International Journal of Research in Engineering, Science and Management
Volume-2, Issue-7, July-2019
www.ijresm.com | ISSN (Online): 2581-5792

Correlation and Regression Analysis

Manisha Puniya[1], R. B. Singh[2]
[1]*M.Sc. Student, Department of Mathematics, Monad University, Hapur, India*
[2]*Professor, Department of Mathematics, Monad University, Hapur, India*

***Abstract*: The present review introduces methods of analyzing the relationship between two quantitative variables. The calculation and interpretation of the sample product moment correlation coefficient and the linear regression equation are discussed and illustrated. Common misuses of the techniques are considered. Tests and confidence intervals for the population parameters are described, and failures of the underlying assumptions are highlighted.**

***Keywords*: Scatter Diagram, Karl Pearson's Coefficient of Correlation, Rank Correlation, Probable Error and Probable Limits**

1. Introduction

Correlation analysis is applied in quantifying the association between two continuous variables, for example a dependent and independent variable or among two independent variables. Regression analysis refers to assessing the relation between the outcome variable and one or more variables. The outcome variable is known as dependent or response variable and the risk elements, and cofounders are known as predictors or independent variables. The dependent variable is shown by "y" and independent variables are shown by "x" in regression analysis.

The sample of a correlation coefficient is estimated in the correlation analysis. It ranges between -1 and +1, denoted by r and quantifies the strength and direction of the linear association among two variables. The correlation among two variables can either be positive, i.e., a higher level of one variable is related to higher level of another) or negative, i.e., a higher level of one variable is related to lower level of the other.

The sign of the coefficient of correlation shows the direction of association. The magnitude of coefficient shows the strength of association.

2. Scatter Diagrams

Scatter Diagrams are convenient mathematical tools to study the correlation between two random variables. As the name suggests, they are a form of a sheet of paper upon which the data points corresponding to the variables of interest, are scattered. Judging by the shape of the pattern that the data points form on this sheet of paper, we can determine the association between the two variables, and can further apply the best suitable correlation analysis technique.

3. Karl Pearson's Coefficient of Correlation

There are many situations in our daily life where we know from experience, the direct association between certain variables but we can't put a certain measure to it. For example, you know that the chances of you going out to watch a newly released movie is directly associated with the number of friends who go with you because the more the merrier. Karl Pearson's Coefficient of Correlation is widely used mathematical method wherein the numerical expression is used to calculate the degree and direction of the relationship between linear related variables.

Pearson's method, popularly known as a Pearsonian Coefficient of Correlation, is the most extensively used quantitative methods in practice. The coefficient of correlation is denoted by "r".

4. Rank Correlation

In statistics, a rank correlation is any of several statistics that measure an ordinal association-the relationship between rankings of different ordinal variables or different rankings of the same variable, where a "ranking" is the assignment of the ordering labels "first", "second", "third", etc. to different observations of a particular variable. A rank correlation coefficient measures the degree of similarity between two rankings, and can be used to assess the significance of the relation between them. For example, two common nonparametric methods of significance that use rank correlation are the Mann–Whitney U test and the Wilcoxon signed-rank test.

5. Provable error and Probable Limits

Probable Error is basically the correlation coefficient that is fully responsible for the value of the coefficients and its accuracy. Let's dig in deeper to know about the concepts of probable error and probable limits in a better way.

As mentioned, probable error is the coefficient of correlation that supports in finding out about the accurate values of the coefficients. It also helps in determining the reliability of the coefficient The calculation of the correlation coefficient usually takes place from the samples.

These samples are in pairs. The pairs generally come from a very large population. It is quite an easy job to find out about the limits and bounds of the correlation coefficient. The

correlation coefficient for a population is usually based on the knowledge and the sample relating to the correlation coefficient. Therefore, probable error is the easy way to find out or obtain the correlation coefficient of any population. Hence, the definition is:

$$\text{Probable Error} = 0.674 \times 1-r2\sqrt{N}$$

Here, r = correlation coefficient of 'n' pairs of observations for any random sample and N = Total number of observations.

6. Probable Limit

To get the upper limit and the lower limit, all we need to do is respectively add and subtract the value of probable error from the value of 'r.' This is exactly where the value of correlation of coefficient lies.

$$\rho \text{ (rho)} = r \pm \text{P.E.}$$

Here, the value of rho is nothing but the correlation coefficient of a population. This is also the limit of the correlation of coefficient. Alongside,

$$\text{Probable Error} = 2/3 \text{ SE}$$

Here, S.E is Standard Error of Correlation Coefficient

$$\text{Standard Error} = (1-r2)/\sqrt{N}$$

Standard Error is basically the standard deviation of any mean. It is the sampling distribution of the standard deviation. The standard error is generally used to refer to any sort of estimate belonging to the standard deviation. Therefore, we use probable error to calculate and check the reliability associated with the coefficient.

EXAMPLES

Question1: Find the probable error. Assume that the correlation coefficient is 0.8 and the pairs of samples are 25.

Solution: We will use the most common method to calculate the outcome of the following. Here, r = 0.8 and n = 25. We know that,

$\text{Probable Error} = 0.674 \times 1-r2\sqrt{N}$

So, on putting the values:

$\text{Probable Error} = 0.674 \times \{(1 – (0.8)2)/\sqrt{25}\}$

$= 0.674 \times \{(1 – 0.64)/5\}$

$= 0.674 \times (0.36/5)$

≈ 0.0486

Therefore, the probable error is: 0.0486.

Question 2: If the value of r = 0.7 and that of n = 64, then find the P. E. of the correlation of coefficient. Furthermore, find the limits for the population correlation coefficient.

Solution: Here, we have to calculate the probable error. Given, r = 0.7 and n = 64. We know that,

$\text{P. E.} = 0.674 \times \{(1\text{-}r2)/\sqrt{N}\}$

$= 0.6745 \times \{(1 – (0.7)2)/\sqrt{64}\}$

$= 0.6745 \times 0.06375$

≈ 0.043

Therefore, the P.E. is 0.043. Now, we have to calculate the limits of the population correlation coefficient. We use the formula,

Probable Limit- ρ (rho) = $r \pm$ P.E.r

Hence, we get, (0.7 ± 0.043) i.e. (0.743, 0.657).

7. Result

Correlation is a measure of association between two variables. The variables are not designated as dependent or independent. The two most popular correlation coefficients are: Spearman's correlation coefficient rho and Pearson's product-moment correlation coefficient.

When calculating a correlation coefficient for ordinal data, select Spearman's technique. For interval or ratio-type data, use Pearson's technique.

The value of a correlation coefficient can vary from minus one to plus one. A minus one indicates a perfect negative correlation, while a plus one indicates a perfect positive correlation. A correlation of zero means there is no relationship between the two variables. When there is a negative correlation between two variables, as the value of one variable increases, the value of the other variable decreases, and vise versa. In other words, for a negative correlation, the variables work opposite each other. When there is a positive correlation between two variables, as the value of one variable increases, the value of the other variable also increases. The variables move together. Simple regression is used to examine the relationship between one dependent and one independent variable. After performing an analysis, the regression statistics can be used to predict the dependent variable when the independent variable is known. Regression goes beyond correlation by adding prediction capabilities.

People use regression on an intuitive level every day. In business, a well-dressed man is thought to be financially successful. A mother knows that more sugar in her children's diet results in higher energy levels. The ease of waking up in the morning often depends on how late you went to bed the night before. Quantitative regression adds precision by developing a mathematical formula that can be used for predictive purposes.

8. Conclusion and Discussion

Correlation and Regression are the two analysis based on multivariate distribution. On the other end, Regression analysis, predicts the value of the dependent variable based on the known value of the independent variable, assuming that average mathematical relationship between two or more variables. When it comes to correlation, there is a relationship between the variables. Regression, on the other hand, puts emphasis on how one variable affects the other. Correlation does not capture

causality, while regression is founded upon it. Correlation between x and y is the same as the one between y and x. Correlation and regression analysis are related in the sense that both deal with relationships among variables. The correlation coefficient is a measure of linear association between two variables. Values of the correlation coefficient are always between -1 and +1. Correlation and linear regression are not the same. Consider these differences: Correlation quantifies the degree to which two variables are related. ... With regression, you do have to think about cause and effect as the regression line is determined as the best way to predict Y from X.

Both correlation and simple linear regression can be used to examine the presence of a linear relationship between two variables providing certain assumptions about the data are satisfied. The results of the analysis, however, need to be interpreted with care, particularly when looking for a causal relationship or when using the regression equation for prediction. Multiple and logistic regression will be the subject of future reviews.

References

[1] Friendly, Michael; Denis, Dan (2005). "The early origins and development of the scatterplot". Journal of the History of the Behavioral Sciences. 41 (2): 103–130.
[2] Visualizations that have been created with Visit at wci.llnl.gov.
[3] Jarrell, Stephen B. (1994). Basic Statistics (Special pre-publication ed.). Dubuque, Iowa: Wm. C. Brown Pub. p. 492. ISBN 978-0-697-21595-6. When we search for a relationship between two quantitative variables, a standard graph of the available data pairs (X,Y), called a scatter diagram, frequently helps.
[4] Utts, Jessica M. Seeing Through Statistics 3rd Edition, Thomson Brooks/Cole, 2005, pp 166-167. ISBN 0-534-39402-7
[5] Cleveland, William (1993). Visualizing data. Murray Hill, N.J. Summit, N.J: At & T Bell Laboratories Published by Hobart Press. ISBN 978-0963488404.
[6] Nancy R. Tague (2004). "Seven Basic Quality Tools". The Quality Toolbox. Milwaukee, Wisconsin: American Society for Quality. p. 15.
[7] Whitley E, Ball J. Statistics review 1: Presenting and summarizing data. Crit Care. 2002; 6:66–71.
[8] Kirkwood BR, Sterne JAC. Essential Medical Statistics. 2. Oxford: Blackwell Science; 2003. [Google Scholar]
[9] Whitley E, Ball J. Statistics review 2: Samples and populations. Crit Care. 2002; 6:143–148.
[10] Bland M. An Introduction to Medical Statistics. 3. Oxford: Oxford University Press; 2001.
[11] Bland M, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986:307–310.
[12] Zar J. H. Biostatistical Analysis. 4. New Jersey, USA: Prentice Hall; 1999.
[13] "Scatter Chart – Any Chart JavaScript Chart Documentation". Any Chart.
[14] Sunil, P.C. Dikshit. Estimation of stature from Hand Length; Journal of Indian Academy of forensic Medicine; 2005; 27(4); 219-221.