A program for approximating R^2 probability values in best subset multiple regression

JEFFREY LEE RASMUSSEN
University of Arizona, Tucson, Arizona

Best subset multiple regression is a controversial statistical procedure due to the fact that the typical F statistic employed to test the significance of the multiple regression coefficient is biased. Best subset regression may lead to overfitting of the data due to capitalizing on chance covariation of predictors with the criterion. Due to this overfitting, subsequent cross-validation or replication often leads to substantial shrinkage in the multiple correlation coefficient.

Diehr and Hoflin (1974) offered a solution to the bias in using the F ratio as a test of statistical significance in multiple regression. They utilized a Monte Carlo simulation in order to approximate the distribution under the null hypothesis of best subset $R^2(k, m, n, a)$, where k is the number of predictors selected, m is the total number of predictors in the regression equation, n is the sample size, and a is the significance level. The maximum value of R^2 was determined for each of 100 simulations of best subset R^2 under the null hypothesis of independence between the predictors and the criterion. These 100 values were ordered from least to greatest, and the 95th R^2 value was taken as an estimate of the R^2 value significant at $a = .05$. The 99th R^2 value was taken as an estimate of the R^2 critical value for $a = .01$. They were able to fit a function of

$$R^2(k, m, n, a) = w(1 - v^k) \quad (1)$$

to approximate the results of the simulation. The values of w and v were determined from the two known values of $R^2(1, m, n, a^*)$ and $R^2(m, m, n, a)$, where $a^* = 1 - (1 - a)^{1/m}$. These R^2 values were derived from the F distribution, where $R^2(k, m, n, a) = mF(m, n - m - 1, a) / [n - m - 1 + mF(m, n - m - 1, a)]$. Diehr and Hoflin (1974) present tables of R^2 critical values for $a = .10$, $.05$, and $.01$ for a very limited number of values of m , n , and k . They suggested that researchers use the critical values tabled in their article or, for those values not tabled, use standard F tables along with Equation (1) to derive the appropriate critical values. Using such critical values, instead of the typical F -statistic critical values, will help researchers guard against using regression equations that overfit the data.

The present program has three advantages over using the tables and equation in the Diehr and Hoflin (1974) article. First, the Diehr and Hoflin article tables only

the R^2 critical values for $a = .10$, $a = .05$, and $a = .01$. Likewise, Equation (1) solves only for an R^2 critical value at a given probability level. The present program solves for α at a given R^2 value. Thus, instead of indicating whether a given R^2 value is significant at the less than $.01$ or less than $.05$ level, the present program iteratively re-solves Equation (1) until it arrives at the exact value of α for a given R^2 value.¹ Second, the program can be made a subprogram of existing software packages that calculate best subset multiple regression (e.g., Hansche & Dunlap, 1976). This obviates the need to consult the Diehr and Hoflin (1974) reference, and can also save computer time if the R^2 probability values are used as a criterion for selection of best subset regression equations. Third, the solution to Equation (1) requires getting critical values of F for α^* . Since F -distribution tables are difficult to locate for values of other than $\alpha = .10$, $.01$, and $.05$, solving Equation (1) for nonstandard α values may be difficult and time-consuming. The present program, however, uses a function that calculates an F -ratio critical value for any α value, thus eliminating the need to look up non-standard α values.

The program consists of a simple example main program and a series of routines that iteratively converge on the probability value for the given R^2 value. The program requires probability functions by Dunlap and Duffy (1975), Dunlap, Marx, and Agamy (1981), and Rasmussen (1983). The program was tested using critical values tabled in the Diehr and Hoflin (1974) article, and the associated probability levels generated by the program agreed within rounding error with the tabled values. The program requires approximately 300 msec per R^2 probability value when run on a Digital Equipment Corporation DEC-10 computer.

Requirements. The main routine and functions are written in single-precision FORTRAN IV, and should run on any computer with a FORTRAN IV compiler. The main program, which contains two features specific to DEC-10 FORTRAN, is included in the listing for illustrative purposes; users probably would want to dispense with the main program and use the functions as part of another program. Those users wanting to keep the main routine would probably need to make two changes. First, on the DEC the terminal is defined as unit 5 in an OPEN statement; terminal definition is probably different on other systems. Second, the use of an asterisk in a READ statement, which causes free-format input on the DEC, must be changed to input with a FORMAT statement on some other computer systems.

Program Availability. A listing of the main program and functions may be obtained free of charge from Jeffrey Lee Rasmussen, Department of Psychology, University of Arizona, Tucson, AZ 85721.

The author's mailing address is: Department of Psychology, University of Arizona, Tucson, AZ 85721.

REFERENCES

- DIEHR, G., & HOFLIN, D. R. (1974). Approximating the distribution of the sample R^2 in best subset regressions. *Technometrics*, 16, 317-320.
- DUNLAP, W. P., & DUFFY, J. A. (1975). FORTRAN IV functions for calculating exact probabilities associated with z , chi-square, t , and F values. *Behavior Research Methods & Instrumentation*, 7, 59-60.
- DUNLAP, W. P., MARX, M. S., & AGAMY, G. J. (1981). FORTRAN IV functions for calculating probabilities associated with Dunnett's test. *Behavior Research Methods & Instrumentation*, 13, 363-366.
- HANSCH, W. J., & DUNLAP, W. P. (1976). A FORTRAN IV program for computing all possible multiple regressions. *Behavior Research Methods & Instrumentation*, 8, 393-394.
- JASPEN, N. (1965). The calculation of probabilities corresponding to values of z , t , F , and chi-square. *Educational and Psychological Measurement*, 25, 877-880.
- RASMUSSEN, J. L. (1983). *A correction of a function for deriving F ratio probability values*. Manuscript submitted for publication.

NOTE

1. The program solves for an exact value of a for Equation 1. Equation 1, however, is an approximate fit to the null distribution of best subset R^2 , and therefore a is an approximate, and slightly conservative, estimate of the actual probability level.

(Revision accepted for publication January 30, 1984.)