Published in final edited form as:

Teach Stat. 2011 September ; 33(3): 81–82. doi:10.1111/j.1467-9639.2010.00455.x.

Use the Correlation Coefficient to Summarize Regression Performance?

Prakash Gorroochurn

Columbia University, New York, USA. pg2113@columbia.edu

Summary

The correlation coefficient is commonly used to indicate the quality of fit in regression. This practice is questionable.

Keywords

Teaching; Goodness of fit

INTRODUCTION

Many regression textbooks portray a close relationship between regression and correlation. For example, one of the primary advocated measures of the goodness of fit in a simple linear regression analysis is R^2 , the square of the correlation coefficient, r , between the independent variable and the dependent variable. Quoting a well-respected textbook (Wonnacott and Wonnacott 1979, p. 152): "...the two [regression and correlation] are so closely related that correlation often becomes a useful aid in interpreting regression. In fact, this is the major reason for studying it." The aim of this article is to argue that using r to summarize the quality of fit of a regression model is not generally appropriate.

CHOOSING INDEPENDENT VARIABLE VALUES

The strategy of selecting a wide range of independent variable values (or 'x-values') in a regression analysis is to be recommended since it increases the precision with which the slope is estimated. In particular, the variance of the usual least squares estimator of the slope in a simple linear regression model $Y = \beta_0 + \beta_1 x + \varepsilon$ is

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2_{r|x}}{S_{xx}}$$

where $\sigma^2_{r|x} \equiv \text{Var}(Y|x)$ is an inherent property of the model and is therefore fixed, while $S_{xx} = \sum_i (x_i - \bar{x})^2$ depends on the sampling scheme of the x -values, increasing as the latter's spread increases. To make this variability small, then, it makes sense to choose a wide range of independent variable values.

THE CORRELATION COEFFICIENT

Now consider the following expression for the sample correlation coefficient r in the case of simple linear regression:

$$|r| = \frac{1}{\sqrt{1 + \frac{(n-2)\sigma_y^2}{\beta_1^2 S_{xx}}}} \cong \frac{1}{\sqrt{1 + \frac{(n-2)\sigma_y^2}{\beta_1^2 S_{xx}}}}$$

The approximation here is “good” for moderate to large sample sizes with the quantities σ_y^2 and β_1 being, of course, properties of the system being studied. Note that $|r| \rightarrow 1$ as $S_{xx}/(n-2)$ increases (van Belle 2008, p. 69). That is, the sample correlation coefficient can be made arbitrarily close to 1 (in magnitude) simply by increasing the spread of the x -values. This means that, if regression sampling is used $R^2 \equiv r^2$ will tend to be an overly optimistic estimate of the goodness of fit of the regression line.

CONCLUSION

The correlation coefficient can be a misleading measure of the quality of fit in a regression. In particular, $|r|$ can be made close to one when a wide range of independent variable values are used.

Acknowledgments

This work was supported in part by NIMH grant MH-48858.

References

- van Belle, G. Statistical Rules of Thumb. 2nd edn. New York: Wiley-Interscience; 2008.
 Wonnacott, RJ.; Wonnacott, TH. Econometrics. New York: John Wiley and Sons; 1979.