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

Gergely Tóth[1]

1 Eötvös Lorand University

Potential competing interests: No potential competing interests to declare.

I found this manuscript rather confusing. Many arguments are according to the omission of other existing statistical quantities or their misinterpretation.

On the incorporation of SD-s in regression assessment – versus covariance:

If we have X and Y vectors, we usually define the covariance of the two set of values, at first. It is an unbounded quantity and the magnitude is closely connected to the standard deviation of both vectors. If we standardize the two vectors (or only scale the vectors with their standard deviations), the covariance of the two vectors become identical to the Pearson correlation coefficient. This correlation coefficient is independent from the 'slope' connecting the two vectors, it is between -1 and 1. If a line could be perfectly fit on the data, we might get -1 or +1 depending only on the sign of the slope. I think, many arguments of the author are mentioned in order to use covariance (an existing feature) as well. But it is a basic step in statistics, therefore, there is no novelty in it.

What is R2:

In the case of regression analysis the coefficient of performance, R2, is a general measure for all type of modelling methods. Its definition is 1-RSS/TSS, where the RSS is the residual sum of square (sum of squared differences of the model given and the original data) and TSS is the total sum of square of the original data (sum of squared difference to the mean Y). I think, using R2 according to this definition is one of the best qualitative measure for a regression, especially if we uses parallel it on fitted data (goodness of fit), on cross-validated data (Q2 or R2cv, robustness) and on separated test data (R2test or Q2F2, prediction). In general, R2 is not related to the Pearson correlation coefficient, r. R2=1 means that the model provides exact answers on the data used in the training set. R2=0 means, that a model provides as good answers as we use only the average of the Y-s in each case. (There are several modifications of R2, I do not want to discuss and evaluate them here, e.g., adjusted R2s, Q2F1-3, Roy-Ojha coefficients, Lin's concordance correlation coefficient. Furthermore the relation of R2 to RMSE is also an interesting question.)

Special case when R2 and r have relation to each other:

They are related to each other only in one case, when a linear regression is performed to determine both slope and intercept parallel and the merit function is to minimalize RSS. In this special case, the square of the Person correlation coefficient is equal to R2. In the meanwhile, the mean of the original and the modelled values are also the same (using the regression method mentioned), therefore, (only here) R2=r^2=1-RSS/TSS=MSS/TSS. Here MSS is the sum of squares explained by the model and R2 is the ratio of the sum of squares explained by the regression. It means, in the case of this special regression (but often this regression is used), r^2 has all advantages of R2 and, I think, in this case r is also an effective measure to quantify the fit. In this case, the absolute value of r cannot be large on bad fits. In this type of regression 0<=R2<=1.

Use of r or r^2 for assessment of other type of model results:

It is a rather dangerous field, because here the absolute value of r might be very large without having a good fit. It might be possible, that the model did not provide any meaningful slope and r is good (or even +1 or - 1) while R2 is small or negative. Furthermore, R2<=1, but the lower limit is minus infinity. r^2 is not equal to R2 and in the meanwhile 0<=r^2<=1 by definition.

I think, the proposed manuscript is somehow confused in these basics of covariance, Pearson correlation coefficient, R2 and in their limits and their roles in different modelling schemes. I should say, that this confusion is somehow present in many ‘inventions’ in regression analysis, e.g., Lin justified his concordance correlation coefficient using exactly the same mismatch of r, r^2 and R2 as the author of this manuscript (see e.g. wikipedia or the original articles of Lin on it)

Furthermore, I think the author does not follow the rules of scientific communication correctly, the manuscript rather resembles an oral note than a written scientific paper.