

Peer Review

Review of: "Graphical Examples Show Why Caution Is Required When Using the Coefficient of Determination (R^2) to Interpret Data for Medical Case Reports"

Bruno Mario Cesana¹

1. University of Milan, Italy

I would like to begin my review by stating that one of the assumptions of OLS (Ordinary Least Squares) regression is that the values of the dependent variable should be uncorrelated (see Draper & Smith, 3rd Ed. Applied Regression Analysis, page 34). Therefore, in my opinion, exposing the well-known pitfalls of the coefficient of determination by incorrectly using the underlying statistical techniques is an unacceptable procedure. If researchers perform this type of analysis, they should be informed that appropriate methods must be used to handle autocorrelation.

Second, it should be noted that the coefficient of determination is biased, and it is possible to obtain the corrected R^2 , which, however, displays all the pitfalls of the uncorrected R^2 . However, I believe a tutorial should provide this information. Third, the model "1.1. Estimating the best-fit line through the data points" proposed in Appendix 1 is unacceptable. The formulas are well-known and must be used. Finally, since only simple regressions (with a single independent variable) are considered, R^2 is equal to Pearson's correlation coefficient, and all examples can be made in terms of the latter, which shares the same disadvantages.

Furthermore, it is well known from the statistical methodological papers/publications that R^2 depends on the number of dependent variables and on the number of observations (with only a few observations more than the number of independent variables, a value of about 1 can be reached); it depends directly on the range of the independent variable (it increases with the increase of this range and otherwise it decreases, assuming that the correct model has been fitted). A high R^2 does not necessarily assure a statistically significant regression equation, and a high R^2 and statistical significance do not assure a

really useful regression or that the chosen regression model adequately represents the true relationship between the variables.

In addition, it is possible to have a very high R^2 even when the real model is non-linear!.

As specific comments:

1)-Fig D in the graphical abstract: values have to be given. The same comment applies also to the remaining figures.

2)-"the line of the best fit" (what does it mean?) The statistical method has to be explained.

3)-The sentences in the "Conclusion" paragraph are too generic.

Frankly, I doubt that a non-technical paper about the pitfalls of the coefficient of determination is necessary. In any case, it has to be written according to sound statistical methodology.

Declarations

Potential competing interests: No potential competing interests to declare.