Staff Papers Series

Staff Paper P77-2 January 1977

THE ACCURACY OF REGRESSION PROGRAMS
UNIVERSITY OF MINNESOTA

by

Maury E. Bredahl and Ann Mylander

Department of Agricultural and Applied Economics

University of Minnesota
Institute of Agriculture, Forestry and Home Economics
St. Paul, Minnesota 55108

THE ACCURACY OF REGRESSION PROGRAMS

UNIVERSITY OF MINNESOTA

Maury E. Bredahl and Ann Mylander

A recent AJAE article compared the accuracy of regression results produced by several regression packages and computer systems [Boehn, Menkhaus, and Penn.] Erroneous results of large magnitude were discovered prompting the authors to recommend

> "Given the serious nature of the issue, perhaps the editor of our journal should require authors to submit regression results from more than one computer program prior to accepting papers for publication."

The accuracy test employed by Boehn et. al. was used to test the accuracy of four computed programs used for economic research at the University of Minnesota. By and large, accurate results were obtained from the regression programs. However, an uninformed user of these programs could very easily produce erroneous results. This paper identifies key program parameters and options that researchers should note in order to produce accurate results.

Multicolinearity. The referenced article suggests an exact specification of an equation (no random error) and highly correlated data. The test equation is

$$Y = 1 + X + X^2 + X^3 + X^4 + X^5 \qquad (X = 0, 1, 2, \ldots, 20)$$

The data constructed in this manner are highly correlated as evidenced by the simple correlation coefficients given in Table 1.

Table 1. Simple Correlation Coefficients

X^2	.971			
X^3	.922	.986		
X^4	.873	.960	.992	
X^5	.829	.929	.976	.995
	X	X^2	X^3	X^4

Highly correlated data (price series for example) are often used in estimation of supply or demand equations. As such, the ability of a computer program to accurately compute regression coefficients and associated standard errors is very important in agricultural economics research.

Most often researchers examine the matrix of simple correlation coefficients in order to identify multicolinearity. However, in many cases, multicolinearity often involves several variables. Serious multicolinearity results when an explanatory variable is an approximate linear function of 2 or more other explanatory variables.

Regression packages which estimate regression equations in a step-wise or forward selection procedure use measures of multicolinearity as a criterion for including variables in the equation. Such computer programs effectively regress each explanatory variable which is not included in the equation on all explanatory variables already included in the equation. The coefficient of determination (R^2) is calculated,

subtracted from one and the resultant value is termed the tolerance of the excluded variable. The tolerance is a measure of the proportion of the variation of an excluded explanatory variable not explained by other variables in the equation. If the excluded explanatory variable is orthogonal to (uncorrelated with) all included explanatory variables, the tolerance would be 1. If, on the other hand, the excluded variable is an exact linear combination of included variables, the tolerance would be zero.

The stepwise or forward selection regression programs utilize an arbitrary default value for the tolerance level for selecting variables for an equation. If highly colinear data are being analyzed, that arbitrary level may be exceeded and variables excluded from the equation which should be included. Indeed, many of the erroneous results reported in the referenced article were produced because the tolerance level was violated and not because the regression programs were inaccurate.

Accuracy of U of M Programs. Four U of M regression packages were tested:

(1) ANALYZE (U of M Economics Dept.)

(2) SP22 (St. Paul Computer Center)

(3) BMD02R (Health Sciences Computing Center, UCLA)

(4) SPSS (Statistical Package for Social Sciences)

The computational methods used by each computer program are not discussed. The estimated coefficients, standard errors, and standard error of the estimates are given in Table 2.

The ANALYZE AND SP22 programs provided acceptable coefficient estimates but standard errors and the estimated standard error of the

Table 2. Estimated Results of Selected U of M Regression Packages

	Constant	X_1	X_2	X_3	X_4	X_5	R^2	S.E.E.
Correct Values								
Coefficient	1	1.0	1.0	1.0	1.0	1.0	1.00	0
S.E.	0	0	0	0	0	0		
ANALYZE Estimated Values								
Coefficient	1.00001508	0.99996084	1.00000389	0.99999858	0.99999999	1.00000000	1.00	95.9528530
S.E.	87.50387874	96.0913545	31.68459305	4.12553651	0.22952728	0.00456662		
SP22 Estimated Values								
Coefficient	1.0000137	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00	4.6156272
S.E.	4.6156272	5.0685966	1.6712890	0.21761251	0.01210703	0.00024087		
BMD02R Estimated Values								
Tolerance Level = .001								
Coefficient	11.39050	-19.85976	8.55626	--	1.05516	0.99892	1.00	5.8144
S.E.	--	2.59053	0.34342	--	0.00181	0.00006		
Tolerance Level = .0001								
Coefficient	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00	0
S.E.	0	0	0	0	0	0		
SPSS Estimated Value								
Coefficient	0.99999834	1.0000023	1.00000000	.99999934	1.00000001	1.00000000	1.00	0
S.E.	0	0	0	0	0	0		
SPSS Stepwise - Estimated Values								
Coefficient	11.390495	-19.859759	8.5562608	--	1.0551647	0.99892245	1.00	5.81
S.E.	4.6230415	2.590786	0.34345050	--	.00180262	0.6332264 E-4		

estimate were not correct. The reported standard errors of the ANALYZE program would have lead one to incorrect specification of the equation.

The SPSS and BMD02R program would give appropriate results if the programs were used correctly. The BMD02R program utilizes a stepwise regression procedure. If the default level for the tolerance is used, erroneous results will be reported. The program excludes X^3 from the estimated equation because the tolerance level is violated. The correct specification includes all variables. Therefore, estimated equations which exclude a variable are necessarily incorrect. If the tolerance level is decreased by the program user, correct results are obtained.

Much the same results are obtained from the SPSS computer program. Unlike BMD02R, which automatically uses a stepwise procedure, SPSS regressions may be run as exactly specified equations or in a stepwise manner. If exact specification is used, the estimated coefficients and standard errors are correct. If the optional stepwise procedure is used, the tolerance level will be violated which excludes appropriate variables and necessarily yields incorrect results.

Thus, the stepwise or forward selection regression packages should be used sparingly in economic analysis. By and large, specification of economic relationships should be based on theoretical considerations and not on a "search" over a list of variables using stepwise or forward selection procedures.

The ANALYZE and SP22 regression packages did not produce acceptable estimated results. BMD02R produced correct results if the tolerance level was decreased. Because BMD02R always uses a stepwise regression procedure, errors due to default values of program parameters may

consistently occur. The SPSS program produced correct estimates and given the simplicity of program use would seem to be the regression package to use for economic analysis at the U of M.

References

Boehm, William T., D.J. Menkhaus and J.B. Penn, Accuracy of Least Squares Computer Programs: Another Reminder, A.J.A.E., 58 (Nov., 1976), 757-760.

Dixon, W.J., ed. BMD Biomedical Computer Programs. Health Sciences Computing Center, University of California, Los Angeles, 1964.

Nie, Norman, D.H. Bent and C.H. Hull, Statistical Package for Social Sciences, New York: McGraw Hill Book Co., 1970.

Time Series Analyzer, Users Manual, mimeo, 1976.

University Computer Center, Statistical Package for the Social Sciences, Version 6.0, University of Minnesota, August, 1975.