ASSESSING UNCERTAINTY OF PREDICTION WHEN ADDITIONAL INDEPENDENT VARIABLES ARE CONSIDERED FOR REGRESSION

Eugene Dutoit, US Army Infantry Center and Columbus College
Douglas Penfield, Rutgers University

I. Introduction.

The purpose of this paper is to discuss another rule for determining whether the addition of another independent variable into a linear multiple regression equation is worth the effort required to obtain data on the variable. In addition, the paper attempts to give a simple explanation of why the standard error of estimate about the regression surface may increase when additional variables are entered.

The paper is divided into three parts. Part I is the introduction, Part II examines the algebraic relationships and Part III discusses applications and gives an example using data.

II. Development.

The form of the linear multiple regression equation considered in this study is:

$$y = b_o + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k \quad (1)$$

The common measure of "goodness of fit" is the coefficient of determination (R^2_k) which is computed as the ratio of explained variance (by regression) to the total variation. Tables have been constructed and published that show what the critical value of R_k would have to be if "k" variables were included in the multiple regression equation. Consider the relationship for computing the standard error of estimate:

$$S^2_{y\downarrow X_k} = \frac{n-1}{n-k-1}(S^2_y)(1-R^2_k) \quad (2)$$

If the analyst has the option to include k, k+1, k+2, . . . number of independent variables in his prediction equation he may want some measure of the change in uncertainty of prediction about the regression. One measure of uncertainty about prediction is the standard error of estimate ($S_{y\downarrow X_k}$) - (Equation 2). Consider the case where the number of observations is fixed. The standard errors of estimate for k and (k+1) independent variables would be $S_{y\downarrow X_k}$ and $S_{y\downarrow X_{k+1}}$ respectively.

It is a common conception that the inclusion of additional variables in the regression equation always tends to reduce the error of estimate somewhat and leads to an increase in R. From the above statement it can be inferred that:

$$S_{y\downarrow X_k} > S_{y\downarrow X_{k+1}} \quad (3)$$

The following ratio can be computed:

$$\frac{S_{y\downarrow X_k}}{S_{y\downarrow X_{k+1}}} \quad (4)$$

This may be interpreted as a ratio of the standard error of estimates computed for (k) and (k+1) independent variables. The importance of this ratio is developed below and illustrated in Figure 1.

Inequality (3) is based on the common conception stated above. However, the underlined statement is not true for all combinations of values for n, k, R^2_k and R^2_{k+1}. The following discussion explores the total relationship in more depth. Consider equation (2) and its extension denoted by (2A).

$$S^2_{y\downarrow X_{k+1}} = \left(\frac{n-1}{n-(k+1)-1}\right)(S^2_y)(1-R^2_{k+1}) \quad (2A)$$

Equation (2A) can be written as:

$$S^2_{y\downarrow X_{k+1}} = \left(\frac{n-1}{n-k-2}\right)(S^2_y)(1-R^2_{k+1}) \quad (2A)$$

When one more independent variable is added to the regression function it follows that:

$$\frac{n-1}{n-k-2} > \frac{n-1}{n-k-1} \quad (5)$$

It is also true that:

$$R^2_{k+1} > R^2_k \quad (6)$$

for an additional independent variable. From equation (6) it then follows that:

$$(1-R^2_k) > (1-R^2_{k+1}) \quad (7)$$

Both (n-1) and (S^2_y) remain the same in equations (2) and (2A). Ratio (4) may be expanded through equations (2) and (2A).

$$\frac{S^2_{y\downarrow X_k}}{S^2_{y\downarrow X_{k+1}}} = \left(\frac{n-k-2}{n-k-1}\right)\cdot\left(\frac{1-R^2_k}{1-R^2_{k+1}}\right) \quad (8)$$

Where:

$$\left.\begin{array}{l}\dfrac{(n-k-2)}{(n-k-1)} < 1.0 \\ \text{and} \\ \dfrac{(1-R^2_k)}{(1-R^2_{k+1})} > 1.0\end{array}\right\} \quad (9)$$

Therefore, the value of the ratio $(S^2_{y\downarrow X_k} / S^2_{y\downarrow Xk+1})$, as given in (8), will depend on how much one factor is less than one in relation to how much the other factor is greater than one--i.e., equations (9).

Equation (8) can be plotted as shown in the following figure.

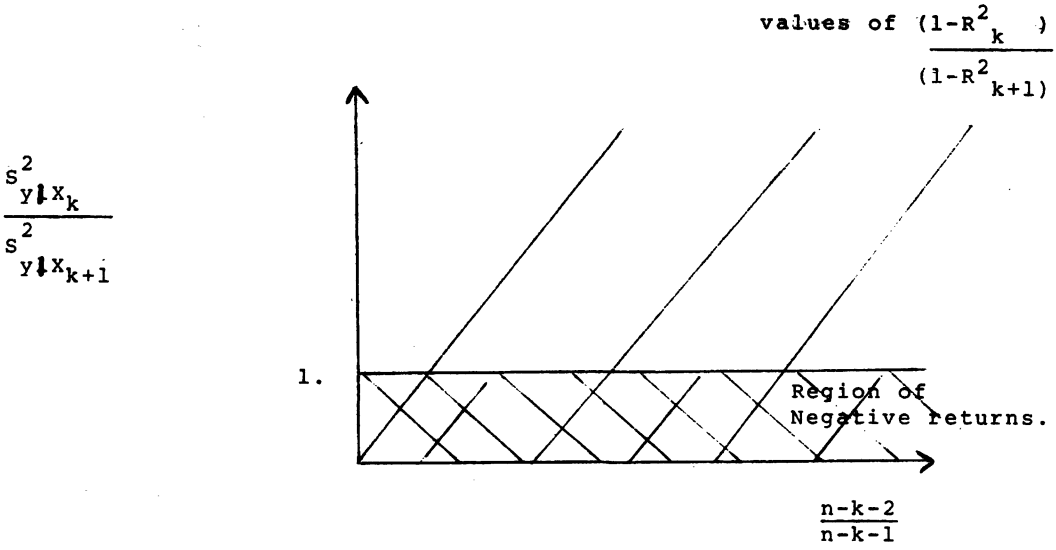


Figure 1

Values characteristic of Figure (1) may be used by the social science researcher to determine if the relationship expressed on inequality (3) is true for his particular experimental situation. If relationship (3) is true; then, the ratio expressed as the ordinate ($S^2_{y\downarrow x_k} / S^2_{y\downarrow x_{k+1}}$) will be greater than one. If relationship (3) is not true, the ratio will be less than one and will fall in the shaded or region of negative returns (i.e., the value of $S^2_{y\downarrow x_k} < S^2_{y\downarrow x_{k+1}}$).

If inequality (3) is expected to be true, then these relationships can be modified to give more information to the analyst. The percentage change (PC) in the standard error of estimate is:

$$PC = \left[\frac{S_{y\downarrow x_k} - S_{y\downarrow x_{k+1}}}{S_{y\downarrow x_k}}\right] * 100 \qquad (10)$$

Equation (10) can be interpreted as some estimate percentage decrease in $S_{y\downarrow x_k}$ which results by adding one more additional independent variable. Substituting equations (2) into equation (10):

$$PC = \frac{\left[\frac{n-1}{n-k-1} \cdot (1-R^2_k)\right]^{\frac{1}{2}} - \left[\frac{n-1}{n-k-2} \cdot (1-R^2_{k+1})\right]^{\frac{1}{2}}}{\left[\frac{n-1}{n-k-1} \quad (1-R^2_k)\right]^{\frac{1}{2}}} \times 100$$

which is equivalent to:

$$PC = \left[1 - \sqrt{\frac{\frac{n-1}{n-k-2} \cdot (1-R^2_{k+1})}{\frac{n-1}{n-k-1} \cdot (1-R^2_k)}}\right] \times 100$$

or finally

$$PC = \left[1 - \sqrt{\frac{n-k-1}{n-k-2} \cdot \frac{(1-R^2_{k+})}{(1-R^2_k)}}\right] * 100 \qquad (11)$$

Equation (11) can be displayed in the same fashion as equation (8). The diagram is shown below:

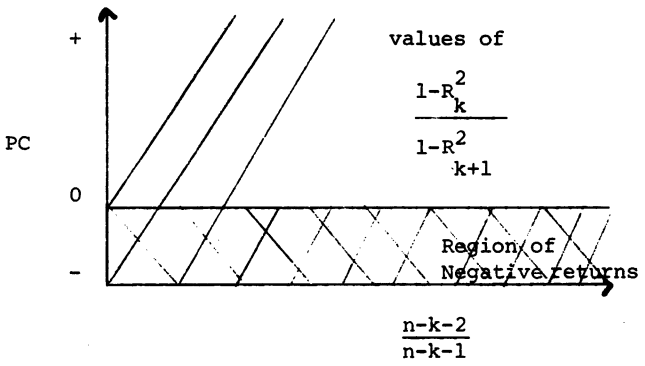


Figure (2)

Figure (2) is computed in table form for selected values of $[(n-k-2)/(n-k-1), R^2_k, R^2_{k+1}]$ - see Table 1. It should be noted that the negative percent reduction indicates that the standard error of estimate actually increases when some additional independent variables are included in the regression equation.

Table 1. Expected Percent Reduction In The Standard Error Of Estimate

$(1-R^2_k)/(1-R^2_{k+1})$

$\frac{n-k-2}{n-k-1}$	1.00	1.02	1.05	1.10	1.15	1.20
.90	-5.41	-4.37	-2.87	- .50	1.71	3.78
.92	-4.26	-3.23	-1.75	.60	2.78	4.83
.94	-3.14	-2.13	- .66	1.66	3.82	5.85
.96	-2.06	-1.06	.40	2.69	4.83	6.83
.98	-1.02	- .02	1.42	3.69	5.80	7.79
1.00	0.00	.99	2.41	4.65	6.75	8.71

III. Application.

A methodology for applying these results is outlined below:

1. For a given set of data of n observations on k independent variables the value of R_k^2 is known.

2. Perhaps the value of R_{k+1}^2 (for an additional independent variable) may be estimated within the context of the experimental situation. In some circumstances the researcher may have some idea of his estimates of R_{k+1}^2 which are based on past experience in observing the relationships between the variables. In any case, the researcher would probably want to consider a range of values (estimates) for R_{k+1}^2.

3. The values of (R_{k+1}^2) are applied in equation (11). If the estimated values of PC do not fall in the region of negative returns (i.e., the region of negative returns = PC <0), then the standard error of estimate would probably be reduced if the (k+1) independent variable were included in the regression. Of course, a value of PC >0 is no firm reason for selecting additional independent variables for regression. The incremental differences $D=(R_{k+1}^2 - R_k^2)$ should also be statistically significant for some value if the type I error equal to α.

Numerical Example:

The following example is taken from Johnson (1950). It consists of a random sample of 50 students taken from a study dealing with the prediction of achievement of freshmen at the University of Minnesota. The variables that were selected were:

Y = honor-point ratio of the end of the freshman year.

X_1 = score on the Johnson Science Application Test.

X_2 = score on an English Test.

X_3 = score on the Cooperative Algebra Test.

X_4 = percentile rank in high school graduation class transformed to probits ($\mu=5, \sigma=1$).

The total set of data for all 50 students can be obtained from Johnson (1950). The following lines summarize the output results of these data by using the program BMD02R.

Step 1
Variable entered = X_4
n = 50
R^2 = .2188
St'd error of estimate = .4848

Step 2
Variable entered = X_1
n = 50
R^2 = .2638
St'd error of estimate = .4757

Step 3
Variable entered = X_3
n = 50
R^2 = .2656
St'd error of estimate = .4802

Step 4
Variable entered = X_2
n = 50
R^2 = .2664
St'd error of estimate = .4853

Equation (11) is verified by example by referring to the four steps of the stepwise regression exercise shown above. Equation (11) is repeated below:

$$PC = \left[1 - \sqrt{\frac{n-k-1}{n-k-2} \cdot \frac{(1-R_{k+1}^2)}{(1-R_k^2)}} \right] X\ 100 \quad . \quad . \quad . \qquad (11)$$

1. Step 1 to step 2

$\frac{1}{}$	$\frac{2}{}$
$n = 50$	$n = 50$
$k = 1$	$k+1 = 2$
$R_k^2 = .2188$	$R_{k+1}^2 = .2638$
$S_{y\cdot X_1} = .4848$	$S_{y\cdot X_2} = .4757$

Actual Change in $S_{y\cdot X}$:

$$\left(\frac{.4848 - .4757}{.4848} \right) \cdot 100 = 1.9\%$$

Using Eq (11)

$$\left[1 - \sqrt{\frac{(48)}{(47)} \frac{(.7362)}{(.7812)}} \right] \cdot 100 = 1.9\%$$

2. Step 2 to step 3

$\frac{2}{}$	$\frac{3}{}$
$n = 50$	$n = 50$
$k = 2$	$k+1 = 3$
$R_k^2 = .2638$	$R_{k+1}^2 = .2656$
$S_{y\cdot X_2} = .4757$	$S_{y\cdot X_3} = .4802$

Actual Change in $S_{y\cdot X}$:

$$\left(\frac{.4757 - .4802}{.4757} \right) X\ 100 = -1\%$$

Using Eq (11)

$$\left[1 - \sqrt{\frac{(47)}{(46)} \frac{(.7344)}{(.7362)}} \right] \quad 100 = -1\%$$

3. Step 3 to step 4

$\frac{3}{}$	$\frac{4}{}$
$n = 50$	$n = 50$
$k = 3$	$k+1 = 4$
$R_k^2 = .2656$	$R_{k+1}^2 = .2664$
$S_{y\cdot X_3} = .4802$	$S_{y\cdot X_4} = .4853$

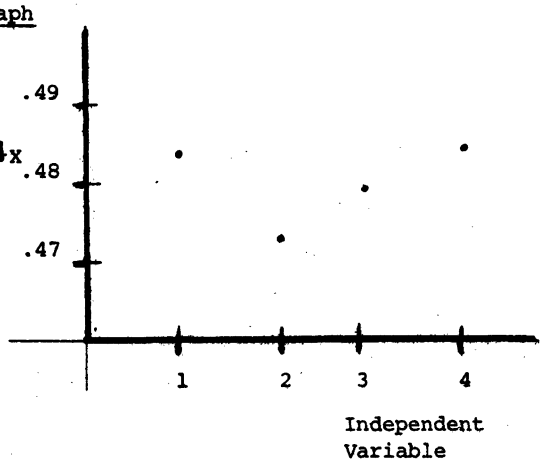
Actual Change in $S_{y\cdot X}$:

$$\left(\frac{.4802 - .4853}{.4802} \right) X\ 100 = -1.1\%$$

Using Eq (11)

$$\left[1 - \sqrt{\frac{(46)}{(45)}\frac{(.7336)}{(.7344)}}\right] \cdot 100 = -1.1\%$$

Graph



References:

BMD, Biomedical Computer Programs, Berkeley, California: University of California Press, 1970

Crow, E. and Davis, F. and Maxfield, M. Statistics Manual. New York: Dover Publications, 1960

Johnson, P. Statistical Methods in Research. New York: Pentice-Hall, Inc., 1950