The American Statistician
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utas20

The Inequality Between the Coefficient of Determination and the Sum of Squared Simple Correlation Coefficients

Gwowen Shieh[a]

[a] Gwowen Shieh is Associate Professor, Department of Management Science, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan 30050, Republic of China . This work was partially supported by National Science Council of Taiwan under contract number NSC89-2118-M-009-016. The author thanks the associate editor and referees for helpful comments that improved the presentation of this article.
Published online: 01 Jan 2012.

The Inequality Between the Coefficient of Determination and the Sum of Squared Simple Correlation Coefficients

Gwowen SHIEH

The inequality between the coefficient of determination and the sum of two squared simple correlation coefficients in a two-variable regression model is reexamined through two relative measures. They are the relative coefficient of determination and the relative simple correlation, which are the ratio of the coefficient of determination to the sum of squares of the two simple correlations and the ratio of two simple correlations, respectively. This approach not only permits new insights into their relationship but also allows clear and informative visual representations of various aspects of the counterintuitive condition. We consider the occurrence and corresponding magnitude, probability, and expected magnitude of the enhancement-synergism situation. Numerical examples are presented to illustrate these phenomena.

KEY WORDS: Coefficient of determination; Multiple regression; Simple correlation coefficient.

1. INTRODUCTION

Hamilton (1987) discussed the counterintuitive nature of multivariate relationship in standard multiple regression models—the coefficient of determination can exceed the sum of the squared correlation coefficients between the response variable and each explanatory variable. To understand this interesting and surprising feature, the theoretical proof and geometrical argument that such inequality can occur were provided for the regression model with two explanatory variables. A slightly simpler proof was given by Bertrand and Holder (1988). Related comments and discussions can be found in Currie and Korabinski (1984), Freund (1988), Hamilton (1988), Mitra (1988), Cuadras (1993) and their references. As visual supplement, Currie and Korabinski (1984) and Freund (1988) contain several diagrams that intend to illustrate when the counterintuitive conditions can occur. Since more than three measures are involved in those plots, their uses are limited to the selected conditional values of the chosen measure. Consequently, the interrelation is not completely shown by single or even several plots together. Hence more concise diagrams are needed to effectively conceive and evaluate the occurrences and magnitudes of these phenomena. Although the existence of the inequality is well presented and recognized in the aforementioned articles, it is still not clear exactly how often it can happen.

The purpose of this article is to provide more informative plots for the existence and subsequent analyses of the inequality and to quantify the probability and expected magnitude of the occurrence.

2. MAIN RESULTS FOR TWO-VARIABLE REGRESSION

Consider the standard regression model with one response variable Y and two explanatory variables X_1 and X_2

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where α, β_1, and β_2 are parameters, and ε_i are iid $N(0, \sigma^2)$ random variables. Let r_{Y1}, r_{Y2}, and r_{12} be the usual simple (product-moment) correlation coefficients between Y and X_1, Y and X_2, and X_1 and X_2, respectively. It follows from Equation (8) of Hamilton (1987) or Equation (3-81) of Johnston (1991) that the coefficient of determination R^2 in terms of the simple correlation coefficients is

$$R^2 = \frac{r_{Y1}^2 + r_{Y2}^2 - 2r_{Y1}r_{Y2}r_{12}}{1 - r_{12}^2}. \tag{2}$$

We will focus on the inequality between the coefficient of determination and the sum of two squared simple correlation coefficients

$$R^2 > r_{Y1}^2 + r_{Y2}^2. \tag{3}$$

The inequality may seem surprising or counterintuitive at first. It will be shown later that it occurs more often than one may think. Currie and Korabinski (1984) called such an occurrence "enhancement," while Hamilton (1988) suggested "synergism" as an alternative. Here it will be termed "enhancement-synergism."

2.1 When Does the Enhancement-Synergism Condition Hold?

Since it is obvious from (2) that $R^2 = r_{Y1}^2 + r_{Y2}^2$ for $r_{12} = 0$, we will assume $r_{12} \neq 0$ in the remainder of this article. It is shown in Hamilton (1987) that the necessary and sufficient condition for (3) in terms of the simple correlation coefficients r_{Y1}, r_{Y2}, and r_{12} is

$$r_{12}\left(r_{12} - \frac{2r_{Y1}r_{Y2}}{r_{Y1}^2 + r_{Y2}^2}\right) > 0. \tag{4}$$

Accordingly, the validity of inequality (4) depends upon the interrelation of r_{Y1}, r_{Y2}, and r_{12}. In this case graphical representation is extremely informative in understanding its occurrence. To lay the basis for developing a simplified view and providing a concise visualization of when the inequality can occur, we first define the relative simple correlation, denoted by Q, as the ratio

Gwowen Shieh is Associate Professor, Department of Management Science, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan 30050, Republic of China (E-mail: gwshieh@cc.nctu.edu.tw). This work was partially supported by National Science Council of Taiwan under contract number NSC89-2118-M-009-016. The author thanks the associate editor and referees for helpful comments that improved the presentation of this article.

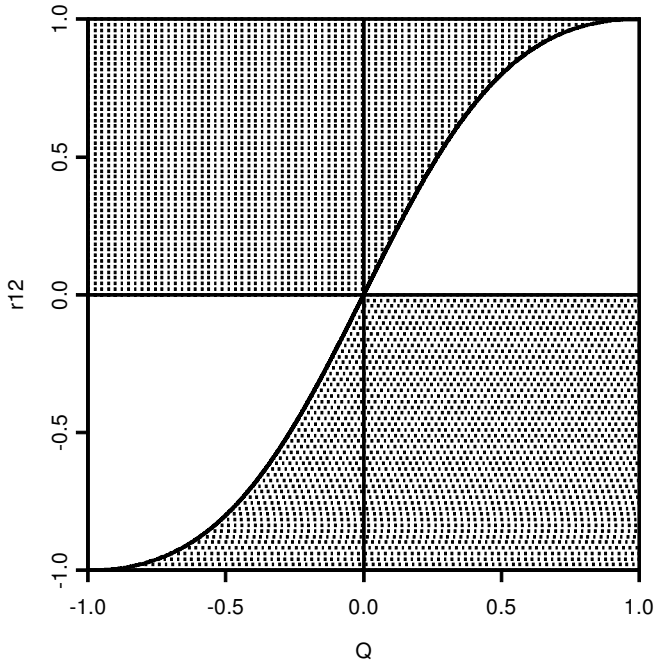

Figure 1. The Regions of Enhancement-Synergism.

of r_{Y2} to r_{Y1}:

$$Q = r_{Y2}/r_{Y1}. \tag{5}$$

Since the designation of X_1 and X_2 is arbitrary, as long as only one of r_{Y1} and r_{Y2} is zero, Q can be set as zero. The case that both r_{Y1} and r_{Y2} are zero will be excluded because R^2 is obviously zero from (2).

Equation (5) enables us to rewrite (4) in terms of Q and r_{12}, and one gets $r_{12} > g(Q)$ for $r_{12} > 0$ and $r_{12} < g(Q)$ for $r_{12} < 0$, where $g(Q) = 2Q/(1+Q^2)$. Equivalently, $Q < q_0$ or $Q > 1/q_0$ for $r_{12} > 0$, and $Q < 1/q_0$ or $Q > q_0$ for $r_{12} < 0$, where $q_0 = (1 - \sqrt{1-r_{12}^2})/r_{12}$. Note that $g(Q) = g(1/Q)$, for $Q \neq 0$. Therefore, the relation between r_{12} and Q can be represented by the relation between r_{12} and Q for $|Q| \leq 1$. Figure 1 presents the occurrence of (4) for combinations of r_{12} and Q for $|Q| \leq 1$. Those four shaded areas stand for the occurrence regions of enhancement-synergism. We believe Figure 1 is more effective for communicating the results than Figure 2 of Currie and Korabinski (1984) where the occurrence of enhancement-synergism was presented with multiple conditional plots (combinations of r_{Y2} and r_{12} for selected values of r_{Y1}).

2.2 What is the Magnitude of Enhancement-Synergism?

Instead of measuring the magnitude of enhancement-synergism with the direct difference of R^2 and $r_{Y1}^2 + r_{Y2}^2$, for the purpose of simplification, we suggest using the relative coefficient of determination, denoted by H,

$$H = R^2/(r_{Y1}^2 + r_{Y2}^2), \tag{6}$$

which is the ratio of R^2 to $r_{Y1}^2 + r_{Y2}^2$. A useful expression for H is

$$H = \frac{1 + Q^2 - 2r_{12}Q}{(1 - r_{12}^2)(1 + Q^2)} \tag{7}$$

as a function of Q and r_{12}. A close examination of (7) shows that H attains its maximum $1/(1-|r_{12}|)$ and minimum $1/(1+|r_{12}|)$ for $Q = -\text{sign}(r_{12})$ and $\text{sign}(r_{12})$, respectively. The function $\text{sign}(x)$ returns respective value of 1, 0, or -1 if $x > 0$, $x = 0$, or $x < 0$. Besides, $H(Q) = H(1/Q)$ for $Q \neq 0$. As in the previous discussion of $g(Q)$, the relation between H and Q can be represented by the relation between H and Q for $|Q| \leq 1$. To visualize these facts, we plot H against Q for $|Q| \leq 1$ in Figure 2 for three different values $r_{12} = .1$, .5, and .9. (The plots for negative values of r_{12} are mirror images of those for positive values.) We believe Figure 2 provides a clearer presentation of the relative magnitude and range of the enhancement-synergism condition than the plots in Freund (1988) and Figure 1 of Currie and Korabinski (1984).

2.3 How Often Does the Enhancement-Synergism Condition Occur and What is the Expected Magnitude of Enhancement-Synergism?

Because we know the enhancement-synergism condition exists, it is of interest to know how often it occurs and what is the expected magnitude in terms of relative coefficient of determination; that is, what are $P(H > 1)$ and $E[H]$? To determine these two quantities, we start with the derivation of the pdf of Q. Assume the true coefficient parameters of β_1 and β_2 in (1) are β_1^* and β_2^*, respectively. It can be shown from the standard assumptions that $Q = Z_2/Z_1$, where $Z_j = S_{Yj}/(\sigma S_j) \sim N(\mu_j, 1), S_{Yj} = \sum_{i=1}^n (Y_i - \overline{Y})(X_{ji} - \overline{X}_j), S_j$ is the square root of $S_j^2 = \sum_{i=1}^n (X_{ji} - \overline{X}_j)^2$, $j = 1$ and 2, $\mu_1 = (\beta_1^* S_1 + \beta_2^* S_2 r_{12})/\sigma$ and $\mu_2 = (\beta_2^* S_2 + \beta_1^* S_1 r_{12})/\sigma$. Note that $\text{corr}(Z_1, Z_2) = r_{12}$. Hence the distribution of Q is exactly the same as the distribution of the ratio of two correlated normal random variables with mean (μ_1, μ_2), variance (1, 1) and correlation r_{12}. This is a special case of the ratio of two correlated normal random variables discussed by Fieller (1932) and Hinkley (1969). Its explicit pdf and cdf were given by Hinkley (1969, eq. (1)–(3)). It appears, however, that there is no analytic form for $P(H > 1) = 1 - P(q_0 < Z_2/Z_1 < 1/q_0)$ for

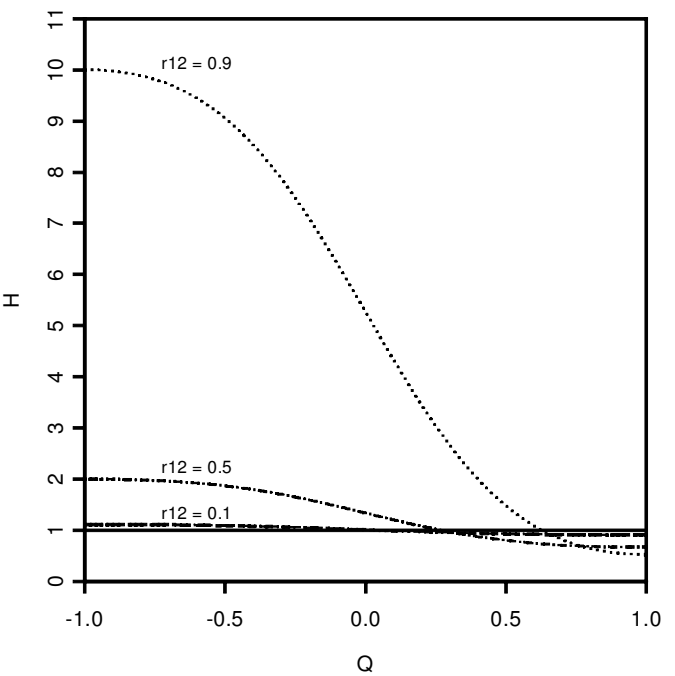

Figure 2. The Relation Between H and Q.

Table 1. The Occurrence Probability of Enhancement-Synergism and the Expected Relative Coefficient of Determination

		$P(H > 1 \mid \mu_1, \mu_2, r_{12})$			$E\{H \mid \mu_1, \mu_2, r_{12}\}$		
μ_1	μ_2	$r_{12} = .1$	$r_{12} = .5$	$r_{12} = .9$	$r_{12} = .1$	$r_{12} = .5$	$r_{12} = .9$
−2	−2	.0549	.0972	.1360	.9329	.7881	.8951
−2	−1	.1992	.3129	.6875	.9583	.9670	2.5033
−2	−1	.5383	.7192	.9925	1.0100	1.3343	5.4156
−2	1	.8508	.9514	1.0000	1.0607	1.6597	7.5596
−2	2	.9656	.9953	1.0000	1.0854	1.8107	8.5373
−1	−2	.1992	.3129	.6875	.9583	.9670	2.5033
−1	−1	.2824	.3284	.3586	.9707	.9922	1.6872
−1	0	.5137	.5864	.8683	1.0071	1.2358	4.2373
−1	1	.7507	.8551	.9984	1.0453	1.5342	6.7198
−1	2	.8508	.9514	1.0000	1.0607	1.6597	7.5596
0	−2	.5383	.7192	.9925	1.0100	1.3343	5.4156
0	−1	.5137	.5864	.8683	1.0071	1.2358	4.2373
0	0	.5000	.5000	.5000	1.0050	1.1547	2.2942
0	1	.5137	.5864	.8683	1.0071	1.2358	4.2373
0	2	.5383	.7192	.9925	1.0100	1.3343	5.4156
1	−2	.8508	.9514	1.0000	1.0607	1.6597	7.5596
1	−1	.7507	.8551	.9984	1.0453	1.5342	6.7198
1	0	.5137	.5864	.8683	1.0071	1.2358	4.2373
1	1	.2824	.3284	.3586	.9707	.9922	1.6872
1	2	.1992	.3129	.6875	.9583	.9670	2.5033
2	−2	.9656	.9953	1.0000	1.0854	1.8107	8.5373
2	−1	.8508	.9514	1.0000	1.0607	1.6597	7.5596
2	0	.5383	.7192	.9925	1.0100	1.3343	5.4156
2	1	.1992	.3129	.6875	.9583	.9670	2.5033
2	2	.0549	.0972	.1360	.9329	.7881	.8951

$r_{12} > 0$, or $1 - P(1/q_0 < Z_2/Z_1 < q_0)$ for $r_{12} < 0$, except in the following special case.

Assume $\beta_1^* = \beta_2^* = 0$, or equivalently $\mu_1 = \mu_2 = 0$. In this case, Q can be viewed as the ratio of two correlated standard normal variables and has a Cauchy distribution with location parameter $\theta = r_{12}$ and scale parameter $\lambda = (1 - r_{12}^2)^{1/2}$; see Johnson, Kotz, and Balakrishnan (1994, eq. 16.1). Since its cdf is of the form $F_Q(q) = .5 + \nu^{-1}\tan^{-1}\{(q - \theta)/\lambda\}$, one has

$$P(H > 1) = 1 - |F_Q(q_0) - F_Q(1/q_0)| = 1 - |\tan^{-1}(-q_0) - \tan^{-1}(1/q_0)|/\nu = .5.$$

The last equality follows from the fact that $-q_0 \times 1/q_0 = -1$. Therefore it is equally likely to have or not to have enhancement-synergism in a two-variable regression with explanatory variables that are absolutely irrelevant for describing the response variable. More importantly, this is true for all $r_{12} \neq 0$. This outcome may be easy to guess; however, it is not as trivial as one may think.

The actual occurrence probability of enhancement-synergism under various values of (μ_1, μ_2) are calculated through numerical integration and are listed in Table 1 for $r_{12} = .1, .5$, and .9. In general, $P(H > 1|\mu_1, \mu_2, r_{12}) = P(H > 1|-\mu_1, -\mu_2, r_{12})$ and $P(H > 1|\mu_1, -\mu_2, r_{12}) = P(H > 1|-\mu_1, \mu_2, r_{12})$ due to symmetry. It is also true that $P(H > 1|\mu_1, \mu_2, r_{12}) = P(H > 1|\mu_1, -\mu_2, -r_{12})$.

Based on the pdf of Q and (7), we can evaluate the expected relative coefficient of determination $E[H]$ for any $r_{12} \neq 0$. Again it does not appear to have a simple analytic form and numerical integration is necessary to carry out the expectation. Table 1 also presents the expected magnitude of enhancement-synergism in terms of H for $r_{12} = .1, .5$, and .9. In particular, when $(\mu_1, \mu_2) = (0, 0)$, the values are 1.0050, 1.1547, and 2.2942 for $r_{12} = .1, .5$, and .9, respectively. This indicates that the relative coefficient of determination is greater than one in "average" for all $r_{12} \neq 0$ when $(\mu_1, \mu_2) = (0, 0)$. So far we are unable to prove that $E[H|\mu_1 = 0, \mu_2 = 0, r_{12}] > 1$ for all $r_{12} \neq 0$. Moreover, Table 1 shows that the differences of $E[H]$ among different values of (μ_1, μ_2) are more dramatic as $|r_{12}|$ gets larger. Overall $E[H|\mu_1, \mu_2, r_{12}] = E[H| - \mu_1, -\mu_2, r_{12}], E[H|\mu_1, -\mu_2, r_{12}] = E[H| - \mu_1, \mu_2, r_{12}]$ and $E[H|\mu_1, \mu_2, r_{12}] = E[H|\mu_1, -\mu_2, -r_{12}]$.

3. CONCLUSION

We provide a simplified and systematic view of the counterintuitive inequality or the enhancement-synergism condition that the coefficient of determination can exceed the sum of two squared simple correlation coefficients. The major difference between our approach and others is that the relative simple correlation Q and the relative coefficient of determination H are the primary tools for analyzing such phenomenon rather than the original measures and their difference. The following four major questions are studied: (1) When does the enhancement-synergism condition occur? (2) What is the magnitude of enhancement-synergism? (3) How often does the enhancement-synergism condition occur? (4) What is the expected magnitude of enhancement-synergism? Both theoretical arguments and graphical presentations are given. Numerical examples are provided to illustrate the levels of the enhancement-

synergism and its dependence on the other measures. In addition to the surprising enhancement-synergism condition itself, we point out two interesting features when the explanatory variables are absolutely irrelevant for describing the response variable. It is shown that even the two true coefficient parameters are indeed zero ($\beta_1^* = \beta_2^* = 0$) the occurrence probability of the enhancement-synergism condition is .5 for all $r_{12} \neq 0$. Furthermore, under the same assumption, it is shown numerically that the expected relative coefficient of determination appears to be greater than 1 for all $r_{12} \neq 0$ and is increasing with $|r_{12}|$.

[Received October 1999. Revised April 2000.]

REFERENCES

Bertrand, P. V., and Holder, R. L. (1988), "A Quirk in Multiple Regression: The Whole Regression can be Greater Than the Sum of its Parts," *The Statistician*, 37, 371–374.

Cuadras, C. M. (1993), "Interpreting an Inequality in Multiple Regression," *The American Statistician*, 47, 256–258.

Currie, I., and Korabinski, A. (1984), "Some Comments on Bivariate Regression," *The Statistician*, 33, 283–293.

Fieller, E. C. (1932), "The Distribution of the Index in a Normal Bivariate Population," *Biometrika*, 24, 428–440.

Freund, R. J. (1988), "When is $R^2 > r_{y1}^2 + r_{y2}^2$ (Revisited)," *The American Statistician*, 42, 89–90.

Hamilton, D. (1987), "Sometimes $R^2 > r_{yx1}^2 + r_{y2}^2$, Correlated Variables are not Always Redundant," *The American Statistician*, 41, 129–132.

——— (1988), "Sometimes $R^2 > r_{yx1}^2 + r_{yx2}^2$, Correlated Variables are not Always Redundant" (Reply), *The American Statistician*, 42, 90–91.

Hinkley, D. V. (1969), "On the Ratio of Two Correlated Normal Random Variables," *Biometrika*, 56, 635–639.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994), *Distributions in Statistics: Continuous Univariate Distributions I*, New York: Wiley.

Johnston, J. (1991), *Econometric Methods*, New York: McGraw-Hill.

Mitra, S. (1988), "The Relationship Between the Multiple and the Zero-Order Correlation Coefficients," *The American Statistician*, 42, 89.