Review of: "Decoding the Correlation Coefficient: A Window into Association, Fit, and Prediction in Linear Bivariate Relationships"

Vishal Ramnath[1]

1 University of South Africa

Potential competing interests: No potential competing interests to declare.

Comments to assist with an author revision

1. In the section *ABSTRACT* the author should avoid ambiguity to aid the reader to understand the purpose of the study. It is suggested that the first sentence read as ``This article examines the relationship between the correlation coefficient (r) and the regression slope (beta_1) in a linear bivariate relationship y=beta_1 x + beta_0''. The author should avoid ambiguous notation as by convention regression analysis uses the symbol beta for the regression coefficients where the subscripts correspond to powers of the observed variable. The use of symbols a and b is potentially confusing to a reader as the regression can be y = ax + b, or y=a + bx. It is suggested that the second sentence remind the reader that the correlation coefficient and coefficient of determination are different concepts and that the second sentence clearly assist readers by incorporating the definition to read as "The article emphasizes that the coefficient of determination defined as r^2 should not be interpreted solely as a measure of fit,…". The second part of this sentence must be reworked as the meaning of "…the association and prediction of change" is unclear. The specific meaning of this term does not appear to be clear to the reviewer and the author must amend this to aid the reader.
2. In the section *INTRODUCTION* paragraph 1 sentence 2 it is redundant to write ``…arguing that the association between the coefficient and slope is influenced by the standard deviations of the variables" since the regression coefficient beta_1 from standard statistics is already well known by definition to take the form \hat{beta}_1 = s_{xy} (s_y/s_x) according to the Wikipedia article [Ref1]. If the author clearly specifies this formula aspects of the later analysis in the paper can be more readily understood.
3. The second last and last sentences in the section *INTRODUCTION* allude to the difference in correlation and causation and it is suggested that the author more clearly alert the reader that these concepts are not synonymous and are technically distinct. Briefly, a correlation is a statistical concept for a statistically complete self-consistent closed mathematical model, whilst a causation is a cause-and-effect system (not a model representation/approximation). It is technically possible for a system to exhibit a non-zero correlation but not causation due to statistical incompleteness through for example the phenomena of so-called "hidden variables" when the chosen statistical model is too simplistic. In more complex models when Principal Component Analysis (PCA) is used to make an analysis more tractable by reducing the model dimensions it becomes more problematic to infer causation from

correlation, and statistical correlations can unintentionally mask artificial behavior and phenomena due to mathematical model incompleteness. The author should allow the reader to be adequately alerted to the scientific literature on the technical differences in correlation and causation and why these concepts can be confusing and misleading if not adequately understood. The present list of references by the author is incomplete and does not offer a meaningful discussion/insight of the underlying technical challenges to infer causation from correlation. It is suggested that the author provide specific counter-examples of when statistical correlation does not logically result in causation in a real system from an appropriate discipline specific context so that the reader can more fully appreciate the underlying technical challenges.

4. The technical reasoning in section *ARGUMENT* paragraph 2 is mathematically erroneous due to ambiguous mathematical notation and symbols. The author states that a 1 SD change in X corresponds to a change of r*SD in Y and that if X is changed by one SDx then Y will change by r*SDy however this is not mathematically possible according to the linear regression model. To prove why this claim is incorrect let the linear regression model be y = beta_1 x + beta_0 where as per standard statistical terminology beta_1 = r (sigma_y/sigma_x) is the linear regression coefficient, beta_0 is the regression intercept, r is the dimensionless regression coefficient where -1 <= r <= +1, sigma_x is the standard deviation in x and sigma_y is the corresponding standard deviation in y. Let the initial value of the independent variable be x = x_1 from which the corresponding observed value will be y = beta_1 x_1 + beta_0. Now calculate the observed value by setting x = x_1 + sigma_x so that the regression model will then predict y_2 = beta_1 (x_1 + sigma_x) + beta_0. It then follows that the difference is (y_2 – y_1)=beta_1 sigma_x. The author states in section *ARGUMENT* paragraph 2 sentence 3 that the change in Y is r*SD and then later in section *ARGUMENT* paragraph 2 sentence 4 that the change is that Y changes by r times SDy. Both sentence 3 and sentence 4 are mathematically incorrect. In sentence 3 the change in Y is not r*SD but beta_1 sigma_x (the change is not the product of the correlation coefficient and the perturbation but the change is the product of the regression parameter and the perturbation) whilst in sentence 4 the change in Y is r sigma_y (the change is not the product of the original perturbation and the regression coefficient, but the product of the regression coefficient and the standard deviation in the observed value). The author must make the necessary mathematical corrections with a consistent mathematical choice of notation/symbols for the regression model for the technical arguments to make sense.
5. In section *ARGUMENT* paragraph 5 on the prediction of a 100% change in X this statement is problematic as whilst the correlation is dimensionless the terms X and Y are dimensional quantities (with physical units). It is not necessarily physically possible in all linear regressions for there to be a 100% change in X and if X is changed by 100% it is not necessarily the case that any associated variation prediction in Y would have any statistical meaning. As an example, X may represent the height of a person, say x_1 = 1.8 meters for the sake of argument, then a 100% change of X is 100*((x_2 – x_1)/x_1) = 100 which implies that x_2 = 2 x_1 = 2*1.8 = 3.6 meters which is not feasible. The linear regression formula and the influence of the standard uncertainties on the value of the regression parameters is known in measurement science technical literature and the influence of uncertainties on confidence intervals for parameters (and confidence regions for probabilities) has been previously documented in the technical literature using multivariate calculus and stochastic Monte Carlo simulation approaches (see [Ref2], [Ref3], [Ref4]). Typically, the percentage variations in X are ± 5% to not more than ± 10% in any linear regression and it is statistically implausible to utilize

variations in X of ± 100% and incorporate that in a percentage variation in Y. The author should advise the reader and caution that if there is such a large variation in X it is more likely that a simple linear regression is not adequate for the particular analysis and either a nonlinear regression is more appropriate or alternatively that noise reduction techniques must first be utilized to pre-process the raw data before attempting any subsequent regression analysis (see [Ref5]). The concept of covariances/correlation is technically restricted to linear/Gaussian models.

6. The technical reasoning in section *ARGUMENT* paragraph 5 that a higher value of b i.e. beta_1 corresponds to a higher correlation coefficient is not strictly mathematically true. Formally beta_1 = r(sigma_y/sigma_x) so whilst a higher value of beta_1 would in general lead to a higher value of r (and vice versa) this is not necessarily always true in all cases. Referring to the formula it may be observed that a smaller value of sigma_x would in general automatically lead to a higher value of beta_1 (division by an arbitrarily small non-zero number yields an arbitrarily large quantity). This means the general assertion is not necessarily true in all cases as all four values (sigma_x, sigma_y & beta_0, beta_1) which logically determine r) are interconnected to each other in a linear regression. It is not possible to make any general observations of the influences of sigma_x and sigma_y on the magnitude of the slope beta_1 as the behavior depends on a case-by-base basis to particular physical model to which the linear regression is considered (the model may be physical, economic, financial or social). It is suggested that the author caution the reader that the interpretation of the observed term X and the predicted term Y are themselves ambiguous based on the underlying context which in term affects the quality of a regression. For example, in a physical laboratory the measurement of physical quantities is straightforward, whilst in a social studies investigation the meaning of variables may be ambiguous. As an example, it may be inappropriate to attempt a simple linear regression between age and conservatism or to make any general inferences as the definition of conservatism is subjective (it varies by country, profession, culture etc.) and can lead to inferences that have no statistical meaning when qualitative inputs are erroneously ascribed to quantitative variables/parameters. In certain cases, it may be more appropriate to utilize a multivariate regression model or a nonlinear regression instead.
7. The author touches on but does not elaborate of the discussion by Rodgers & Nicewander [Ref6] on the 13 different interpretations of the correlation coefficient. It is suggested that the author make a graphical summary of these 13 different interpretations (using appropriate case study data from a discipline/field in which the author works) to assist the reader to understand how the relationship between the correlation coefficient influences the slope based on the magnitudes on the standard uncertainties of the observed input X and the predicted output Y respectively. It is suggested that the author incorporate and utilize appropriate technical/statistical supporting evidence (tables, graphs, plots etc.) in the exposition to aid with reader comprehension of the work in the paper.
8. It is suggested that the author incorporate more mathematical/statistical rigor in the analysis of how the existing technical literature has gaps and what these gaps are to assist the reader. The topic of linear bivariate regression can be tackled in various ways at varying levels of rigor from different disciplines. It is suggested that the author revise the manuscript for a specific target reader audience in mind and offer more tangible numerical/graphical evidence for that that particular targeted audience background (rewrite for a reader from a social science field in more qualitative terms with evidence from the social sciences field versus rewrite for a reader from the physical sciences field in more quantitative terms with evidence from the physical sciences field) so the reader can more fully appreciate the

exposition.

9. The author is thanked for their research work and the above comments are offered in the spirit of positive constructive inputs/suggestions to make the article better.

References

[Ref1] Wikipedia, "*Simple linear regression*", Simple linear regression - Wikipedia (Accessed: 17 July 2023)

[Ref2] BIPM et al, "*Evaluation of measurement data – Guide to the expression of uncertainty in measurement*", Technical Report JCGM/WG1 GUM, 2008, https://www.bipm.org/en/publications/guides/ (Accessed: 17 July 2023)

[Ref3] M. Krystek & M. Anton, "*A least-squares algorithm for fitting data points with mutually correlated coordinates to a straight line*", Meas. Sci. Technol., Vol. 22, pp. 035101 (2011). DOI: 10.1088/0957-0233/22/3/035101

[Ref4] V. Ramnath, "*Comparison of straight lien curve fit approaches for determining parameter variances and covariances*", Int. J. Metrol. & Qual. Eng., Vol. 11, no. 14 (2020). DOI: 10.1051/ijmqe/2020011

[Ref5] N. Moriya, "*Noise related multivariate optimal joint analysis in longitudinal stochastic processes*", In: Fengshan Yang (ed.) Progress in Applied Mathematical Modeling, pp. 223-260. Nova Science Publishers, Inc. ISBN 978-1-60021-976-4

[Ref6] L. Rodgers, W. A. Nicewander, "*Thirteen ways to look at the correlation coefficient", The American Statistician*, Vol. 42, No. 1, pp. 59-66 (1988). DOI: 10.1080/00031305.1988.10475524