On a Correspondence Between *t* and *F* Values in Multiple Regressions

POTLURI RAO*

In empirical research we sometimes face the problem of selecting a subset of independent variables from a given list according to an objective. One objective that is commonly used is the maximization of $\bar{R}^2$, the square of the multiple correlation coefficient adjusted for the number of degrees of freedom. Haitovsky [3] showed that discarding an independent variable with t-value less than unity in magnitude increases the $\bar{R}^2$ of a multiple regression equation. This rule—elegant and simple—is useful only when just one variable is to be discarded. That is, if we employ the t-value rule in a sequential manner to discard variables one at a time we do not necessarily obtain the highest possible $\bar{R}^2$. Edwards [1] showed that when more than one variable is being discarded the relevant statistic to be considered is the F-value. Though the F-value rule gives us a correct answer, it is computationally tedious; with n independent variables we have to compute $2^n - 1$ different F-values to obtain the subset with highest $\bar{R}^2$. The t-values are a standard feature on most regression programs and are easily available to researchers. In this paper we derive a necessary condition relating t and F-values in order to generalize the t-value rule applicable when discarding more than one independent variable at a time. This condition is stated in general terms so that it also may be used when the objective is to discard any number of variables on the basis of F-values with a stated probability of a Type I error.

Consider the two equations estimated by least squares, expressed here in Yule's notation (see Yule and Kendall [4]) as

$$y = b_{y1\cdot 23\ldots n}x_1 + \cdots + b_{yn\cdot 123\ldots(n-1)}x_n + e_{y\cdot 123\ldots n}, \quad (1)$$

$$y = b_{y(k+1)\cdot(k+2)\ldots n}x_{k+1} + \cdots + b_{yn\cdot(k+1)(k+2)\ldots(n-1)}x_n + e_{y\cdot(k+1)(k+2)\ldots n}, \quad (2)$$

where equation (2) is obtained by discarding the first k independent variables, $x_1, x_2, \ldots, x_k$, of equation (1). All variables (x's and y) are measured as deviations from their respective means; hence the constant term is implicit. The b are the least squares estimates and the e's are residuals. The observation subscript is suppressed for simplicity of notation. These equations are estimated from data on T observations. The number of degrees of freedom of equation (1) is $\nu = T - n - 1$ and that of equation (2) is $T - (n - k) - 1$.

Let us consider the F-value used in testing the null hypothesis that the parameters associated with the k discarded variables are simultaneously zero. The F-value is defined as

$$F(k, \nu) = \frac{\{\Sigma e^2_{y\cdot(k+1)(k+2)\ldots n} - \Sigma e^2_{y\cdot 123\ldots n}\}/k}{(\Sigma e^2_{y\cdot 123\ldots n})/\nu}, \quad (3)$$

where the summation (Σ) is carried over all T observations.

The F-value rule for maximization of $\bar{R}^2$ states that whenever $F(k, \nu) < 1$ equation (2) will have a higher $\bar{R}^2$ than equation (1). In a general context we may want to discard the k independent variables whenever $F(k, \nu) < c$, where c is a critical value corresponding to a stated probability of a Type I error under the assumptions of classical regression specification.

The F-value defined in equation (3) is functionally related to the t-values of the k discarded variables in equation (1). It is possible to infer the F-value from the t-values of the discarded variables without having to estimate equation (2). A necessary condition relating the F and the t-values may be expressed by the following t-value rule:

If $F(k, \nu) \leq c$ then the absolute t-value of each of the k discarded variables must be less than $\sqrt{kc}$.

Let us consider the case of maximizing $\bar{R}^2$. In this case we want to discard the k independent variables whenever $F(k, \nu) < 1$. This sets the value of c to unity. Whenever $F(k, \nu)$ is less than unity the absolute t-value of each of the k discarded variables must be less than $\sqrt{k}$. That is, *if we do not have at least k independent variables with absolute t-value less than $\sqrt{k}$, $\bar{R}^2$ cannot be increased by discarding k independent variables at a time.* This is a necessary but not a sufficient condition. If there are k or more independent variables with absolute t-value less than $\sqrt{k}$, we may or may not be able to increase $\bar{R}^2$ by discarding k independent variables. But should $\bar{R}^2$ increase by discarding k independent variables, the variables to be discarded must come from the set of independent variables with absolute t-values less than $\sqrt{k}$.

Let us illustrate the rule by a numerical example. Let a regression equation with 5 independent variables have absolute t-values arranged in ascending order of magnitude as 1.40, 1.41, 1.74, 2.01 and 2.24. In this example we see that the t-value rule is met for $k = 1, 3, 4, 5$. That is, if we discard any 1, or 3, or 4, or 5 variables at a time $F(k, \nu)$ will not be less than unity, and hence $\bar{R}^2$ will not increase. The rule is not met for $k = 2$. We have at least two independent variables with absolute t-values less than $\sqrt{2}$, namely x_1 and x_2. $\bar{R}^2$ may increase when two independent variables are discarded at a time. Should $\bar{R}^2$ increase as a result of discarding two variables at a time, the variables discarded cannot be anything else but x_1 and x_2. Since all the other possible regressions have lower $\bar{R}^2$ than the existing equation it follows that the subset

* Dept. of Economics, Univ. of Washington, Seattle, WA 98195.

with highest $\bar{R}^2$ corresponds to $(x_1, \ldots, x_5)$ or (x_3, x_4, x_5). Notice that in this example even though we have five independent variables we can obtain the subset with highest $\bar{R}^2$ by estimating only two regression equations.

A proof for the necessary condition is given below. In order to conserve space let us use Q in place of the expression $(k + 1)(k + 2) \ldots n$ in our notation. For consistency of notation let us rewrite equation (3) as

$$F(k, \nu) = \frac{\{\Sigma e^2_{y \cdot Q} - \Sigma e^2_{y \cdot 12 \ldots kQ}\}/k}{(\Sigma e^2_{y \cdot 12 \ldots kQ})/\nu} \tag{4}$$

The condition that $F(k, \nu) \leq c$ may be restated as

$$\{\Sigma e^2_{y \cdot Q} - \Sigma e^2_{y \cdot 12 \ldots kQ}\}/k \leq c(\Sigma e^2_{y \cdot 12 \ldots kQ})/\nu,$$

which in turn may be rewritten as

$$\nu \Sigma e^2_{y \cdot Q} \leq (\nu + kc)\Sigma e^2_{y \cdot 12 \ldots kQ}. \tag{5}$$

Using the properties of least squares derived in Yule and Kendall [4] we have

$$\Sigma e^2_{y \cdot 12 \ldots kQ} = \Sigma e^2_{y \cdot Q}(1 - r^2_{y1 \cdot Q}) \cdots (1 - r^2_{yk \cdot 12 \ldots (k-1)Q}) \tag{6}$$

where the r's are partial correlation coefficients. Since

$$r^2_{yi \cdot 12 \ldots p} = t^2_{yi \cdot 12 \ldots p}/(t^2_{yi \cdot 12 \ldots p} + \mu),$$

where $t_{yi \cdot 12 \ldots p}$ is the t-value corresponding to $b_{yi \cdot 12 \ldots p}$ and μ is the number of degrees of freedom of the corresponding regression equation (see Gustafson [2]), we have

$$\Sigma e^2_{y \cdot 12 \ldots kQ} = \Sigma e^2_{y \cdot Q}(\nu + k - 1)(\nu + k - 2) \cdots \nu/\{(t^2_{y1 \cdot Q} + \nu + k - 1) \cdot (t^2_{y2 \cdot 1Q} + \nu + k - 2) \cdots (t^2_{yk \cdot 12 \ldots (k-1)Q} + \nu)\} \tag{7}$$

By substituting equation (7) in (5) the F condition may be restated as

$$(t^2_{y1 \cdot Q} + \nu + k - 1) \cdots (t^2_{yk \cdot 12 \ldots (k-1)Q} + \nu) \leq (\nu + kc)\{(\nu + k - 1)(\nu + k - 2) \cdots (\nu + 1)\}. \tag{8}$$

The left hand side of the inequality may be rewritten as

$$\{\phi + (\nu + k - 1)(\nu + k - 2) \cdots (\nu + 1)\}(t^2_{yk \cdot 12 \ldots (k-1)Q} + \nu), \tag{9}$$

where ϕ stands for the remaining terms in the expansion of the product of the first $(k - 1)$ terms of the left hand side. Since each term in the summation is a positive quantity it follows that $\phi > 0$. We may rewrite equation (8) as

$$\{\phi + (\nu + k - 1)(\nu + k - 2) \cdots (\nu + 1)\} (t^2_{yk \cdot 12 \ldots (k-1)Q} + \nu) \leq (\nu + kc) \cdot \{(\nu + k - 1)(\nu + k - 2) \cdots (\nu + 1)\},$$

which reduces to

$$\phi(t^2_{yk \cdot 12 \ldots (k-1)Q} + \nu) \leq (kc - t^2_{yk \cdot 12 \ldots (k-1)Q}) \{(\nu + k - 1)(\nu + k - 2) \cdots (\nu + 1)\} \tag{10}$$

The condition expressed in equation (10) is nothing but a restatement of $F(\kappa, \nu) \leq c$. Equation (10) is computationally tedious. We shall use only a part of the information contained in equation (10) to derive a necessary condition for $F(k, \nu) \leq c$.

It is obvious that the expression on the left hand side of the inequality in equation (10) is positive. If the inequality is met then the right hand side of the inequality must be positive, which is possible only when $(kc - t^2_{yk \cdot 12 \ldots (k-1)Q})$ is positive. Hence a necessary condition for the inequality $F(k, \nu) \leq c$ is that

$$t^2_{yk \cdot 12 \ldots (k-1)Q} < kc. \tag{11}$$

But $t_{yk \cdot 12 \ldots (k-1)Q}$ is nothing but the t value of x_k in equation (1). If $F(k, \nu) \leq c$ then the absolute t-value of x_k must be less than $\sqrt{kc}$. Since the choice of x_k is arbitrary, by rearranging the order of the k discarded variables it follows that the absolute t-value of each of the k discarded variables must be less than $\sqrt{kc}$.

Whenever the inequality (10) holds the condition stated in equation (11) is met. But whenever equation (11) is met the inequality (10) need not hold. Hence the t-value rule stated here is necessary but not sufficient.

REFERENCES

[1] Edwards, J. B., "The relation between the F-test and $\bar{R}^2$," *The American Statistician,* Vol. 23, No. 5, (Dec., 1969), p. 28.

[2] Gustafson, R. L., "Partial correlations in Regression Computations," *Journal of American Statistical Association,* Vol. 56, No. 294, (June, 1961), pp. 363–367.

[3] Haitovsky, Y., "A Note on the Maximization of $\bar{R}^2$," *The American Statistician,* Vol. 23, No. 1, (Feb, 1969), pp. 20–21.

[4] Yule, G. U. and M. G. Kendall, *An Introduction to the Theory of Statistics,* Charles Griffin & Company Limited (London), 1950, Fourteenth Edition.