Is that correlation really significant? The author shows how one statistical parameter can give misleading answers

The coefficient of determination exposed!

Gerald J. Hahn

The accessibility to computers, especially time-sharing varieties, has made regression analysis a frequently used tool for estimating the relationship between an observed response (dependent variable) and factors (independent variables) that may be related to the response. As a result, regression analysis is used by statistician and non-statistician alike. However, after fitting a regression equation to his data the user faces the question of how good a job the fitted equation does. This question is often answered by examination of the quantity known as the coefficient of determination, R^2, or its square root, the coefficient of multiple correlation, which can be generated by many computer programs.

The coefficient of determination, its interpretation, *and its limitations,* are the subject of this article.

What is the coefficient of determination?

The coefficient of determination is that proportion of the total variability in the dependent variable that is accounted for by the regression equation in the independent variable(s). A value for R^2 of 1 indicates that the fitted regression equation accounts for all the variability of the values of the dependent variable in the sample data. At the other extreme, a value of 0 for R^2 indicates that the regression equation accounts for none of the variability. Further elaboration is provided in Appendix 1.

Does a high value of R^2 assure a statistically significant regression equation?

Not necessarily! In fact, one can obtain a value of 1 for R^2 by simply fitting a regression equation that includes as many (statistically estimable) terms as there are observations (i.e., data points). When the number of observations exceeds the number of terms in the regression equation by only a small number then the coefficient of determination might be large, even if there is no true relationship between the independent and dependent variables. For example, the chances are one in ten of obtaining a value of R^2 as high as 0.9756 in fitting a simple linear regression equation to the relationship between an independent variable X and a normally distributed dependent variable Y based on only 3 observations, even if X is totally unrelated to Y, i.e., this result can occur 10% of the time, even if the two variables are unrelated. On the other hand, with 100 observations a coefficient of determination of 0.07 is sufficient to establish statistical significance of a linear regression at the 1% level.

More generally, Table 1 indicates the values of R^2 required to establish statistical significance for a simple linear regression equation. This tabulation gives values at the 10%, 5%, and 1% significance levels, corresponding, respectively, to the situations where one is ready to take one chance in ten, one chance in twenty, and one chance in a hundred of *incorrectly* concluding there is evidence of a statistically significant linear regression relationship when in fact X and Y are unrelated. The previously quoted statistically significant values for R^2 for 3 and 100 observations were taken from Table 1. *Note that Table 1 applies only for a simple linear regression equation.* For the case of multiple regression, statistical significance of the overall regression equation can be determined by the F-ratio in the analysis-of-variance table [see Draper and Smith (1966)]. Such a table is contained in many computer programs for regression analysis.

Gerald J. Hahn, who is Manager of the Statistics Program at GE's Research and Development Center has, since 1955, provided consulting, development, and problem-solving services on a variety of statistical problems throughout the company. He also is an adjunct professor in Union College's Graduate Studies Division. His previous employment was with Biow Advertising, and the United States Army Chemical Corps. Dr. Hahn holds degrees from City College of New York (1952, BBA), Columbia University (1953, MS statistics), Union College (1965, MS mathematics) and Rensselaer Polytechnic Institute (1971, PhD, operations research and statistics). He has published extensively in the statistical and engineering literature, and is coauthor of *Statistical Models in Engineering*, (John Wiley, 1967). He is a member of the Institute of Mathematical Statistics, the American Statistical Association, and the American Society for Quality Control and is a recipient of the ASQC Jack Youden Prize. He will shortly succeed Dr. Youden as a regular author of an applied statistics column in an ACS periodical—namely CHEMTECH.

Reprinted from **CHEMICAL TECHNOLOGY**, Vol. 3, No. 10, October 1973, Page 609

Table 1. Values of R^2 required to establish the statistical significance of a simple regression equation for various sample sizes

Sample size	Statistical significance level		
	10%	5%	1%
3	0.9756	0.9938	0.9998
4	0.810	0.903	0.980
5	0.65	0.77	0.92
6	0.53	0.66	0.84
7	0.45	0.57	0.77
8	0.39	0.50	0.70
9	0.34	0.44	0.64
10	0.30	0.40	0.59
11	0.27	0.36	0.54
12	0.25	0.33	0.50
13	0.23	0.31	0.47
14	0.21	0.28	0.44
15	0.19	0.26	0.41
20	0.14	0.20	0.31
25	0.11	0.16	0.26
30	0.09	0.13	0.22
40	0.07	0.10	0.16
50	0.05	0.08	0.13
100	0.03	0.04	0.07

Does a statistically significant R^2 assure a useful regression equation?

Not necessarily! *Practical* significance and statistical significance are not equivalent. With a small sample, it is possible not to obtain any evidence of a statistically significant regression relationship between two variables even if their true relationship is quite strong. This is because, as seen above, a relatively high value of R^2 is required to show a regression equation to be statistically significant when only a small number of observations are used. On the other hand, a regression equation based on only a modest (and practically unimportant) true relationship may be established as statistically significant if a sufficiently large number of observations are available. For example, it was seen that with 100 observations, a value for R^2 of 0.07 was sufficient to establish a highly significant statistical linear relationship between two variables.

What if R^2 is both large and of statistical significance?

That's better news, but it still may not be sufficient to assure a really useful regression equation, especially if the equation is to be used for prediction. One reason is that *the coefficient of determination is not expressed on the same scale as the dependent variable Y.* A particular regression equation may explain a large proportion of the total variability in the dependent variable, thus yielding a high value of R^2. Yet the *total* variability may be so large that the remaining unexplained variability left after the regression equation is fitted will still be larger than is tolerable for useful prediction. Thus, it is not possible to tell solely from the magnitude of R^2 how accurate the predictions will be.

Furthermore, the magnitude of R^2 depends directly on the range of variation of the independent variables for the given data. The coefficient of determination thus decreases with a decrease in the range of variation of the independent variables, assuming the correct regression model is being fitted to the data. For example, Figure 1*a* shows the fitted regression equation between an independent variable, X, and a dependent variable, Y, based on 110 equally spaced values of X over the range from 10 to 20. The estimated coefficient of determination is $R^2 = 0.89$. However, if one had available only the 30 observations in the range 14 to 16 (see Figure 1*b*), the resulting coefficient of determination from the fitted regression equation would be only $R^2 = 0.21$. (Further details concerning this example are given in Appendix 2.)

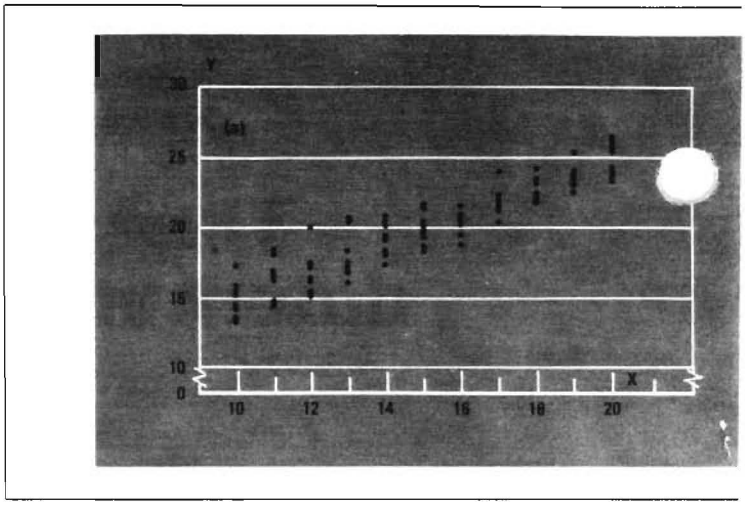


Figure 1. Data plots with $R^2 = 0.89$ (*a*), and $R^2 = 0.21$ (*b*)

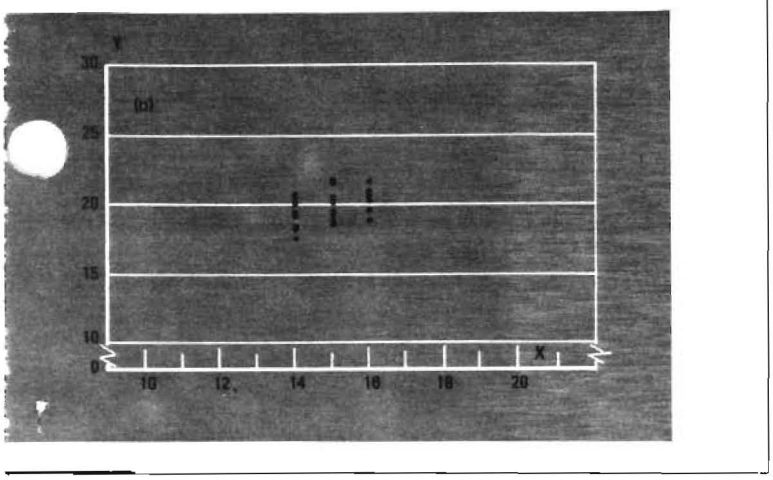
Thus a large value of R^2 might reflect the fact that the data had been obtained over an unrealistically large range of variation. Conversely, a small value of R^2 might be due to the limited range of the independent variables. This is sometimes the case in analyzing data from a manufacturing process in which normal plant practice restricts the range of the process variables.

Note also that a large and statistically significant coefficient of determination does not assure that the chosen regression model adequately represents the true relationship for all purposes. A coefficient of determination of 0.99, even if statistically significant, for a regression model involving only linear terms for each of k independent variables, does not mean that a model that also includes quadratic and interaction terms could not conceivably yield a significantly better fit, nor that the "real cause variables" have been included in the regression equation.

Standard error of estimate

Actually, a more useful single measure of the prediction capability of a k-variable regression equation than the coefficient of determination is the sample standard deviation of the values of the dependent variable about the regression equation. This statistic, known as the standard error of estimate, is obtained by many computer programs for regression analysis.

The standard error of estimate enters directly into expressions for obtaining a variety of statistical in-



tervals. These important intervals, which are also calculated by many computer programs, often contain more meaning than the common R^2. Roughly speaking, their meaning is as follows.

Statistical intervals

1. **A confidence interval for the dependent variable** is an interval that one expects with a specified degree of confidence to contain the *average value* of the dependent variable at a set of specified values of the independent variables. For example, assume that a linear regression relationship has been fitted between height (independent variable) and weight (dependent variable) based on measuring a random sample of 100 people from a given population. A 95% confidence interval for the average weight of 6-ft individuals would then be an interval constructed from the fitted regression relationship that one would expect, with 95% confidence, to contain the *average weight* of all 6-ft individuals in the population. (More precisely, a 95% confidence interval contains the true average value of the dependent variable approximately 95% of the time, if such intervals are constructed in many independent regression analyses.)

2. **A prediction interval** for the dependent variable is an interval that one expects with a specified degree of confidence to contain a *single future value* of the dependent variable from the sampled population at a set of specified values of the independent variables. Thus, in the preceding example, a 95% prediction interval for the weight of a single, future 6-ft individual is an interval constructed from the fitted regression relationship that one would expect with 95% confidence to contain the weight of a randomly selected *single future individual* who is six feet tall.

3. **A confidence interval around a regression coefficient** is an interval that one expects, with a specified degree of confidence, to contain the *true regression coefficient*. Thus in the preceding example, a 95% confidence interval to contain the regression coefficient is an interval constructed from the fitted linear regression relationship that one would expect with 95% confidence to contain the average unit increase in weight per unit increase in height. Since regression coefficients often have physical meaning, this is a particularly useful statistic in chemical technology. Linear free-energy relationships provide many examples: $\ln x = \Delta H/RT + \Delta S$. Here the "coefficients" ΔH and ΔS are subject to physical interpretation.

The exact method for obtaining the preceding intervals, using the standard error of estimate, and more detailed discussions of their interpretations are provided in texts such as Draper and Smith (1966). Which of these intervals one need be concerned with in practice depends on the specific application at hand.

A warning—Indiscriminate use of regression analysis can be hazardous

There are many hazards in the application of regression analysis of which the user needs to beware. These hazards include drawing conclusions concerning cause and effect (all we've "proven" is *correlation,* not *causality*) taking remedial action as a result of such conclusions, and extrapolating beyond the range of the given data. The extent of the hazard depends on the use to which the results of the regression analysis are being put. A particular regression equation may be appropriate to make predictions of further observations from the population from which the given data were obtained, but may be quite unsafe for other purposes, such as explaining the "reasons" for the variability in the dependent variable [see Box (1966) and Hahn and Shapiro (1966)].

Concluding remarks

The coefficient of determination, R^2, measures the proportion of variability in the dependent variable that is accounted for by the regression equation. Since R^2 is independent of scale, it can be used to describe the results of the regression analysis without requiring knowledge of the nature of the dependent variable. On the other hand, unlike the standard error of estimate and derived confidence intervals, R^2 alone does *not* provide direct information as to how well the regression equation can be used for prediction.

This article is limited to a discussion of one aspect of interpreting the results of a regression analysis. Other aspects and further details are provided in the excellent books of Daniel and Wood (1971) and Draper and Smith (1966). For example, an important part of such interpretations is to obtain various graphical displays of the data and of the residual variation *after* fitting the regression equation, [see Anscombe (1973), and the foregoing].

Acknowledgment. The author wishes to express his appreciation to Dr. Richard L. Shuey, of General Electric Corporate Research and Development for his support and encouragement of this work and to Dr. Paul Feder and Dr. Wayne Nelson for their constructive comments which led to important improvements.

Author's address: Research and Development Center, General Electric Co., P.O. Box 43, Schenectady, N.Y. 12301.

References

Anscombe, F. J., Graphs in statistical analysis, *Amer. Statist.*, 27 (1), 17 (1973).

Box, G. E. P., Use and abuse of regression, *Technometrics*, 8, 625 (1966).

Daniel, C. and Wood, F. S., *Fitting Equations to Data*, John Wiley and Sons, New York, N.Y., 1971.

Draper, N. R., and Smith, H., *Applied Regression Analysis*, John Wiley and Sons, Inc., New York, N.Y., 1966.

Hahn, G. J., and Shapiro, S. S., The use and misuse of multiple regression analysis, *Ind. Qual. Contr.*, 23, 184 (1966).

Appendix 1. On the meaning of R^2

Without any information about the independent variables, the best prediction for a future value of the dependent variable Y is $\bar{Y}$, the sample average of the n past observations. For a particular past value, Y_i, the error in using $\bar{Y}$ as a predictor is the deviation (initial variation) $Y_i - \bar{Y}$ (see Figure 2); and a measure of the total variability for the n given observations, Y_i, $i = 1, \ldots, n$, is the sum of the squares of the deviations around $\bar{Y}$:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

When the associated values of the independent variables are known, the best estimate of a future value for Y is the value $\hat{Y}$ obtained from the fitted regression equation. This leads to a prediction error (or residual variation) $Y_i - \hat{Y}_i$ for the ith given observation (see Figure 2). Thus, a measure of the variability in Y that remains after the regression equation is fitted is given by the sum of the squares of the deviations around the $\hat{Y}_i$'s:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

The ratio of the preceding two quantities,

$$\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

is the proportion of the total variability unexplained by the fitted regression equation. The coefficient of determination, R^2, is the preceding quantity subtracted from 1. Thus, the proportion of the total variability accounted for by the regression equation is:

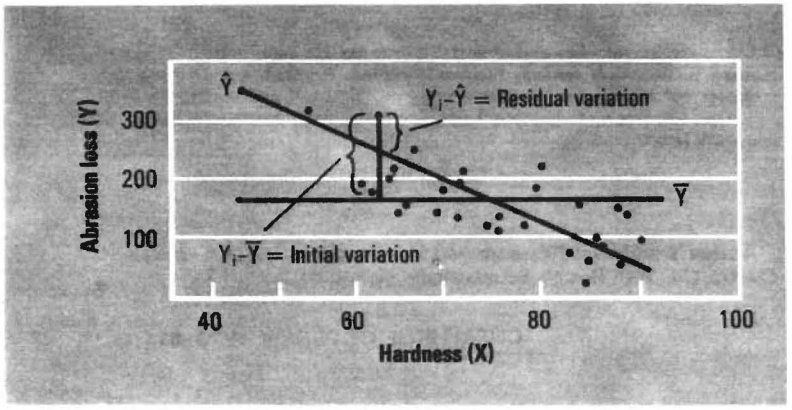
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

$$= \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

The values R^2, s (the sample standard deviation of the original values of Y), and s_E (the standard error of estimate of the regression equation) are related as follows:

$$R^2 = 1 - \frac{s_E^2}{s^2}\left(\frac{n - k}{n - 1}\right)$$

Figure 2. Use of regression line vs. sample mean to predict Y



where, k denotes the number of terms (including the intercept) in the fitted regression equation and

$$s_E = \left[\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n - k}\right]^{1/2}$$

Appendix 2. About the effect on R^2 of the range of the independent variables in the regression equation

The observations in Figure 1 were randomly generated, using the model

$$Y = 5 + 1X + e$$

where the components of random variation e are independently, normally distributed about the regression equation with mean 0 and standard deviation $\sigma_E = 1$ (estimated from a fitted regression equation by s_E).

More generally, it can be shown that for the relationship

$$Y = \beta_0 + \beta_1 X + e$$

the expected value of R^2 from a simple linear regression fit approximately equals

$$\frac{\beta_1^2 s_X^2}{\beta_1^2 s_X^2 + \sigma_E^2}$$

where

$$s_X^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/n \qquad \left(\bar{X} = \sum_{i=1}^{n}X_i/n\right)$$

and X_i, $i = 1, \ldots, n$ are the values of the independent variable X. This expression leads to approximate expected values for R^2 of 0.91 and 0.40 for the data generated in Figures 1*a* and 1*b*, respectively, as compared to the calculated values of 0.89 and 0.21. Thus both the expected and observed values of R^2 decrease sharply with a decrease in the variation of the independent variable for a simple linear regression equation. Similar results apply for multiple regression equations.

Questions to test the reader's understanding

The following data have been obtained on yearly sales:

Year (X)	*Total sales in $ millions (Y)*
1969	8.0
1970	10.0
1971	10.0
1972	8.0

Without performing any calculations, answer these questions.

1. The following model is fitted to the data:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + e$$

What is the resulting value of R^2?

2. The following model is now fitted to the data:

$$Y = \beta_0 + \beta_1 X + e$$

What is the resulting value of R^2?

The answers to these questions are hidden elsewhere in this issue—hidden to encourage readers to do "their homework" before peeking.