한국통계학회논문집 제2권 1호
1995년 4월 pp. 32-47

A Study on the Coefficient of Determination in Linear Regression Analysis

S.H. Park[1] and Sung-im Lee[2]

Abstract

The coefficient of determination R^2, as the proprtation of y explained by a set of independent variables $x_1, x_2, \cdots, x_k$ through a linear regression model, is a very useful tool in linear regression analysis. Suppose $R^2{}_{yx_i}$ is the coefficient of determination when y is regressed only on x_i alone. If the independent variables are correlated, the sum, $R^2{}_{yx_1} + R^2{}_{yx_2} + \cdots + R^2{}_{yx_k}$, is not equal to $R^2{}_{yx_1x_2\cdots x_k}$, where $R^2{}_{yx_1x_2\cdots x_k}$ is the coefficient of determination when y is regressed simultaneously on $x_1, x_2, \cdots, x_k$. In this paper it is discussed that under what conditions the sum is greater than, equal to, or less than $R^2{}_{yx_1x_2\cdots x_k}$, and then the proofs for these conditions are given. Also illustrated examples are provided. In addition, we will discuss about inequality between $R^2{}_{yx_1x_2\cdots x_k}$ and the sum, $R^2{}_{yx_1} + R^2{}_{yx_2} + \cdots + R^2{}_{yx_k}$.

1. INTRODUCTION

Suppose we have a general linear regression model

$$y_j = \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj} + \varepsilon_j \; j=1,2, \cdots ,n \qquad (1)$$

where y_j is the jth response for the dependent variable y, $(x_{1j}, x_{2j}, \cdots, x_{kj})$ are the jth values of k independent variables, $\beta_i' s$ are regression coefficients and $\varepsilon_j' s$ are error terms which are identically and independently distributed with mean zero and variance σ^2.

We assume that all variables are standardized such that

$$\sum_j y_j = 0 \qquad \sum_j y_j^2 = 1,$$

[1] Department of Computer Science and Statistics, Seoul National University, Seoul, 151-742, KOREA.

[2] Department of Computer Science and Statistics, Seoul National University, Seoul, 151-742, KOREA

$$\sum_j x_{ij}=0, \qquad \sum_j x_{ij}^2=1, \ i = 1, 2, \cdots, k . \tag{2}$$

Note that in the model (1) there is no intercept term β_0, since all variables are standardized.

In regression analysis the coefficient of determination R^2 is perhaps one of the most often quoted statistics in the modeling of relationships between a dependent variable and a set of independent variables. This coefficient, specific to a given response variable y, is generally defined as the proportion of the variation of y explained by a set of independent variables $x_1, x_2, \cdots, x_k$. As a general rule, the addition of variables will remove unexplained variation in the response, that is, as the number of independent variables increases the value of R^2 in regression equation will increase. Let us consider first the simple case when there are only two independent variables.

It is desirable to consider the value of R^2 in the following case :

Regression Equation	Coefficient of Determination	
$y=\alpha_1 x_1+\varepsilon_1$	$R^2_{yx_1}=r^2_{yx_1}$	(3)
$y=\alpha_2 x_2+\varepsilon_2$	$R^2_{yx_2}=r^2_{yx_2}$	(4)
$y=\beta_1 x_1+\beta_2 x_2+\varepsilon$	$R^2_{yx_1x_2}=\dfrac{1}{1-r^2_{x_1x_2}}(r^2_{yx_1}+r^2_{yx_2}-2r_{x_1x_2}r_{yx_1}r_{yx_2})$	(5)

where $r_{x_1x_2}$ is the sample correlation coefficient between x_1 and x_2, r_{yx_i} is the sample correlation coefficient between x_i and y. When there is only one independent variable in the regression equation, the coefficient of determination is $r^2_{yx_i}$, the square of the simple correlation coefficient between y and x_i as shown in the above. The results of R^2 in these equations are obtained in Appendix A. If the independent variables, x_1 and x_2, are uncorrelated, then the sum, $R^2_{yx_1}+R^2_{yx_2}$, is $R^2_{yx_1x_2}$. But, what if they are correlated? It is highly probable that the sum, $R^2_{yx_1}+R^2_{yx_2}$, is larger than $R^2_{yx_1x_2}$. However, it is possible that the sum, $R^2_{yx_1}+R^2_{yx_2}$, may be smaller than $R^2_{yx_1x_2}$ for some particular situation. For this latter situation, Hamilton(1987) gave a necessary and sufficient condition, which is

$$r_{x_1x_2}\,(r_{x_1x_2} - \frac{2r_{yx_1}r_{yx_2}}{r^2_{yx_1} + r^2_{yx_2}}) > 0. \tag{6}$$

He mentioned that the inequality $R^2{}_{yx_1x_2} > R^2{}_{yx_1} + R^2{}_{yx_2}$ issues from multicollinearity. In this paper, we will mainly consider the problem of making a comparison between $R^2{}_{yx_1x_2}$ and the sum, $R^2{}_{yx_1} + R^2{}_{yx_2}$ according to the relationship among dependent and independent variables. Especially, using the result of simulation we will examine when the situation, $R^2{}_{yx_1x_2} > R^2{}_{yx_1} + R^2{}_{yx_2}$, can happen. And using that result we will discuss how it can be extended to the cases where more variables are involved. It will be also discussed about inequality between $R^2{}_{yx_1 \cdots x_k}$ and the sum $r^2{}_{yx_1} + \cdots + r^2{}_{yx_k}$. Notice that for $k=2$, Hamilton(1987) proved that $R^2{}_{yx_1x_2} \leq r^2{}_{yx_1} + r^2{}_{yx_2|x_1}$ where $r_{yx_2|x_1}$ denotes the sample partial correlation coefficient between two variables x_2 and y given x_1. Using partial sum approach, we will extend this result into general k.

2. Regression with Two Predictor Variables

Consider the following identity in the general linear regression model to give the measure of precision for the estimated regression equation :

$$(y_j - \bar{y}) = (y_j - \hat{y_j}) + (\hat{y_j} - \bar{y}) \tag{7}$$

where $\bar{y}$ is the average of $y_j's$ and $\hat{y_j}$ is the least squares estimate of the jth response. If we square both sides of this and sum over $j = 1, 2, \cdots, n$, we obtain

$$\Sigma(y_j - \bar{y})^2 = \Sigma(\hat{y_j} - \bar{y})^2 + \Sigma(y_j - \hat{y_j})^2. \tag{8}$$

We can express (8) in words as follows :

$$\begin{matrix} \text{Total} & & \text{Sum of Squares} & & \text{Sum of Squares} \\ \text{Sum of Squares} & = & \text{due to regression} & + & \text{due to residual errors} \end{matrix}$$

We shall be pleased if the Sum of Squares(SS) due to regression is much greater than the SS due to residual errors. So we can define the R^2 measure to see that how useful the regression line will be as a prediction model. We define

$$R^2 = \frac{SS \;\; due \; to \; regression}{Total \;\; SS}$$

$$= \frac{\sum (\widehat{y_j} - \bar{y})^2}{\sum (y_j - \bar{y})^2} \qquad (9)$$

where both summations are over $j=1,2,\cdots,n$. The quantity R^2 is commonly called the coefficient of determination. If we employ the conditions in (2), R^2 can be written as $R^2 = \sum_j \widehat{y_j}^2$ since $\bar{y}=0$ and $\sum_j y_j^2 = 1$.

For related discussion with R^2 and correlations, see, for instance, Draper and Smith(1981), Montgomery(1982), Kleinbaum and Kupper(1988), Hamilton(1992), Park(1991), Neter, Wasserman and Kutner(1990) and among many references.

Now consider the problem of making a comparison between the sum, $R^2_{yx_1} + R^2_{yx_2}$, and $R^2_{yx_1x_2}$ in terms of correlations. To simply compare them, we can propose the measure Q which is defined as follows :

$$\begin{aligned} Q &= R^2_{yx_1} + R^2_{yx_2} - R^2_{yx_1x_2} \\ &= r^2_{yx_1} + r^2_{yx_2} - \frac{1}{1 - r^2_{x_1x_2}} (r^2_{yx_1} + r^2_{yx_2} - 2r_{x_1x_2} r_{yx_1} r_{yx_2}) \\ &= \frac{-r_{x_1x_2}}{1 - r^2_{x_1x_2}} \{ r_{x_1x_2} (r^2_{yx_1} + r^2_{yx_2}) - 2r_{yx_1} r_{yx_2} \} \qquad (10) \end{aligned}$$

It is obvious that if $r_{x_1x_2}=0$ in (10) that $R^2_{yx_1x_2} = R^2_{yx_1} + R^2_{yx_2}$. In the following we suppress the case that $r_{x_1x_2}$ is ±1 because it is necessary to obtain a non-zero value of the determinant $D=1-r^2{}_{x_1x_2}$.

Proposition 1. Consider the models (3),(4) and (5) with the coefficients of determination at the right-hand side. A sufficient condition for the inequality, $R^2{}_{yx_1}+R^2{}_{yx_2}< R^2{}_{yx_1x_2}$, is $|r_{x_1x_2}| > \dfrac{2|r_{yx_1}r_{yx_2}|}{r^2{}_{yx_1}+r^2{}_{yx_2}}$ provided $r_{x_1x_2}r_{yx_1}r_{yx_2}>0$.

Proof. Using the fact that $1-r^2{}_{x_1x_2}>0$ for all $r_{x_1x_2}$ $(-1<r_{x_1x_2}<1)$, it is enough to consider $Q'=(1-r^2{}_{x_1x_2})Q$ which is as follows :

$$Q'=-r_{x_1x_2}\{r_{x_1x_2}(r^2{}_{yx_1}+r^2{}_{yx_2})-2r_{yx_1}r_{yx_2}\}$$

From the Hamilton's inequality, then, we can easily find the sufficient condition such that $Q'<0$.

$$Q'=-r_{x_1x_2}\{r_{x_1x_2}(r^2{}_{yx_1}+r^2{}_{yx_2})-2r_{yx_1}r_{yx_2}\}<0$$

$$\Leftrightarrow \quad r_{x_1x_2}\left(r_{x_1x_2}-\frac{2r_{yx_1}r_{yx_2}}{r^2{}_{yx_1}+r^2{}_{yx_2}}\right)>0 \quad (\text{ Hamilton's inequality })$$

In most practical cases, the condition $r_{x_1x_2}r_{yx_1}r_{yx_2}>0$ holds, since, if $r_{yx_1}r_{yx_2}>0$, then usually $r_{x_1x_2}>0$, and if $r_{yx_1}r_{yx_2}<0$, then usually $r_{x_1x_2}<0$. Under this condition it often happens that $R^2{}_{yx_1}+R^2{}_{yx_2}<R^2{}_{yx_1x_2}$ if $|r_{x_1x_2}|>\dfrac{2|r_{yx_1}r_{yx_2}|}{r^2{}_{yx_1}+r^2{}_{yx_2}}$.

Proposition 2. A sufficient condition for the inequality, $R^2{}_{yx_1}+R^2{}_{yx_2}<R^2{}_{yx_1x_2}$, is $r_{x_1x_2}r_{yx_1}r_{yx_2}<0$.

Proof. Similarly, as shown in the proof of Proposition 1, consider the sufficient condition such that $Q'<0$.

$$Q' = -r_{x_1x_2}\{r_{x_1x_2}(r^2_{yx_1} + r^2_{yx_2}) - 2r_{yx_1}r_{yx_2}\} < 0$$

$$\Leftrightarrow \quad r_{x_1x_2}\,(r_{x_1x_2} - \frac{2r_{yx_1}r_{yx_2}}{r^2_{yx_1} + r^2_{yx_2}}) > 0 \quad \text{(Hamilton's inequality)}$$

$$\Leftrightarrow \quad r^2_{x_1x_2} > \frac{2r_{x_1x_2}r_{yx_1}r_{yx_2}}{r^2_{yx_1} + r^2_{yx_2}}$$

$$\Leftarrow \quad r_{x_1x_2}r_{yx_1}r_{yx_2} < 0$$

The condition $r_{x_1x_2}r_{yx_1}r_{yx_2} < 0$ as shown above, is the opposite to that of Proposition 1. But this condition almost covers the situation when the inequality, $R^2_{yx_1} + R^2_{yx_2} < R^2_{yx_1x_2}$, can occur. Notice that the union of two sufficient conditions given in Propositions 1 and 2 is equivalent to a necessary and sufficient condition for the inequality, $R^2_{yx_1} + R^2_{yx_2} < R^2_{yx_1x_2}$.

3. Regression with Three and More Independent Variables

In a similar way in Section 2, consider the following models with the coefficients of determination.

Regression Equation	Coefficient of Determination	
$y = a_1x_1 + \varepsilon_1$	$R^2_{yx_1} = r^2_{yx_1}$	(11)
$y = a_2x_2 + \varepsilon_2$	$R^2_{yx_2} = r^2_{yx_2}$	(12)
$y = a_3x_3 + \varepsilon_3$	$R^2_{yx_3} = r^2_{yx_3}$	(13)
$y = \beta_1x_1 + \beta_2x_2$ $+\beta_3x_3 + \varepsilon$	$R^2_{yx_1x_2x_3} = \frac{1}{D}\{r^2_{yx_1}(1 - r^2_{x_2x_3})$ $+ r^2_{yx_2}(1 - r^2_{x_1x_3})$ $+ r^2_{yx_3}(1 - r^2_{x_1x_2})$ $+2r_{yx_1}r_{yx_2}(r_{x_1x_3}r_{x_2x_3} - r_{x_1x_2})$ $+2r_{yx_1}r_{yx_3}(r_{x_1x_2}r_{x_2x_3} - r_{x_1x_3})$ $+2r_{yx_2}r_{yx_3}(r_{x_1x_2}r_{x_1x_3} - r_{x_2x_3})\}$	(14)

where $D = 1 + 2r_{x_1x_2}r_{x_1x_3}r_{x_2x_3} - r^2_{x_1x_2} - r^2_{x_1x_3} - r^2_{x_2x_3}$.The results of R^2 in these equations are obtained in Appendix A.

Consider a sufficient conditio for the inequality, $R^2_{yx_1} + R^2_{yx_2} + R^2_{yx_3} < R^2_{yx_1x_2x_3}$, in terms of correlations. We can propose the measure Q which is similarly defined as shown in Section 2. That is,

$$\begin{aligned} Q &= R^2_{yx_1} + R^2_{yx_2} + R^2_{yx_3} - R^2_{yx_1x_2x_3} \\ &= \frac{1}{D}\{2r_{x_1x_2}r_{x_1x_3}r_{x_2x_3}(r^2_{yx_1} + r^2_{yx_2} + r^2_{yx_3}) \\ &\quad -(r^2_{x_1x_3} + r^2_{x_2x_3} + r^2_{x_1x_2})(r^2_{yx_1} + r^2_{yx_2} + r^2_{yx_3}) \\ &\quad +(r^2_{yx_1} \ r^2_{x_2x_3} + r^2_{yx_2})(r^2_{x_1x_3} + r^2_{yx_3} \ r^2_{x_1x_2}) \\ &\quad -2r_{yx_1}r_{yx_2}(r_{x_1x_3} r_{x_2x_3} - r_{x_1x_2}) - 2r_{yx_1}r_{yx_3}(r_{x_1x_2} r_{x_2x_3} - r_{x_1x_3}) \\ &\quad -2r_{yx_2}r_{yx_3}(r_{x_1x_2} r_{x_1x_3} - r_{x_2x_3}))\} \end{aligned} \tag{15}$$

Note that the determinant D becomes larger than zero except that the $X'X$ matrix is singular.

Proposition 3. Consider the models (11)-(14). If the conditions, $r_{x_1x_2}r_{yx_1}r_{yx_2} < 0$, $r_{x_1x_3}r_{yx_1}r_{yx_3} < 0$ and $r_{x_2x_3}r_{yx_2}r_{yx_3} < 0$, are satisfied, then we obtain $R^2_{yx_1} + R^2_{yx_2} + R^2_{yx_3} < R^2_{yx_1x_2x_3}$.

The proof of the above proposition is given in Appendix B.

In case that more variables are involved, considering Propositions 1,2 and 3, we can assume that we have $R^2_{yx_1} + R^2_{yx_2} + \cdots + R^2_{yx_k} < R^2_{yx_1x_2\cdots x_k}$ under the conditions $r_{x_ix_j}r_{yx_i}r_{yx_j} < 0$, or $|r_{x_ix_j}| > \frac{2|r_{yx_i}r_{yx_j}|}{r^2_{yx_i} + r^2_{yx_j}}$ provided $r_{x_ix_j}r_{yx_i}r_{yx_j} > 0$, for all $i \neq j$, $i, j = 1, 2, \cdots, k$. However, these conditions are so conservative that we need further study using geometric approach or anything else.

4. Examples

In this section, we will introduce examples of applying Propositions 1,2, and 3. And through the simulation of correlations we will also show that under what conditions we can encounter $R^2_{yx_1x_2} > R^2_{yx_1} + R^2_{yx_2}$.

4.1 Example for Proposition 1

1) A simulated set of data is given below :

No.	x_1	x_2	y
1	178	182	44
2	185	185	40
3	156	168	44
4	166	172	42
5	178	180	38
6	176	176	47
7	176	180	40
8	162	170	43
9	174	176	44
10	170	186	38
11	168	168	44
12	186	192	45
13	176	176	45
14	162	164	47
15	166	170	54
16	180	185	49
17	168	172	51
18	162	168	51
19	162	164	48
20	168	168	49
21	174	176	57
22	156	165	54
23	164	166	52
24	146	155	50
25	172	172	51
26	168	172	54
27	186	188	51
28	148	155	57
29	186	188	49
30	170	176	48
31	170	172	52

2) Correlation Matrix

	y	x_1	x_2
y	1	-.23674	-.39797
x_1	-.23674	1	.92975
x_2	-.39797	.92975	1

$\Rightarrow r_{yx_1}=-.23674,\ r_{yx_2}=-.39797,$

$r_{x_1x_2}=.92975$

$$\frac{2|r_{yx_1}r_{yx_2}|}{r^2_{yx_1}+r^2_{yx_2}}=.8792$$

3) For this example, we have the values of R^2 as follows :

Regression Equation	Coefficient of Determination
$y=\alpha_1x_1+\varepsilon_1$	$R^2_{yx_1}=.0560$
$y=\alpha_2x_2+\varepsilon_2$	$R^2_{yx_2}=.1584$
	sum : .2144
$y=\beta_1x_1+\beta_2x_2+\varepsilon$	$R^2_{yx_1x_2}=.2894$

From the examination of the correlation matrix it reveals that the conditions, $r_{x_1x_2}r_{yx_1}r_{yx_2}>0$ & $|r_{x_1x_2}| > \frac{2|r_{yx_1}r_{yx_2}|}{r^2_{yx_1}+r^2_{yx_2}}$ are satisfied. Under this situation we confirm that $R^2_{yx_1x_2}$ is larger than the sum, $R^2_{yx_1} + R^2_{yx_2}$. For several other cases of Proposition 1, see Table 1.

Table 1. Some cases of Proposition 1

$r_{x_1x_2}$	r_{yx_1}	r_{yx_2}	$R^2_{yx_1x_2}$	$R^2_{yx_1} + R^2_{yx_2}$	*diff 1*	*diff2*
					$(= R^2_{yx_1} + R^2_{yx_2} - R^2_{yx_1x_2})$	$(= \lvert r_{x_1x_2} \rvert - \frac{2\lvert r_{yx_1} r_{yx_2} \rvert}{r^2_{yx_1} + r^2_{yx_2}})$
0.9	0.1	0.2	.074	.05	-.024	.100
0.9	0.1	0.4	.516	.17	-.346	.429
0.7	0.1	0.4	.224	.17	-.054	.229
0.7	0.1	0.6	.561	.37	-.191	.376
0.5	0.1	0.4	.173	.17	-.003	.029
0.5	0.1	0.6	.413	.37	-.043	.176
0.5	0.1	0.8	.760	.65	-.110	.254
0.3	0.1	0.8	.662	.65	-.012	.054

4.2 Example for Proposition 2

1) A simulated set of data is given below :

No.	x_1	x_2	y
1	44	44	182
2	45	40	185
3	54	44	168
4	59	42	172
5	49	38	180
6	44	47	176
7	45	40	180
8	49	43	170
9	39	44	176
10	60	38	186

2) Correlation Matrix

	y	x_1	x_2
y	1	-.12804	-.56287
x_1	-.12804	1	-.39972
x_2	-.56287	-.39972	1

$\Rightarrow r_{yx_1}$=-.12804, r_{yx_2}=-.56287,
$r_{x_1x_2}$=-.39972

3) For this example, we have the values of R^2 as follows :

Regression Equation	Coefficient of Determination
$y = \alpha_1 x_1 + \varepsilon_1$	$R^2_{yx_1} = .0164$
$y = \alpha_2 x_2 + \varepsilon_2$	$R^2_{yx_2} = .3168$
	sum : .3332
$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$	$R^2_{yx_1x_2} = .4652$

From the examination of the correlation matrix it reveals that the condition, $r_{x_1x_2} r_{yx_1} r_{yx_2} < 0$, is satisfied. Under this condition we confirm that $R^2_{yx_1x_2}$ is larger than the sum, $R^2_{yx_1} + R^2_{yx_2}$. For several other cases of Proposition 2, see Table 2.

Table 2. Some cases of Proposition 2

$r_{x_1x_2}$	r_{yx_1}	r_{yx_2}	$R^2_{yx_1x_2}$	$R^2_{yx_1} + R^2_{yx_2}$	*Diff 1* $(= R^2_{yx_1} + R^2_{yx_2} - R^2_{yx_1x_2})$
0.9	0.1	-0.2	.453	.05	-.403
0.7	0.1	-0.4	.443	.17	-.273
0.7	0.1	-0.6	.890	.37	-.520
0.5	0.1	-0.4	.280	.17	-.110
0.5	0.1	-0.6	.573	.37	-.203
0.5	0.1	-0.8	.973	.65	-.323
0.3	0.1	-0.4	.213	.17	-.043
0.3	0.1	-0.6	.446	.37	-.076
0.3	0.1	-0.8	.767	.65	-.117

4.3 Example for Proposition 3

1) A simulated set of data is given below :

No	x_1	x_2	x_3	y
1	44	44	98	182
2	45	40	57	185
3	54	44	58	168
4	59	42	86	172
5	49	38	98	180
6	44	47	77	176
7	45	40	57	180

2) Correlation Matrix

	y	x_1	x_2	x_3
y	1	-.75367	-.43359	.11742
x_1	-.75367	1	-.07926	.07061
x_2	-.43359	-.07926	1	.00330
x_3	.11742	.07061	.00330	1

$\Rightarrow$ $r_{yx_1}=-.75367$, $r_{yx_2}=-.43359$,
$r_{yx_3}=.11742$, $r_{x_1x_2}=-.07926$,
$r_{x_1x_3}=.07061$, $r_{x_2x_3}=.00330$

3) For this example, we have the values of R^2 as follows :

Regression Equation	Coefficient of Determination
$y=\alpha_1x_1+\varepsilon_1$	$R^2{}_{yx_1}=.5680$
$y=\alpha_2x_2+\varepsilon_2$	$R^2{}_{yx_2}=.1880$
$y=\alpha_3x_3+\varepsilon_3$	$R^2{}_{yx_3}=.0138$
	sum : .7698
$y=\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon$	$R^2{}_{yx_1x_2x_3}=.8437$

It is clear from the correlation matrix that the conditions of Proposition 3 are satisfied. That is, $r_{x_1x_2}r_{yx_1}r_{yx_2}<0$, $r_{x_1x_3}r_{yx_1}r_{yx_3}<0$ and $r_{x_2x_3}r_{yx_2}r_{yx_3}<0$. Under these conditions we confirm that $R^2{}_{yx_1x_2x_3}$ is larger than the sum, $R^2{}_{yx_1}+R^2{}_{yx_2}+R^2{}_{yx_3}$.

5. Remarks

Ott(1993) stated general linear model in his textbook as follows: " When the independent variables are themselves correlated, it is difficult to separate R^2 into the independent contribution of each independent variable. For these situations, where the independent variables account for overlapping pieces of the variability in the y-values, we often find that

$$R^2{}_{yx_1x_2\cdots x_k} < r_{yx_1}{}^2 + r_{yx_1}{}^2 + \cdots + r_{yx_k}{}^2. \qquad (16)$$

Hamilton(1987) stated that statement (16) (Hamilton quoted this statement in Ott(1984)) reflected erroneous belief in such a way that correlated explanatory variables contain only redundant information about y. Thus he corrected (16) for k=2, $R^2{}_{yx_1x_2} \le r_{yx_1}{}^2 + r_{yx_2|x_1}{}^2$, as presented in the introduction. Similarly we will correct (16) for general k. It is equal to find the upper bound for $R^2{}_{yx_1x_2\cdots x_k}$. Consider first the statement (16) for k=3. The sum of squares for regression, $SSR(x_1 ,x_2 ,x_3)$ can be partitioned as

$$SSR(x_1 ,x_2 ,x_3) \;=\; SSR(x_1|\; x_2 ,x_3) \;+\; SSR(x_2|\; x_1) \;+\; SSR(x_1), \qquad (17)$$

where $SSR(x_1)$ is the Sum of Squares (SS) due to regressing y on x_1, $SSR(x_2|\; x_1)$ is the extra SS due to adding x_2 to a model with x_1, and $SSR(x_1|\; x_2 ,x_3)$ is the extra SS due to adding x_3 to a model with x_1 and x_2 . If we express each SS in statement (17) in terms of sample correlation coefficients and sample partial correlation coefficients, then we can represent it as follows :

$$SSR(x_1)= r^2{}_{yx_1}\, SST$$

$$SSR(x_2|\; x_1) \;=\; r^2{}_{yx_2|x_1}(1- r^2{}_{yx_1})\; SST$$

$$SSR(x_1 ,x_2 ,x_3) \;=\; r^2{}_{yx_1|x_2x_3}(1- r^2{}_{yx_2|x_1})(1- r^2{}_{yx_1})\; SST$$

where $r_{yx_3|x_1x_2}$ measures the contribution of adding the first order term x_3 to the model after the effects of the first order terms x_1, x_2 are controlled for. Kleinbaum and Kupper(1978) called this "multiple-partial correlation coefficient". Relating $R^2{}_{yx_1x_2x_3}$, we can obtain

$$R^2{}_{yx_1x_2x_3} \;=\; r^2{}_{yx_1} + r^2{}_{yx_2|x_1}(1- r^2{}_{yx_1}) + r^2{}_{yx_3|x_1x_2}(1- r^2{}_{yx_2|x_1})(1- r^2{}_{yx_1})$$

Therefore, a correct version of (16) for $k=3$ is $R^2{}_{yx_1x_2x_3} \le r^2{}_{yx_1} + r^2{}_{yx_2|x_1} + r^2{}_{yx_3|x_1x_2}$. If it is repeatedly calculated in the same way, we can obtain the following for general k :

$$\begin{aligned} R^2{}_{yx_1x_2\cdots x_k} \;=\; & r^2{}_{yx_1} + r^2{}_{yx_2|x_1}(1- r^2{}_{yx_1}) + r^2{}_{yx_3|x_1x_2}(1- r^2{}_{yx_2|x_1})(1- r^2{}_{yx_1}) \\ & + r^2{}_{yx_4|x_1x_2x_3}(1- r^2{}_{yx_3|x_1x_2})(1- r^2{}_{yx_2|x_1})(1- r^2{}_{yx_1}) \\ & + \;\cdots\; + \\ & + r^2{}_{yx_k|x_1\cdots x_{k-1}}(1- r^2{}_{yx_{k-1}|x_1\cdots x_{k-2}})(1- r^2{}_{yx_{k-2}|x_1\cdots x_{k-3}}) \cdots (1- r^2{}_{yx_1}) \end{aligned}$$

Thus, when the independent variables are correlated, namely, there is multicollinearity in multiple regression, we can state that

$$R^2{}_{yx_1x_2\cdots x_k} \leq r^2{}_{yx_1} + r^2{}_{yx_2|x_1} + r^2{}_{yx_3|x_1x_2} + \cdots + r^2{}_{yx_k|x_1\cdots x_{k-1}}.$$

Appendix A

Note that with the underlying variables R^2 is expressed in matrix forms as follows :

(1) $y = a_1x_1 + \varepsilon_1$

$$\Rightarrow R^2{}_{yx_1} = \underline{y}'X\,(X'X)^{-1}X'\underline{y}$$
$$= r^2{}_{yx_1}$$

where $\underline{y} = (y_1, y_2, \cdots, y_n)$

$$X = \begin{bmatrix} x_{11}, & x_{21}, \cdots, & x_{k1} \\ x_{12}, & x_{22}, \cdots, & x_{k2} \\ x_{13}, & x_{23}, \cdots, & x_{k3} \\ \vdots & \vdots \quad \vdots & \vdots \\ x_{1n}, & x_{2n}, \cdots, & x_{kn} \end{bmatrix}$$

(2) $y = a_2x_2 + \varepsilon_2$

$$\Rightarrow R^2{}_{yx_2} = r^2{}_{yx_2}$$

(3) $y = \beta_1x_1 + \beta_2x_2 + \varepsilon$

$$\Rightarrow R^2{}_{yx_1x_2} = (r_{yx_1} \;\; r_{yx_2}) \begin{pmatrix} 1 & r_{x_1x_2} \\ r_{x_1x_2} & 1 \end{pmatrix}^{-1} \begin{pmatrix} r_{yx_1} \\ r_{yx_2} \end{pmatrix}$$
$$= \frac{1}{1 - r_{x_1x_2}{}^2}\{r_{yx_1}(r_{yx_1} - r_{x_1x_2}r_{yx_2}) + r_{yx_2}(r_{yx_2} - r_{x_1x_2}r_{yx_1})\}$$

(4) $y = a_3x_3 + \varepsilon_3$

$$\Rightarrow R^2{}_{yx_3} = r^2{}_{yx_3}$$

(5) $y = \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \varepsilon$

$$\Rightarrow R^2{}_{yx_1x_2x_3} = (r_{yx_1} \;\; r_{yx_2} \;\; r_{yx_3}) \begin{pmatrix} 1 & r_{x_1x_2} & r_{x_1x_3} \\ r_{x_1x_2} & 1 & r_{x_2x_3} \\ r_{x_1x_3} & r_{x_2x_3} & 1 \end{pmatrix}^{-1} \begin{pmatrix} r_{yx_1} \\ r_{yx_2} \\ r_{yx_3} \end{pmatrix}$$

$$\begin{aligned} = \frac{1}{D}\{ \; & r^2{}_{yx_1}(1- r^2{}_{x_2x_3})+ r^2{}_{yx_2}(1- r^2{}_{x_1x_3}) \\ & + r^2{}_{yx_3}(1- r^2{}_{x_1x_2})+2r_{yx_1}r_{yx_2}(r_{x_1x_3}r_{x_2x_3} - r_{x_1x_2}) \\ & +2r_{yx_1}r_{yx_3}(r_{x_1x_2}r_{x_2x_3} - r_{x_1x_3}) \\ & +2r_{yx_2}r_{yx_3}(r_{x_1x_2}r_{x_1x_3} - r_{x_2x_3}) \; \} \end{aligned}$$

where $D = 1+2r_{x_1x_2}r_{x_1x_3}r_{x_2x_3} - r^2{}_{x_1x_2} - r^2{}_{x_1x_3} - r^2{}_{x_2x_3}$.

Appendix B

Proof of Proposition 3.

From (13) we only want to check the sign of the measure Q. So using the fact that if $r_{x_1x_2}$, $r_{x_1x_3}$ and $r_{x_2x_3}$ are near to zero then $D = 1 + 2r_{x_1x_2}r_{x_1x_3}r_{x_2x_3} - r^2_{x_1x_2} - r^2_{x_1x_3} - r^2_{x_2x_3} > 0$.

We can rewrite Q as follows:

$$Q = - r_{x_1x_2}{}^2(r_{yx_1}{}^2 + r_{yx_2}{}^2) - r_{x_2x_3}{}^2(r_{yx_2}{}^2 + r_{yx_3}{}^2) - r_{x_1x_3}{}^2(r_{yx_1}{}^2 + r_{yx_3}{}^2) \tag{①}$$

$$+2\,(r_{x_1x_2}r_{yx_1}r_{yx_2} + r_{x_1x_3}r_{yx_1}r_{yx_3} + r_{x_2x_3}r_{yx_2}r_{yx_3}) \tag{②}$$

$$+2r_{yx_1}r_{x_1x_2}(r_{x_1x_3}r_{x_2x_3}r_{yx_1} - r_{yx_3}r_{x_2x_3}) \tag{③}$$

$$+2r_{yx_2}r_{x_2x_3}(r_{x_1x_2}r_{x_1x_3}r_{yx_2} - r_{yx_1}r_{x_1x_3}) \tag{④}$$

$$-2r_{yx_3}r_{x_1x_3}(r_{x_1x_2}r_{x_2x_3}r_{yx_3} - r_{yx_2}r_{x_1x_2}) \tag{⑤}$$

Here, suppose that $$\begin{cases} r_{x_1x_2}r_{yx_1}r_{yx_2} < 0 \\ r_{x_1x_3}r_{yx_1}r_{yx_3} < 0 \\ r_{x_2x_3}r_{yx_2}r_{yx_3} < 0 \; . \end{cases}$$

Then in the above we can obtain the minus sign of Q by showing that all the terms ① - ⑤ will be less than zero, respectively, under the conditions in Proposition 5.

(i) Proof of ① and ②

It is trivial that the equations, ① and ②,are less than zero under the conditions.

(ii) Proof of ③

Consider all possible signs of correlations satisfying the conditions, and let

$$s = r_{yx_1} r_{x_1x_2}$$
$$t = r_{x_1x_3} r_{x_2x_3} r_{yx_1} - r_{yx_3} r_{x_2x_3} \ .$$

We then prove that the equation ③ is less than zero by showing that the sign of multiplication s by t is minus. That is ,we can obtain the following:

r_{yx_1}	$r_{x_1x_2}$	r_{yx_2}	r_{yx_3}	$r_{x_1x_3}$	$r_{x_2x_3}$	$r_{x_1x_3} r_{x_2x_3} r_{yx_1}$	$-r_{yx_3} r_{x_2x_3}$	s	t	③$(=s\times t)$
+	−	+	+	−	−	+	+	−	+	−
+	−	+	−	+	+	+	+	−	+	−
−	+	+	+	+	−	+	+	−	+	−
−	+	+	−	−	+	+	+	−	+	−
+	+	−	+	−	+	−	−	+	−	−
+	+	−	+	−	+	−	−	+	−	−
−	−	−	+	+	+	−	−	+	−	−
−	−	−	−	−	+	−	−	+	−	−

This result shows that the equation ③ is less than zero under the conditions.

(iii) Proof of ④

It can be easily proved in the same way as shown in (ii). Let

$$s = r_{yx_2} r_{x_2x_3}$$
$$t = r_{x_1x_2} r_{x_1x_3} r_{yx_2} - r_{yx_1} r_{x_1x_3}.$$

Then we have the following :

r_{yx_2}	$r_{x_2x_3}$	r_{yx_3}	r_{yx_1}	$r_{x_1x_3}$	$r_{x_1x_2}$	$r_{x_1x_2} r_{x_1x_3} r_{yx_2}$	$-r_{yx_1} r_{x_1x_3}$	s	t	④$(=s\times t)$
+	−	+	+	−	−	+	+	−	+	−
+	−	+	−	+	+	+	+	−	+	−
−	+	+	+	+	−	+	+	−	+	−
−	+	+	−	−	+	+	+	−	+	−
+	+	−	+	−	+	−	−	+	−	−
+	+	−	+	−	+	−	−	+	−	−
−	−	−	+	+	+	−	−	+	−	−
−	−	−	−	−	+	−	−	+	−	−

(iv) Proof of ⑤

It can be easily proved in the similar way as shown in (ii) and (iii).

Let $s = r_{yx_3} r_{x_1x_3}$

$$t = r_{x_1x_2} r_{x_2x_3} r_{yx_3} - r_{yx_2} r_{x_1x_2}.$$

Then we have the following :

r_{yx_3}	$r_{x_1x_3}$	r_{yx_1}	$r_{x_1x_2}$	$r_{x_2x_3}$	r_{yx_2}	$r_{x_1x_2}r_{x_2x_3}r_{yx_3}$	$-r_{yx_2}r_{x_1x_2}$	s	t	⑤$(=s\times t)$
+	−	+	+	+	−	+	+	−	+	−
+	−	+	−	−	+	+	+	−	+	−
−	+	+	+	−	−	+	+	−	+	−
−	+	+	−	+	+	+	+	−	+	−
+	+	−	+	−	+	−	−	+	−	−
+	+	−	−	−	+	−	−	+	−	−
−	−	−	+	+	+	−	−	+	−	−
−	−	−	−	−	−	−	−	+	−	−

References

[1] Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis,* 2nd ed., John Wiley & Sons, Inc., New York.

[2] Hamilton, L.C. (1992). *Regression with Graphics,* Brooks/Cole Publishing Company Pacific Grove, California.

[3] Hamilton, D.(1987). Sometimes $R^2 > r^2_{yx_1} + r^2_{yx_2}$: Correlated Variables Are Not Always Redundant, *The American Statistician,* 41.

[4] Kleinbaum, D.G. and Kupper, L.L. (1988). *Applied Regression Analysis and Other Multivariate Methods,* 2nd ed., PWS-Kent Publishing Co., Boston.

[5] Montgomery, D.C. and Peck, E.A. (1982). *Introduction to Linear Regression Analysis,* John Wiley & Sons, Inc., New York.

[6] Neter, J., Wasserman, W., and Kutner, M.H. (1990). *Applied Linear Statistical Models,* 3rd ed., Irwin, Homewood, Il.

[7] Park, S.H. (1991). *Regression Analysis,* Dae Young Sa.

[8] Ott, L. (1984). *An Introduction to Statistical Methods and Data Analysis,* 2nd ed., Duxbury Press, Boston.

[9] Ott, L. (1993). *An Introduction to Statistical Methods and Data Analysis,* 4th ed., Duxbury Press, Boston.