International Journal of Social Impact
ISSN: 2455-670X
Volume 5, Issue 2, DIP: 18.02.015/20200502
DOI: 10.25215/2455/0502015
www.ijsi.in |April-June, 2020

Correlation Vs. Regression: A Review

Dr. Arati Shah[1]*

ABSTRACT

When researching the connection between at least two numeric variables, know the distinction between correlation and regression. The likenesses/contrasts and benefits/burdens of these apparatuses are discussed here alongside each instance. Correlation evaluates the course and strength of the connection between two numeric variables, X and Y, and consistently lies between - 1.0 and 1.0. Straightforward direct regression relates X to Y through a condition of the structure Y = a + bX. The review article has shown the difference between correlation and regression.

Keywords: *Correlation, Regression, Variables*

Correlation investigation is applied in evaluating the relationship between two nonstop variables, for instance, a reliant and free factor or among two autonomous variables.

Regression examination alludes to surveying the connection between the result variable and at least one variable. The result variable is known as the ward or reaction variable and the danger components, and fellow benefactors are known as indicators or free variables. The reliant variable is displayed by "y" and autonomous variables are displayed by "x" in regression examination.

Examination Between Correlation and Regression

Basis	**Correlation**	**Regression**
Meaning	A statistical measure that defines co-relationship or association of two variables.	Describes how an independent variable is associated with the dependent variable.
Dependent and Independent variables	No difference	Both variables are different.
Usage	To describe a linear relationship between two variables.	To fit the best line and estimate one variable based on another variable.
Objective	To find a value expressing the relationship between variables.	To estimate values of a random variable based on the values of a fixed variable.

[1]Assistant Professor at JG College of Commerce, Gujarat (India)
*Responding Author

Received: April 23, 2020; Revision Received: May 12, 2020; Accepted: June 25, 2020

THE CORRELATION EQUATION

The word correlation is used in everyday life to denote some form of association. We might say that we have noticed a correlation between foggy days and attacks of wheeziness. However, in statistical terms we use correlation to denote association between two quantitative variables. We also assume that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other. The other technique that is often used in these circumstances is regression, which involves estimating the best straight line to summarise the association.

Correlation coefficient

The degree of association is measured by a correlation coefficient, denoted by r. It is sometimes called Pearson's correlation coefficient after its originator and is a measure of linear association. If a curved line is needed to express the relationship, other and more complicated measures of the correlation must be used.

The correlation coefficient is measured on a scale that varies from + 1 through 0 to – 1. Complete correlation between two variables is expressed by either + 1 or -1. When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative. Complete absence of correlation is represented by 0. Figure 11.1 gives some graphical representations of correlation.

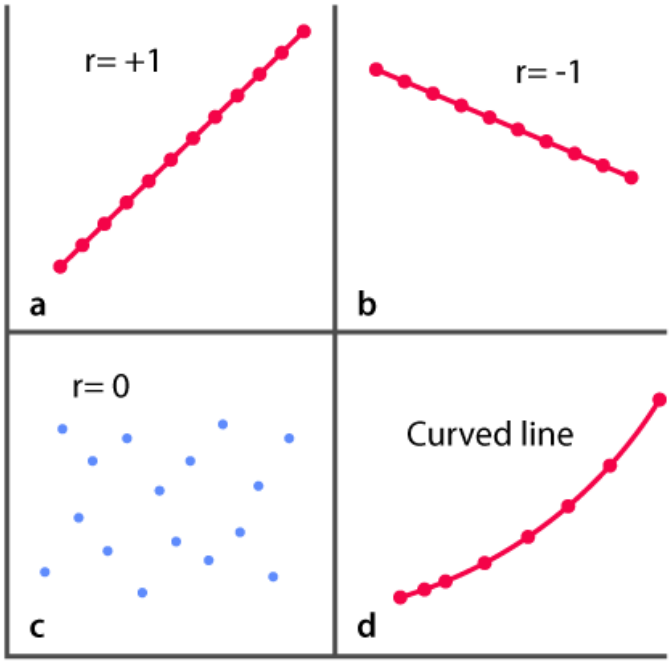

Figure 11.1 Correlation illustrated.
Looking at data: Byjus

When an investigator has collected two series of observations and wishes to see whether there is a relationship between them, he or she should first construct a scatter diagram. The vertical scale represents one set of measurements and the horizontal scale the other. If one set of observations consists of experimental results and the other consists of a time scale or observed classification of some kind, it is usual to put the experimental results on the vertical axis. These represent what is called the "dependent variable". The "independent variable", such as time or height or some other observed classification, is measured along the horizontal axis, or baseline.

The words "independent" and "dependent" could puzzle the beginner because it is sometimes not clear what is dependent on what. This confusion is a triumph of common sense over misleading terminology, because often each variable is dependent on some third variable, which may or may not be mentioned. It is reasonable, for instance, to think of the height of children as dependent on age rather than the converse but consider a positive correlation between mean tar yield and nicotine yield of certain brands of cigarette.' The nicotine liberated is unlikely to have its origin in the tar: both vary in parallel with some other factor or factors in the composition of the cigarettes. The yield of the one does not seem to be "dependent" on the other in the sense that, on average, the height of a child depends on his age. In such cases it often does not matter which scale is put on which axis of the scatter diagram. However, if the intention is to make inferences about one variable from the other, the observations from which the inferences are to be made are usually put on the baseline. As a further example, a plot of monthly deaths from heart disease against monthly sales of ice cream would show a negative association. However, it is hardly likely that eating ice cream protects from heart disease! It is simply that the mortality rate from heart disease is inversely related – and ice cream consumption positively related – to a third factor, namely environmental temperature.

Calculation of the correlation coefficient

The data are given in table 11.1 and the scatter diagram shown in figure 11.2 Each dot represents one variable, and it is placed at the point corresponding to the measurement of the studding hours (horizontal axis) and the exam score (vertical axis). The registrar now inspects the pattern to see whether it seems likely that the area covered by the dots centres on a straight line or whether a curved line is needed. In this case the paediatrician decides that a straight line can adequately describe the general trend of the dots. His next step will therefore be to calculate the correlation coefficient.

	Hours spent studying	Exam score	IQ score
Hours spent studying	1.00	0.82	0.48
Exam score	0.82	1.00	0.33
IQ score	0.08	0.33	1.00
Hours spent sleeping	-0.22	-0.04	0.06
School rating	0.36	0.23	0.02

When making the scatter diagram (figure 11.2) to show the studding hours and the exam score in the sample, the studding hours figures as in columns (1), (2), and (3) of table 11.1 . It is helpful to

arrange the observations in serial order of the independent variable when one of the two variables is clearly identifiable as independent. The corresponding figures for the dependent variable can then be examined in relation to the increasing series for the independent variable. In this way we get the same picture, but in numerical form, as appears in the scatter diagram.

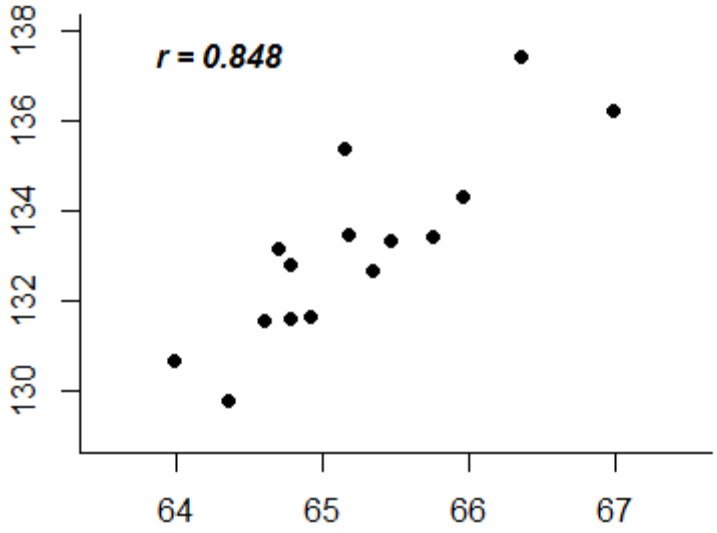

Figure 11.2 Scatter diagram of relation in samples between studding hours and exam score.

The calculation of the correlation coefficient is as follows, with x representing the values of the independent variable (in this case studding hours) and y representing the values of the dependent variable (in this case exam score). The formula to be used is:

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{[\Sigma(x - \bar{x})^2 (y - \bar{y})^2]}}$$

which can be shown to be equal to:

$$r = \frac{\Sigma xy - \bar{n}\bar{x}\bar{y}}{(n-1)SD\,(x)SD(y)}$$

Calculator procedure

Find the mean and standard deviation of x, as described in

$$\bar{x}, SD(x)$$

$$\bar{x} = 144.6, SD(x) = 19.3769$$

Find the mean and standard deviation of y:

$$\bar{y}, SD(y)$$
$$\bar{y} = 66.93, SD(y) = 23.6476$$

Subtract 1 from n and multiply by SD(x) and SD(y), (n – 1)SD(x)SD(y)
$14 \times 19.3679 \times 23.6976 (6412.0609)$

This gives us the denominator of the formula. (Remember to exit from "Stat" mode.)
For the numerator multiply each value of x by the corresponding value of y, add these values together and store them.
110 x 44 = *Min*
116 x 31 = *M+*
etc.

This stores $\Sigma xy\ (150605)$ in memory. Subtract $n\bar{x}\bar{y}$

MR – 15 x 144.6 x 66.93 (5426.6)

Finally divide the numerator by the denominator.
r = 5426.6/6412.0609 = 0.846.

The correlation coefficient of 0.846 indicates a strong positive correlation between size of studding hours and exam score. But in interpreting correlation it is important to remember that correlation is not causation. There may or may not be a causative connection between the two correlated variables. Moreover, if there is a connection it may be indirect.

A part of the variation in one of the variables (as measured by its variance) can be thought of as being due to its relationship with the other variable and another part as due to undetermined (often "random") causes. The part due to the dependence of one variable on the other is measured by Rho. For these data Rho= 0.716 so we can say that 72% of the variation between sample in size of the studding hours is accounted for by the exam score of the sample. If we wish to label the strength of the association, for absolute values of r, 0-0.19 is regarded as very weak, 0.2-0.39 as weak, 0.40-0.59 as moderate, 0.6-0.79 as strong and 0.8-1 as very strong correlation, but these are rather arbitrary limits, and the context of the results should be considered.

Significance test
To test whether the association is merely apparent, and might have arisen by chance use the *t* test in the following calculation:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

The *t* Appendix Table B
is entered at n – 2 degrees of freedom.
For example, the correlation coefficient for these data was 0.846.
The number of pairs of observations was 15.

Entering table B at 15 – 2 = 13 degrees of freedom we find that at t = 5.72, P < 0.001 so the correlation coefficient may be regarded as highly significant. Thus (as could be seen immediately from the scatter plot) we have a very strong correlation between dead space and height which is most unlikely to have arisen by chance.

The assumptions governing this test are:

1. That both variables are plausibly Normally distributed.
2. That there is a linear relationship between them.
3. The null hypothesis is that there is no association between them.

The test should not be used for comparing two methods of measuring the same quantity, such as two methods of measuring peak expiratory flow rate. Its use in this way appears to be a common mistake, with a significant result being interpreted as meaning that one method is equivalent to the other. The reasons have been extensively discussed(2) but it is worth recalling that a significant result tells us little about the strength of a relationship. From the formula it should be clear that with even with a very weak relationship (say r = 0.1) we would get a significant result with a large enough sample (say n over 1000).

Spearman rank correlation

A plot of the data may reveal outlying points well away from the main body of the data, which could unduly influence the calculation of the correlation coefficient. Alternatively the variables may be quantitative discrete such as a mole count, or ordered categorical such as a pain score. A non-parametric procedure, due to Spearman, is to replace the observations by their ranks in the calculation of the correlation coefficient.

This results in a simple formula for Spearman's rank correlation, Rho.

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

where d is the difference in the ranks of the two variables for a given individual. Thus we can derive table 11.2 from the data in table 11.1 .

In this case the value is very close to that of the Pearson correlation coefficient. For n> 10, the Spearman rank correlation coefficient can be tested for significance using the t test given earlier.

THE REGRESSION EQUATION

Correlation describes the strength of an association between two variables, and is completely symmetrical, the correlation between A and B is the same as the correlation between B and A. However, if the two variables are related it means that when one changes by a certain amount the other changes on an average by a certain amount. For instance, in the sample described earlier greater exam score is associated, on average, with greater studding hours. If y represents the dependent variable and x the independent variable, this relationship is described as the regression of y on x.

The relationship can be represented by a simple equation called the regression equation. In this context "regression" (the term is a historical anomaly) simply means that the average value of y is a "function" of x, that is, it changes with x.

The regression equation representing how much y changes with any given change of x can be used to construct a regression line on a scatter diagram, and in the simplest case this is assumed to be a straight line. The direction in which the line slopes depends on whether the correlation is positive or negative. When the two sets of observations increase or decrease together (positive) the line slopes upwards from left to right; when one set decreases as the other increases the line slopes downwards from left to right. As the line must be straight, it will probably pass through few, if any, of the dots. Given that the association is well described by a straight line we have to define two features of the line if we are to place it correctly on the diagram. The first of these is its distance above the baseline; the second is its slope. They are expressed in the following *regression equation* :

$$Y = a + bX$$

$$b = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} \qquad a = \frac{\sum Y - b\sum X}{N}$$

Where,
N = number of observations, or years
X = a year index (decade)
Y = population size for given census years

With this equation we can find a series of values of y_{fit} the variable, that correspond to each of a series of values of x, the independent variable. The parameters α and β have to be estimated from the data. The parameter signifies the distance above the baseline at which the regression line cuts the vertical (y) axis; that is, when y = 0. The parameter β (the *regression coefficient*) signifies the amount by which change in x must be multiplied to give the corresponding average change in y, or the amount y changes for a unit increase in x. In this way it represents the degree to which the line slopes upwards or downwards.

The regression equation is often more useful than the correlation coefficient. It enables us to predict y from x and gives us a better summary of the relationship between the two variables. If, for a particular value of x, x i, the regression equation predicts a value of y fit , the prediction error is $y_1 - y_{fit}$. It can easily be shown that any straight line passing through the mean values x and y will give a total prediction error $\Sigma(y_1 - y_{fit})$ of zero because the positive and negative terms exactly cancel. To remove the negative signs we square the differences and the regression equation chosen to minimise the sum of squares of the prediction errors, $S^2 = \Sigma(y_1 - y_{fit})^2$ We denote the sample estimates of Alpha and Beta by a and b. It can be shown that the one straight line that minimises s_2 , the least squares estimate, is given by

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

and

$$a = \bar{y} - b\bar{x}$$

it can be shown that

$$b = \frac{\Sigma xy - n\bar{x}\bar{y}}{(n-1)SD(x)^2}$$

which is of use because we have calculated all the components of equation (11.2) in the calculation of the correlation coefficient.

The calculation of the correlation coefficient on the data in table 11.2 gave the following:

$\Sigma xy = 150605, SD(x) = 19.3679, \bar{y} = 66.93, \bar{x} = 144.6$

Applying these figures to the formulae for the regression coefficients, we have:

$$b = \frac{150605 - 15 \times 66.93 \times 144.6}{14 \times 19.3679^2} = \frac{5426.6}{5251.6} = 1.033 ml/cm$$

$$a = 66.39 - (1.033 \times 144.6) = -82.4$$

Therefore, in this case, the equation for the regression of y on x becomes

$$y = -82.4 + 1.033x$$

This means that, on average, for every increase in height of 1 cm the increase in anatomical dead space is 1.033 ml *over the range of measurements made*.

The line representing the equation is shown superimposed on the scatter diagram of the data in figure 11.2. The way to draw the line is to take three values of x, one on the left side of the scatter diagram, one in the middle and one on the right, and substitute these in the equation, as follows:
If x = 110, y = (1.033 x 110) – 82.4 = 31.2
If x = 140, y = (1.033 x 140) – 82.4 = 62.2
If x = 170, y = (1.033 x 170) – 82.4 = 93.2

Although two points are enough to define the line, three are better as a check. Having put them on a scatter diagram, we simply draw the line through them.

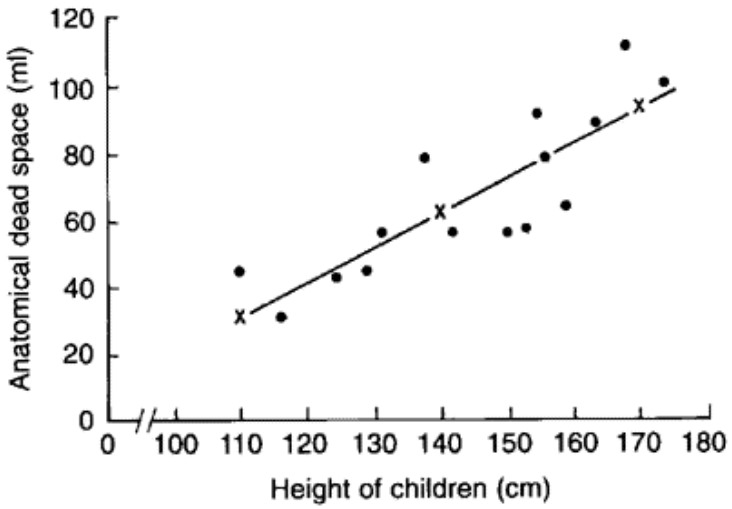

Figure 11.3 Regression line drawn on scatter diagram relating height and pulmonaiy anatomical dead space in 15 children

The standard error of the slope SE(b) is given by:

$$SE_{(b)} = \frac{S_{res}}{\sqrt{\Sigma(x - \bar{x})^2}}$$

where S_{res} is the residual standard deviation, given by:

$$S_{res} = \sqrt{\frac{\Sigma(y - y_{fit})^2}{n - 2}}$$

This can be shown to be algebraically equal to

$$\sqrt{((SD((y)^2(1 - r^2)(n - 1) / (n - 2)}$$

We already have to hand all of the terms in this expression. Thus S_{res} is the square root of $23.6476^2(1 + -0.846^2)14 / 13 = \sqrt{171.2029} = 13.08445$. The denominator of (11.3) is 72.4680. Thus SE(b) = 13.08445/72.4680 = 0.18055. We can test whether the slope is significantly different from zero by: t = b/SE(b) = 1.033/0.18055 = 5.72.

Again, this has n – 2 = 15 – 2 = 13 degrees of freedom. The assumptions governing this test are:

1. That the prediction errors are approximately Normally distributed. Note this does not mean that the x or y variables have to be Normally distributed.
2. That the relationship between the two variables is linear.
3. That the scatter of points about the line is approximately constant – we would not wish the variability of the dependent variable to be growing as the independent variable increases. If this is the case try taking logarithms of both the x and y variables.

Note that the test of significance for the slope gives exactly the same value of P as the test of significance for the correlation coefficient. Although the two tests are derived differently, they are algebraically equivalent, which makes intuitive sense.

We can obtain a 95% confidence interval for b from $b - t_{0.05} \times SE(b)$ to $b + t_{0.05} \times SE(b)$ where the t statistic from has 13 degrees of freedom, and is equal to 2.160. Thus the 95% confidence interval is l.033 – 2.160 x 0.18055 to l.033 + 2.160 x 0.18055 = 0.643 to 1.422.

Regression lines give us useful information about the data they are collected from. They show how one variable changes on average with another, and they can be used to find out what one variable is likely to be when we know the other – provided that we ask this question within the limits of the scatter diagram. To project the line at either end – to extrapolate – is always risky because the relationship between x and y may change or some kind of cut off point may exist. For instance, a regression line might be drawn relating the chronological age of some children to their bone age, and it might be a straight line between, say, the ages of 5 and 10 years, but to project it up to the age of 30 would clearly lead to error. Computer packages will often produce the intercept from a regression equation, with no warning that it may be totally meaningless. Consider a regression of blood pressure against age in middle aged men. The regression coefficient is often positive, indicating that blood pressure increases with age. The intercept is often close to zero, but it would be wrong to conclude that this is a reliable estimate of the blood pressure in newly born male infants!

MORE ADVANCED METHODS

More than one independent variable is possible – in such a case the method is known as multiple regression. (3, 4) This is the most versatile of statistical methods and can be used in many situations. Examples include: to allow for more than one predictor, age as well as height in the above example; to allow for covariates – in a clinical trial the dependent variable may be outcome after treatment, the first independent variable can be binary, 0 for placebo and 1 for active treatment and the second independent variable may be a baseline variable, measured before treatment, but likely to affect outcome.

CONCLUSION

It is a common error to confuse correlation and causation. All that correlation shows is that the two variables are associated. There may be a third variable, a confounding variable that is related to both of them. For example, monthly deaths by drowning and monthly sales of ice-cream are positively correlated, but no-one would say the relationship was causal! Firstly always look at the scatter plot and ask, is it linear? Having obtained the regression equation, calculate the residuals $e_1 = y_1 - y_{fit}$ A histogram of e_1 will reveal departures from Normality and a plot of versus y_{fit} will reveal whether the residuals increase in size as y_{fit} increases.

REFERENCES

1. Russell MAH, Cole PY, Idle MS, Adams L. Carbon monoxide yields of cigarettes and their relation to nicotine yield and type of filter. BMJ 1975; 3:713.
2. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; i:307-10.

3. Brown RA, Swanson-Beck J. Medical Statistics on Personal Computers , 2nd edn. London: BMJ Publishing Group, 1993.
4. Armitage P, Berry G. In: Statistical Methods in Medical Research , 3rd edn. Oxford: Blackwell Scientific Publications, 1994:312-41.
5. Croxton, Frederick Emory; Cowden, Dudley Johnstone; Klein, Sidney (1968) Applied General Statistics, Pitman. ISBN 9780273403159 (page 625)
6. Dietrich, Cornelius Frank (1991) Uncertainty, Calibration and Probability: The Statistics of Scientific and Industrial Measurement 2nd Edition, A. Higler. ISBN 9780750300605 (Page 331)
7. Aitken, Alexander Craig (1957) Statistical Mathematics 8th Edition. Oliver & Boyd. ISBN 9780050013007 (Page 95)
8. Rodgers, J. L.; Nicewander, W. A. (1988). "Thirteen ways to look at the correlation coefficient". The American Statistician. 42 (1): 59–66. doi:10.1080/00031305.1988.10475524. JSTOR 2685263.
9. Dowdy, S. and Wearden, S. (1983). "Statistics for Research", Wiley. ISBN 0-471-08602-9 pp 230
10. Francis, DP; Coats AJ; Gibson D (1999). "How high can a correlation coefficient be?". Int J Cardiol. 69 (2): 185–199. doi:10.1016/S0167-5273(99)00028-5.
11. Yule, G.U and Kendall, M.G. (1950), "An Introduction to the Theory of Statistics", 14th Edition (5th Impression 1968). Charles Griffin & Co. pp 258–270
12. Kendall, M. G. (1955) "Rank Correlation Methods", Charles Griffin & Co.
13. Mahdavi Damghani B. (2013). "The Non-Misleading Value of Inferred Correlation: An Introduction to the Cointelation Model". Wilmott Magazine. 2013 (67): 50–61. doi:10.1002/wilm.10252.
14. Székely, G. J. Rizzo; Bakirov, N. K. (2007). "Measuring and testing independence by correlation of distances". Annals of Statistics. 35 (6): 2769–2794. arXiv:0803.4101. doi:10.1214/009053607000000505.
15. Székely, G. J.; Rizzo, M. L. (2009). "Brownian distance covariance". Annals of Applied Statistics. 3 (4): 1233–1303. arXiv:1010.0297. doi:10.1214/09-AOAS312. PMC 2889501. PMID 20574547.
16. Lopez-Paz D. and Hennig P. and Schölkopf B. (2013). "The Randomized Dependence Coefficient", "Conference on Neural Information Processing Systems" Reprint
17. Thorndike, Robert Ladd (1947). Research problems and techniques (Report No. 3). Washington DC: US Govt. print. off.
18. Nikolić, D; Muresan, RC; Feng, W; Singer, W (2012). "Scaled correlation analysis: a better way to compute a cross-correlogram". European Journal of Neuroscience. 35 (5): 1–21. doi:10.1111/j.1460-9568.2011.07987.x. PMID 22324876.
19. Higham, Nicholas J. (2002). "Computing the nearest correlation matrix—a problem from finance". IMA Journal of Numerical Analysis. 22 (3): 329–343. CiteSeerX 10.1.1.661.2180. doi:10.1093/imanum/22.3.329.
20. "Portfolio Optimizer". portfoliooptimizer.io/.
21. Borsdorf, Rudiger; Higham, Nicholas J.; Raydan, Marcos (2010). "Computing a Nearest Correlation Matrix with Factor Structure". SIAM J. Matrix Anal. Appl. 31 (5): 2603–2622. doi:10.1137/090776718.

22. Qi, HOUDUO; Sun, DEFENG (2006). "A quadratically convergent Newton method for computing the nearest correlation matrix". SIAM J. Matrix Anal. Appl. 28 (2): 360–385. doi:10.1137/050624509.
23. Park, Kun Il (2018). Fundamentals of Probability and Stochastic Processes with Applications to Communications. Springer. ISBN 978-3-319-68074-3.
24. Aldrich, John (1995). "Correlations Genuine and Spurious in Pearson and Yule". Statistical Science. 10 (4): 364–376. doi:10.1214/ss/1177009870. JSTOR 2246135.
25. Mahdavi Damghani, Babak (2012). "The Misleading Value of Measured Correlation". Wilmott Magazine. 2012 (1): 64–73. doi:10.1002/wilm.10167.
26. Anscombe, Francis J. (1973). "Graphs in statistical analysis". The American Statistician. 27 (1): 17–21. doi:10.2307/2682899. JSTOR 2682899

Acknowledgements
The author profoundly appreciates all the people who have successfully contributed to ensuring this paper in place. Their contributions are acknowledged however their names cannot be mentioned.

Conflict of Interest
The author declared no conflict of interest.

How to cite this article: Shah. A. (2020). Correlation Vs. Regression: A Review. *International Journal of Social Impact, 5*(2), 151-162. DIP: 18.02.015/20200502, DOI: 10.25215/2455/0502015