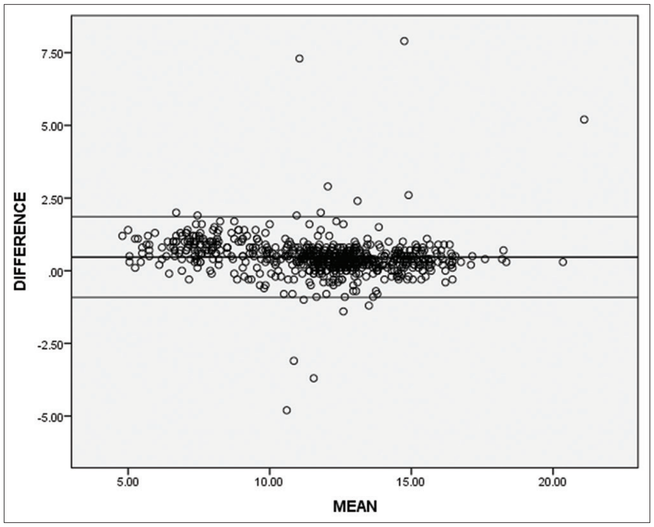


Figure 1: Bland–Altman plot.

Intraclass Correlation Coefficient

A numerical way to assess the agreement is the intraclass correlation coefficient (ICC). This was originally introduced in genetics to judge sibling correlations, but now the ICC statistic is now most often used to assess the consistency or conformity of measurements made by multiple observers measuring the same parameter or two or more raters scoring the same set of subjects. The methods of ICC calculation have evolved over time. The earliest work on intraclass correlations focused on paired measurements, and the first ICC statistics to be proposed was the modifications of the Pearson’s correlation coefficient calculations. The calculation of ICC is now based on the true (between subjects) variance and variance of the measurement error (during repeat measurement). It takes a value between 0 and 1. Complete interrater agreement is indicated by a value of 1, but this is seldom achieved. Arbitrarily, the agreement boundaries proposed are as follows: <0.40: poor, 0.40–0.60: fair, 0.60–0.74: good, and >0.75: strong. Software may report two coefficients with their respective 95% confidence intervals. ICC for single measures is an index for the reliability of the multiple ratings by a single typical rater. ICC for average measures is an index for the reliability of different raters averaged together. This ICC is always slightly higher than the single measures ICC.

Linear Regression

Correlation analysis is seldom used alone and is usually accompanied by the regression analysis. The difference between the correlation and regression lies in the fact that while a correlation analysis stops with the calculation of the correlation coefficient and perhaps a test of significance, a regression analysis goes ahead to expresses the relationship in the form of an equation and moves into the realm of prediction. Furthermore, correlation coefficients do not give the information about whether one variable moves in response to another. There is no attempt to establish one variable as “dependent” and the other as “independent.” This is done in a regression analysis. In simple linear regression, the value of one variable (x) is used to predict the value of the other variable (y) by means of a simple mathematical function, the linear regression equation, which quantifies the straight-line relationship between the two variables. This straight line, or regression line, is actually the “line of best fit” for the data points on the scatter plot showing the relationship between the variables in question. The reader may wonder why the statistical procedure of fitting a line is called “regression” which in common usage means “going backward.” Interestingly, the term was used neither by Legendre or Gauss but is attributed to the English scientist Francis Galton who had a keen interest in heredity. In Victorian England, Galton measured the heights of 202 fathers and their first born adult sons and plotted them on a graph of median height versus height group. The scatter for fathers and sons approximated to two lines that intersected at a point representing the average height of the adult English population. Studying this plot, Galton made the very interesting observation that tall fathers tend to have tall sons but they are not as tall as their fathers, and short fathers tend to have short sons but they are not as short as their fathers, and in the course of just two or three generations, the height of individuals tended to go back or "regress" to the mean population height. He published a famous article titled "Regression toward mediocrity in hereditary stature."[6] This phenomenon of regression to the mean can be observed in many biological variables.

The regression line has the general formula: Y = a + bx, where "a" and "b" are two constants denoting the intercept of the line on the Y-axis (y-intercept) and the gradient (slope) of the line, respectively. The other name for b is the "regression coefficient." Physically, "b" represents the change in *y*, for every 1 unit change in *x*, while "a" represents the value that *y* would take, if *x* is 0. Once the values of a and b have been established, the expected value of *y* can be predicted for any given value of *x*, and vice versa. Thus, a model for predicting *y* from *x* is established. There may be situations, in which a straight line passing through the origin will be appropriate for the data, and in these cases, the equation of the regression line simplifies to y = bx. But how do we fit a straight line to a scattered set of points which seem to be in linear relationship? If the points are not all on a single straight line, we can, by eye estimation, draw multiple lines that seem to fit the series of data points on the scatter diagram. But which is the line of best fit? The solution was in the form of the method of least squares, which was first published by the French mathematician Adrien-Marie Legendre in 1805 but used earlier by Carl Friedrich Gauss in Germany in 1795.[7] There are several types of regression analysis [Table 1].

Learning with an Example

The hypothetical data correspond to a medical experiment during which the concentration of an antibody is measured for 8 mice submitted to 8 different doses of a new molecule being tested. For each mouse, a blood sample has been taken and divided into four homogeneous subsamples. Two methods are being tested each on 2 of the 4 subsamples [Table 2]. The first method is currently considered the reference, but it is much more expensive than the second and new method. Our goal is to check if it is possible to use the new method instead of the reference one. Further with reference to this example, more than 50% of the values lie outside the limit which indicates that there is a fair agreement between the tests [Figure 2]. This analysis can be done through a freely available excel add on "Analyze-it" (https://analyse-it.com). The option for Bland Altman Plot can be accessed, as shown in Figure 3. Correlation and regression can be done through a "Analysis ToolPak" add-on from the "option" tab within the Excel [Figure 4]. The regression coefficient (labeled "Coefficients" in Figure) quantifies the average change in systolic blood pressure (SBP) (measured in mmHg) per unit change in body mass index (BMI) (measured in kg/m.sq). The heading labeled "R-square" signifies the proportion of variation in the dependent (i.e., SBP) variable explained by the independent (i.e., BMI) variable.

Conclusion

In summary, the correlation coefficients are used to assess the strength and direction of the linear relationships between the pairs of continuous variables. When both variables are normally distributed a Pearson correlation is used while for nonnormal data or for low sample sizes, spearman correlation is used. Correlation does not imply agreement and there are ranges of other statistical methods which have been discussed to assess the agreement between two raters. Correlation analysis is seldom used alone and is usually accompanied by the regression analysis, as regression adjusts for multiple variables together. The other difference between the correlation and regression lies in the fact that while a correlation analysis stops with the calculation of the correlation coefficient and perhaps a test of significance, a regression analysis goes ahead to expresses the relationship in the form of an equation

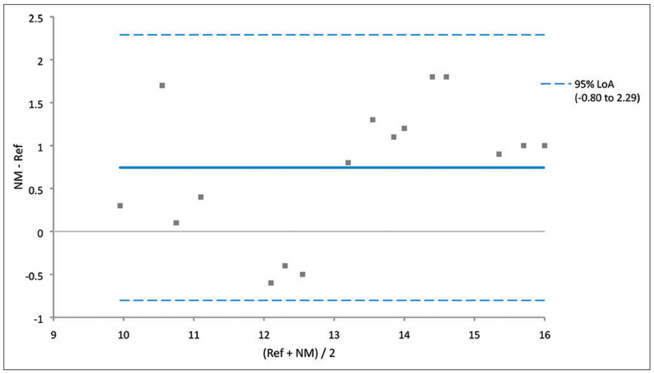


Figure 2: Bland–Altman plot in Excel.

Table 1: Types of regression

Type of regression	Dependent variable and its nature	Independent variable and its nature	Relationship between the variables
Simple linear	One, continuous, normally distributed	One, continuous, normally distributed	Linear
Multiple linear	One, continuous	Two or more, may be continuous, or categorical	Linear
Logistic	One, binary	Two or more, may be continuous, or categorical	Linearity not required
Polynomial (logistic) also known as multinomial	Nonbinary	Two or more, may be continuous, or categorical	Linearity not required
Cox-proportional hazards regression	Time to an event	Two or more, may be continuous, or categorical	Nonlinear

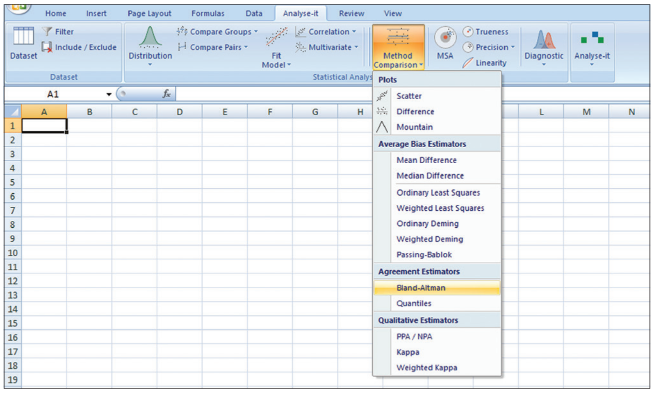


Figure 3: Option for Bland–Altman analysis in Excel (analyze-it add-on).

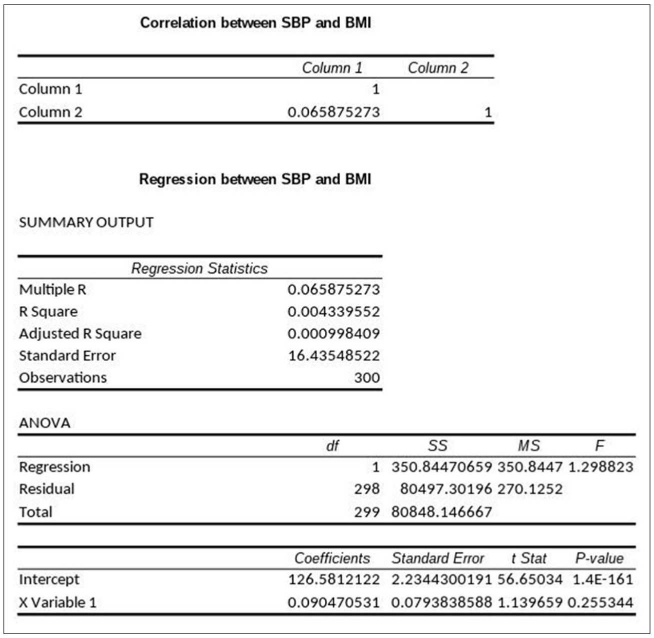
Table 2: Hypothetical data for Bland-Altman analysis

ID	Rater	Reference	New method
1	1	9.8	10.1
1	2	9.7	11.4
2	1	10.7	10.8
2	2	10.9	11.3
3	1	12.4	11.8
3	2	12.5	12.1
4	1	12.8	12.3
4	2	12.8	13.6
5	1	12.9	14.2
5	2	13.3	14.4
6	1	13.4	14.6
6	2	13.5	15.3
7	1	13.7	15.5
7	2	14.9	15.8
8	1	15.2	16.2
8	2	15.5	16.5

and explores the concept of prediction through a regression equation.

Financial support and sponsorship

Nil.



Correlation between SBP and BMI

	Column 1	*Column 2*
Column 1	1	
Column 2	0.065875273	1

Regression between SBP and BMI

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.065875273
R Square	0.004339552
Adjusted R Square	0.000998409
Standard Error	16.43548522
Observations	300

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	350.84470659	350.8447	1.298823
Residual	298	80497.30196	270.1252	
Total	299	80848.146667		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	126.5812122	2.2344300191	56.65034	1.4E-161
X Variable 1	0.090470531	0.0793838588	1.139659	0.255344

Figure 4: Correlation and regression analysis in Excel.

Conflicts of interest

There are no conflicts of interest.

References

1. Deshpande S, Gogtay NJ, Thatte UM. Data types. J Assoc Phy Ind 2016;64:64-5.
2. David FN. Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples. Cambridge: Cambridge University Press; 1938.
3. Aggarwal R, Ranganathan P. Common pitfalls in statistical analysis: The use of correlation techniques. Perspect Clin Res 2016;7:187-90.
4. Altman DG, Bland JM. Measurement in medicine: The analysis of method comparison studies. Statistician 1983;32:307-17.
5. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
6. Galton F. Regression towards mediocrity in hereditary stature. J Anthropol Inst 1886;15:246-63.
7. Gauss CF. Theoria combinationis obsevationum erroribus Minimis Obnoxiae. Vol. 4. Göttingen, Germany: Werke; 1823.