TABLES FOR DETERMINING THE MINIMUM INCREMENTAL SIGNIFICANCE OF THE MULTIPLE CORRELATION COEFFICIENT

Eugene F. Dutoit, US Army Infantry School
Douglas Penfield, Rutgers, The State University

STATEMENT OF THE PROBLEM AND BACKGROUND

The purpose of this paper is to develop a set of tables to allow the researcher to assess the significance of the increase of R (multiple correlation coefficient) for a linear regression function when additional independent variables are to be considered and collected for the regression. McNemar (1949) points out that:

> A practical question of considerable importance arises when one wonders whether inclusion of additional variables in the multiple regression equation leads to a significant increase in the accuracy of prediction or when one wishes to know whether the dropping of certain variables results in a significant decrease in the amount of variance predicted. The inclusion of additional variables in the equation always tends to reduce the error of estimate somewhat and leads to an increase in R (multiple correlation coefficient).

Methods presently available can determine the critical value of R^2_k (coefficient of determination for a linear multiple regression equation with k independent variables) for some given sample size n, number of independent variables k and some predetermined level of significance (α). These critical values have already been computed and published in Crow, Davis and Maxfield (1960). Existing stepwise regression procedures will process collected multivariable data and enter the independent variables into the regression in a hierarchical sequence of significance (most significant to least significant). An analysis of variance of total regression is computed as each independent variable is included in the regression analysis. The existing Biomedical Computer Programs (1970) also compute the appropriate F test statistic to determine if the increase in R^2 indicates statistical significance. However, it has to be emphasized that this procedure advises the researcher of this additional "net regression worth" only in the context of collected data. This study proposes a methodology for determining the minimum significant value of R^2_{k+1} prior to the data collection on some additional independent variable(s). This information could be used by the researcher (now a decision maker) to determine if the statistically significant increase in the multiple correlation coefficient is feasible and worth the time, effort and costs involved in obtaining the additional data.

The form of a linear multiple regression equation is:

$$Y = b_0+b_1X_1+b_2X_2+\ldots+b_kX_k \qquad (1)$$

The common measure of "goodness of fit" is the coefficient of determination which is computed as the ratio of explained variance (by regression) to the total variation. The notation for this ratio is: R^2_k (2)

It can be shown that a test of significance for regression is:

$$F=\frac{(n-1)(S^2_y)(R^2_k)}{k} \Big/ \frac{(n-1)(S^2_y)(1-R^2_k)}{n-k-1} \qquad (3)$$

$$R_k=\left[\frac{F_{(\alpha),(k),(n-k-1)}}{F_{\text{(as above)}}+\frac{n-k-1}{k}}\right]^{1/2} \qquad (4)$$

Therefore, given α, k (preselected), and n (the sample size), the critical values of R_k can be determined by equation (4).

Tables based on Equation 4 have already been constructed and published in Crow, Davis and Maxfield (1960). They show that the critical value of R_k would have to be if "k" variables were to be included in the prediction multiple regression equation.

MINIMUM INCREMENTAL SIGNIFICANCE

Crocker (1972), Cohen (1968) and McNemar (1949) give the formula for testing the statistical significance of additional independent variables.

This expression is given below:

$$F=\frac{(R^2_{A,B,}-R^2_A) \Big/ P_B}{(1-R^2_{A,B,}) \Big/ n-p-1} \qquad (5)$$

where

(1) R^2_A is the coefficient of determination for a function containing P_A independent variables.

(2) $R^2_{A,B}$ is the coefficient of determination for a function containing P_B independent variables.

(3) p is the sum of P_A and P_B.

(4) n is the sample size.

Equation (5) is the general solution for testing the improvement beyond R^2_A obtained by adding the independent variables from the B group. This equation can be modified to the situation outlined in this paper. The situations are:

(1) R^2_A becomes R^2_k. R^2_k is the

coefficient of determination for a function containing k independent variables.

(2) $R^2_{A,B}$ becomes $R^2_{(k+1)}$. $R^2_{(k+1)}$ is the coefficient of determination for a function containing (k+1) independent variables.

(3) $p = k + (k+1) = 2k+1$.

(4) n is still the sample size.

Equation (5) can now be written as

$$F = \frac{R^2_{k+1} - R^2_k \Big/ (k+1)}{(1-R^2_{k+1}) \Big/ n-(2k+1)-1}$$

$$F = \frac{(R^2_{k+1} - R^2_k)/(k+1)}{(1-R^2_{k+1}) \Big/ (n-2k-2)} \qquad (6)$$

The educational researcher can now assess the statistical significance of the increase in R^2 resulting from the addition of one more independent variable that results in a significant value of $R^2_{(k+1)}$. The methodology will proceed as follows:

(1) Given some sample size n and observations on k independent variables, the value of R^2_k is computed.

(2) For some given level of significance (α), the statistical significance of R^2_k can be tested using equations (3) or (4).

(3) The experimenter is considering collecting data on an additional independent variable. However, it would be practical to determine what the value of $R^2_{(k+1)(min)}$ must be in order for the new variable to be considered to have contributed significant information. This question can be answered by solving equation (6) for $R^2_{(k+1)(min)}$ by proceeding as follows:

$$F_{(\alpha),(k+1),(n-2k-2)} = \frac{R^2_{(k+1)(min)} - R^2_k}{(k+1)} \times \frac{n-2k-2}{1-R^2_{(k+1)(min)}}$$

Let: $\frac{n-2k-2}{k+1} = Z$

$$F_{(\alpha),(k+1),(n-2k-2)} = F$$

Then: $$F = \frac{R^2_{(k+1)(min)} - R^2_k}{1-R^2_{(k+1)(min)}} \cdot Z$$

Therefore solving for $R^2_{(k+1)(min)}$

$$R^2_{(k+1)(min)} = \frac{F+ZR^2_k}{F+Z}$$

Substituting the original variables for F and Z the solution is:

$$R^2_{(k+1)(min)} = \left[F_{(\alpha),(k+1),(n-2k+2)} + \frac{n-2k-2}{k+1} \cdot R^2_k \right] \Big/ F_{(above)} + \frac{n-2k-2}{k+1} \qquad (7)$$

Because all the parameters on the right side of equation (7) are known or can be determined, the researcher can compute the minimum incremental value of $R^2_{(k+1)(min)}$. Although equation (7) can be computed without too much difficulty, it might be desirable to express the values of $R^2_{(k+1)(min)}$ in some tabular form. Given values of α, n, R^2_k and k, all terms except the value of $R^2_{(k+1)(min)}$ are determined.

Table 1 determines the values of $R^2_{(k+1)(min)}$ for $\alpha = .05, .01$ and sets of values for n and k.

Use of Table 1 Example (Incremental Significance) Table 1 is based on equation (7).

1. The following information is known by the researcher:
   (a) The sample size n = 50.
   (b) The equation contains one independent variable.
   (c) The correlation coefficient is equal to .4678 ($R^2 = .2188$).
2. The experimenter wishes to include an additional independent variable only if the increment is statistically significant at $\alpha = .05$.

Therefore:

(a) The sample size remains the same (n = 50).

(b) The five percent level of significance is chosen. The pieces to equation (7) become:

$$F_{(.05),(2),(50-2-2)} = F_{(.05),(2),(46)} = 3.21$$

$$\frac{n-2k-2}{k+1} = \frac{50-2-2}{2} = \frac{46}{2} = \frac{23}{}$$

$$R^2_k = (.4678)^2 = .2188$$

Using equation (7)

$$R^2_{(k+1)(min)} = \frac{3.21+(23)(.2188)}{3.21+23} = \frac{8.24}{26.21} = .3143$$

The same results can be obtained from Table (1):

(c) The level of significance (α) = .05.

(d) n = 50, k = 1, $R^2_{k=1} = .22$.

(e) The value for $R^2_{(k=1)(min)}$ that is obtained from the above coordinates is approximately 0.30. Actual

interpolation would be:

R^2_k	$R^2_{(k+1)}$
.20	.298
.22	x
.25	.342

Therefore by interpolating directly:

$$\frac{.25 - .20}{.25 - .22} = \frac{.342 - .298}{.342 - x}$$

x = .3156. This agrees with the result obtained above.

The value of $R_{(k+1)(min)} = \sqrt{R^2_{(k+1)(min)}}$

Therefore:

$R_{(k+1)(min)} = .5607.$

REFERENCES

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

Cohen, J. Multiple Regression As A General Data-Analytic Model. Psychological Bulletin, 1968, 70, 426-443.

Crocker, D. Some Interpretations of the Multiple Correlation Coefficient. The American Statistician, 1972, ____, 31-33.

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

NcNemar, Q. Psychological Statistics. New York: John Wiley and Sons, 1949.

**Because of space limitations, it was not feasible to include the tables. Copies of the tables will be furnished upon request by writing to:

Eugene Dutoit
US Army Infantry School
ATZB-CD-CS
Fort Benning, GA 31905

or

Douglas Penfield
Graduate School of Education
Rutgers, The State University
New Brunswick, NJ 07003