Open Peer Review on Qeios

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

S A Hamed Hosseini[1]

1 University of Newcastle

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

This article examines the relationship between the correlation coefficient (r) and the regression slope in a linear bivariate relationship. The article emphasizes that the coefficient of determination should not be interpreted solely as a measure of fit, as it also signifies the association and prediction of change. The estimation of the percentage of variation in Y for a given change in X is explored, as well as the role of the correlation coefficient in prediction. The limitations of using r alone for prediction are discussed, and the importance of considering standard deviations is emphasized. The article concludes by suggesting further acknowledging the misuse and broader potential of correlation in understanding causality and predictability.

Introduction

This article delves into the relationship between the correlation coefficient (r) and the regression slope (b) in linear bivariate relationships. It challenges the assumption that a higher correlation coefficient indicates a stronger correspondence in change, arguing that the association between the coefficient and slope is influenced by the standard deviations of the variables. The coefficient of determination is examined as a measure of fit, association, and prediction.

The article highlights the significance of considering the estimation of variation in Y and explores the use of the correlation coefficient for prediction. Limitations in using r alone for accurate prediction are discussed, emphasizing the importance of incorporating standard deviations into the analysis. In conclusion, this article highlights that the correlation coefficient is more than a measure of the tightness of fit in linear bivariate relationships. It demonstrates the interdependence between the coefficient and the slope, influenced by the standard deviations of the variables. The article emphasizes the importance of considering standard deviations in accurately estimating variation and predicting outcomes. While correlation alone cannot confirm causation, it can provide valuable insights when examined alongside more advanced modelling methods. The article encourages further exploration and discussion on the multifaceted nature of the correlation coefficient in understanding causality and predictability.

Methodology

The argument put forth in this article draws upon the understanding of linear bivariate relationships and the principles of regression analysis. It examines the relationship between the correlation coefficient and the regression slope, considering the influence of standard deviations on this relationship. The estimation of variation in Y and the prediction using the correlation coefficient are discussed in relation to the underlying assumptions and limitations.

Argument

In a linear bivariate relationship, the coefficient and the regression slope are not completely independent, and thus we cannot assert with certainty that a higher correlation coefficient (r) never implies a stronger correspondence in change. The relationship between the coefficient and the slope in a linear regression depends on the standard deviations (SDs) of the variables. In a simple linear regression with two variables (Y and X), the coefficient (b) is a function of the correlation coefficient (r) multiplied by the ratio of the SDs. This means that a 1 SD change in X corresponds to a change of (r * SD) in Y. By using the equation of the regression line, if you change X by one SDx, you will observe that Y changes by r times SDy.

Instead of stating that a higher coefficient of determination should never be interpreted as indicating stronger conservatism in older people compared to younger people (in comparison to a lower coefficient), it would be more accurate to say that “it may not necessarily always imply that... .”

Upon reflection, it seems contradictory to claim that r squared demonstrates how much variation in Y can be explained by X and then dismiss it as merely a measure of the fit to the line, devoid of any association or predictive power in their changes.

The accurate measurement of the “percentage of variation in Y” for a 100 percent change in X can be achieved by using the coefficient (b) in the equation (b * (X1/(a+bX1)) * 100%), where X1 represents any point on the regression line rather than an observed case.

Even without conducting regression analysis, the correlation coefficient (r) can still provide an estimation of prediction, although not always accurately and reliably.

It is important to remember the relationship between the coefficient (b) and the correlation coefficient (r), where b can be replaced by r times the ratio of the standard deviations (SDy/SDx), whether standardized or not. Therefore, a higher

coefficient (b) corresponds to a higher correlation coefficient (r).

However, it is crucial to note that using r alone does not provide a certain and accurate prediction of Y, unless the units of change are represented by the standard deviations (SDs) and their ratio is taken into account. This consideration depends on the specific SDs involved. When the SD of X is much larger than the SD of Y, it indicates that the cases are closer to a nearly horizontal line, resulting in a smaller slope and a higher correlation coefficient. This situation prevents us from concluding that higher age is strongly associated with higher conservatism, but this is applicable only under such circumstances.

In the scenario where the standard deviation of X (SDx) is much smaller than the standard deviation of Y (SDy), even if the correlation is relatively small, the slope becomes steeper. This indicates that the cases are more dispersed vertically than horizontally, deviating from the regression line. Despite the smaller correlation, we can observe that within a certain age range, older people are significantly more conservative based on the higher slope.

As a thought experiment, let's consider a scenario where we are only provided with the correlation coefficient (r) to teach students about predictability, without introducing the concept of standard deviations. In this case, the best approximation for predictability would be to examine the "ratio of the standard deviations" multiplied by X1/(a+bX1) in the aforementioned equation. This could potentially yield a rough estimation of r. If this rough estimation shows some positive indication, multiplying it by r itself would lead us to a prediction that is not too far from reality: $[r. (SDy/SDx) * (X1/(a+bX1)] \approx r . r \approx r2$.

Discussion

Correlation is often misused but possesses interesting qualities beyond a simple measure of association or tightness. Moreover, we need to bear in mind that when there is a high degree of tightness around a relatively horizontal regression line, in many cases, both the correlation coefficient (r) and the slope tend to be closer to zero.

Joseph Lee Rodgers and W. Alan Nicewander (1988) in their article "Thirteen Ways to Look at the Correlation Coefficient" explore various functions and features of the correlation coefficient, including its relationship with the regression slope or the correlation coefficient as a factor of slope. While correlation alone cannot confirm causation, it can be indicative of it under certain conditions. There are more sophisticated modelling methods that utilize correlation to approach a better understanding of causation. Therefore, it is important to tolerate non-definitive reflections on causality and predictability, particularly when reflections on correlation are supported by more advanced literature and research (Hosseini, 2021).

References

Hosseini, S. A. H. (2021) The correlation coefficient is more than a measure of the tightness of fit!, ResearchGate.

Lee Rodgers, J. & Nicewander, W. A. (1988) Thirteen Ways to Look at the Correlation Coefficient, The American Statistician, 42(1), 59-66. https://doi.org/10.1080/00031305.1988.10475524.