THE DELETION OF VARIABLES FROM REGRESSION MODELS BASED ON TESTS OF SIGNIFICANCE: A STATISTICAL AND MORAL ISSUE*

David L. Debertin and R. J. Freund

The purpose of this paper is to illustrate some of the dangers inherent in use of statistical tests as a criterion for deleting variables from regression models. The deletion of variables from regression models based on t or F tests of regression coefficients has been a procedure widely followed by applied economists and other researchers. When economic theory does not provide an adequate conceptual basis for rigorous *a priori* specification of the regression model, one approach to model specification has been to include in the regression equation all variables thought to be "somehow" related to the dependent variable of interest. Subsets of variables with statistically significant coefficients are identified, with the aid of a stepwise regression routine.[1] Truncated models consisting of only those variables with statistically significant regression coefficients are sometimes presented in the published research without reference to the initial data dredging that took place.[2]

Such procedures entail the dredging of data for type I error, therefore altering the performance of statistical tests of coefficients of the truncated model. Parameter estimates from the truncated model are no longer unbiased. Statisticians are quite aware of these problems. Selvin and Stuart [8] have commented that ". . . if we decide on the basis of the data to discard one or more variables from an explanatory equation, we cannot apply standard statistical tests to retained variables in the equation as though nothing has happened." Draper and Smith [1] have argued that variable selection procedures, such as stepwise, can easily be abused by the amateur statistician.

Warnings by statisticians appear to have been largely unheeded by some researchers in the profession. Statistical tests continue to be applied by some researchers to truncated models as if the tests were valid. It is important to recognize that biased coefficients and invalid statistical tests are of importance to applied economic research. Results presented in this paper illustrate that the dangers inherent in variable selection procedures, based on significance tests of regression coefficients, can substantially influence the validity of the research.

METHODOLOGY

To illustrate effects of variable selection based on statistical tests of significance, a linear model was fitted to sets of numbers produced with a random number generator. Both the dependent and independent variables were random numbers produced by a random number generator. The familiar statistical model was assumed.

$$Y = \beta o + \beta_1 X_1 + \ldots \beta_g X_g + \varepsilon \quad (1)$$

where

Y = an endogenous (dependent) variable

$X_1, \ldots X_g$ = predetermined (independent) variables

$\beta o \ldots \beta_g$ = structural parameters to be estimated

ε = an error term

David L. Debertin is Assistant Professor, Department of Agricultural Economics, University of Kentucky. R. J. Freund is Professor, Institute of Statistics at Texas A&M University. Statistical analysis was conducted while Dr. Debertin was at Purdue University. Journal paper No. 5117 of the Purdue Agricultural Experiment Station.

* The authors are indebted for the assistance provided by E. W. Kehrberg, T. K. White, J. Havlicek, G. L. Bradford, Alan J. Randall, Eldon D. Smith and L. D. Jones. Any errors remain the responsibility of the authors.

1 There are a number of alternative regression routines for identifying subsets of explanatory variables with statistically significant coefficients. All achieve similar results.

2 See, for example [2, 3, 4, 5, 6, 7]. The technique is used every day by agricultural economists and other researchers.

The sample used to estimate (1) consisted of 25 observations on a normally distributed (0, 1) dependent variable and 20 uniformly distributed (0-1) independent variables.[3] Under ordinary least squares assumptions, significant regression coefficients should occur only as a result of type I error, and the number of significant regression coefficients found in each regression equation should follow the binomial distribution.

$$b(z;\ n, \alpha) = \binom{n}{z}\ \alpha^{z}\ (1 - \alpha)^{n - z} \qquad (2)$$

where

- b = the probability of x significant regression coefficients in n trials
- n = the number of trials (equal to the number of variables initially presented for possible inclusion in the regression models.)
- z = the number of significant regression coefficients
- α = selected significanece level

A binomial probability table for selected significance levels therefore indicated probabilities of finding exactly z significant regression coefficients (at alternative α levels) if n (in this case, 20) variables are presented for possible inclusion in the regression equation (Table 1). One hundred regression equations were run. Expected numbers of significant regression coefficients occurring as a result of type I error are merely probabilities from Table 1 multiplied by 100. Counts of "significant" regression coefficients were done after stopping the stepwise regression routine at 7 steps, a procedure similar to that followed by researchers.

Table 1. PROBABILITIES OF FINDING EXACTLY x SIGNIFICANT REGRESSION COEFFICIENTS WHEN 20 VARIABLES ARE PRESENTED FOR POSSIBLE INCLUSION IN THE MODEL (VALUES OF THE BINOMIAL DISTRIBUTION)

Expected Number of Significant Variables (z)	Selected Significance Levels			
	.20	.10	.05	.025
0	.012	.122	.358	.603
1	.058	.270	.377	.309
2	.137	.285	.189	.075
3	.205	.190	.060	.012
4	.218	.090	.013	.001
5	.175	.032	.002	.000
6	.109	.009	.000	.000
Over 6	.086	.003	.000	.000

(Columns may not total to 1.00 due to rounding.)

[3] The use of normally distributed y ensures that the ordinary least squares assumption that ε is normally distributed is met. The X matrix is considered to be a set of fixed numbers and β a set of constants (parameters). Hence, ε takes on the same assumed distribution as y. The y vector is $N(\mu_y, \sigma^2)$ while ε is $N(o, \sigma^2)$. There is no difficulty with the assumption that uniformly distributed random numbers used in forming the X matrix when generated take on fixed values for purposes of running the regression. Hence, the regressors themselves are actually non-stochastic, but independent.

APPRAISAL OF RESULTS

Observed numbers of significant regression coefficients far exceed expected numbers based on binomial probabilities when the degrees of freedom associated with final, rather than original, variable set is used. For example, at the 0.20 probability level, 35 of the 100 equations were found to have over 6 significant coefficients. The binomial formula predicted 8.6 equations with over 6 significant coefficients.

Hence, the appropriate degrees of freedom for testing the significance of regression coefficients is not degrees of freedom associated with the final regression equation after nonsignificant variables have been deleted (Tables 2 and 3). Use of degrees of freedom associated with the final regression equation for t tests of regression coefficients leads to far too many variables being called significant. Use of degrees of freedom associated with the final regression equation for tests of regression coefficients has been a widespread and serious error in applied statistical research.

Table 2. EXPECTED AND ACTUAL NUMBERS OF SIGNIFICANT REGRESSION COEFFICIENTS, 100 EQUATIONS, .20 PROBABILITY LEVEL

Number of Significant Coefficients	Assumed Degrees of Freedom		Expected Number
	(25 - 7 - 1)[a]	(25 - 20 -1)[b]	
0	0	0	1.2
1	1	1	5.8
2	3	12	13.7
3	3	6	20.5
4	16	20	21.8
5	16	13	17.5
6	27	27	10.9
Over 6	35	21	8.6
Total	100	100	100.0

[a] The degrees of freedom associated with the final variable set.
[b] The degrees of freedom associated with the original variable set.

Furthermore, results indicate that use of degrees of freedom associated with the original variable set also leads to greater than the expected number of "significant" regression coefficients. Regression equations were also estimated using random numbers from a table, rather than numbers generated by a random number generator, with similar results. Wallace and Ashar [9] suggest one possible explanation why binomial probabilities may not strictly apply to a truncated regression model, even if degrees of freedom associated with the original rather than final variable set are used in performing the t-test. They argue that if variables have been sequentially deleted, estimated variances of remaining coefficients are no longer estimates of variances of coefficients when all variables are in the model. Hence, binomial probabilities no longer apply as

Table 3. EXPECTED AND ACTUAL NUMBERS OF SIGNIFICANT REGRESSION COEFFICIENTS, 100 EQUATIONS, .05 PROBABILITY LEVEL

Number of Significant Coefficients	Assumed Degrees of Freedom		Expected Number
	(25 - 7 - 1)[a]	(25 - 20 - 1)[b]	
0	6	25	35.8
1	14	27	37.7
2	14	18	18.9
3	19	16	6.0
4	23	7	1.3
5	13	4	0.2
6	9	3	0.0
Over 6	2	0	0.0
Total	100	100	100.0

[a] The degrees of freedom associated with the final variable set.

[b] The degrees of freedom associated with the original variable set.

they would to a regression model in which variables had not been sequentially deleted.

THE MORAL DILEMMA

It might be useful to distinguish between two types of data dredging. The first is dredging for informational purposes. The researcher uses stepwise regression routines to dredge data for hypotheses to be tested with new data at some future time. The second is the dredging of data to test hypotheses. There is nothing particularly wrong with the former approach, provided that the researcher indicates in the published findings that hypotheses were being generated rather than tested. It is the second approach — using the same data which generated the hypotheses to test the hypotheses — that is contrary to scientific method.

It is evident from the preceeding results that, if a researcher dredges data on enough variables, application of the stepwise regression routine will eventually lead to a model specification in which most if not all estimated parameters are "Statistically different from zero." Much of the purported statistical significance will be type I error. The researcher therefore faces a serious "moral" dilemma. Reviewers for the "major" journals tend to be highly critical of regression models in which most, if not all, coefficients on explanatory variables are not larger in absolute value than the respective standard errors and when all coefficients of determination are not of an "acceptable magnitude."

Data dredging will eventually lead to a model specification consistent with these criteria. The resultant model will not only be an inaccurate representation of whatever structural phenomena it is supposed to represent, but it may be useless as a predictive model, since predictions will be based on meaningless coefficients.

The researcher who uses a stepwise regression routine must therefore make a "moral" decision as to whether or not to admit to dredging data

when reporting results of research efforts. If the researcher admits to dredging data, he risks the wrath of the "rigorous *a priori* model specification" proponents in the profession and hence find it difficult to get a journal to publish his research. If the researcher does not admit to the dredging of data, journals may be willing to accept the researchers "positive" findings (that may have occurred largely as a result of type I error). However, other research and extension personnel reading the published findings are not adequately warned that the published results consist primarily of type I error. In the case of the extension specialist, there is added danger that "incorrect" conclusions arising from data dredging with the stepwise will lead to "incorrect" decisions adversely affecting large numbers of lay citizens.

Given these alternatives, most researchers are hesitant to admit to data dredging practices. Most researchers' salaries in agricultural economics are functionally related to the number of journal articles published and not at all related to the *ex post* accuracy of their research findings. Hence, the "moral" dilemma is perhaps similar to tax evasion. It is clearly "wrong" to publish research findings based on truncated regression models without mention of the initial data dredging that took place. However, the expected penalty for the researcher is usually quite low.

SOME RECOMMENDATIONS

This appraisal suggests a number of recommendations for professional agricultural economists using these statistical techniques:

(1) Degrees of freedom associated with the original variable set, while not entirely satisfactory, are more appropriate than degrees of freedom associated with the final variable set when individual regression coefficients are tested for truncated regression models. It should become standard procedure to use degrees of freedom associated with original rather than final variable sets when tests of individual regression coefficients are made.

(2) Early in the empirical analysis, researchers should decide whether they are going to either *generate* or *test* hypotheses. Stepwise regression routines are quite useful for generating hypotheses to be tested subsequently with new data. It is only attempting to present hypotheses that were *generated* by dredging data as if they were hypotheses that were rigorously *tested* that is contrary to the scientific method.

(3) Journal reviewers who accept or reject articles based solely on acceptability of statistical results as measured by aforementioned criteria are naive. More effort needs to be directed toward publication of research results which are useful, but negative. The screening of articles by a journal reviewer based on statistical criteria of acceptability ensures that type I error will predominate in a journal.

(4) The computer is not an economist. Perhaps we did a better job of specifying our models when we were forced to estimate regression parameters by hand! Widespread use of the computer as both a fast and inexpensive means for estimating alternative regression equations has meant that many economists are willing to let the computer specify the model. There is no substitute for a well — planned conceptual model.

REFERENCES

[1] Draper, Norman and H. Smith. *Applied Regression Analysis*, John Wiley and Sons, New York, 1966.

[2] Egbert, Alvin C. "An Aggregate Model of Agriculture — Empirical Estimates and Some Policy Implications," *Journal of Farm Economics*. 51: 71-86, February 1969.

[3] Gustman, Alan L., and George B. Pidot, Jr. "Interactions Between Educational Spending and Student Enrollment," *Journal of Human Resources*. 8: 3-23, Winter, 1973.

[4] Ladd, George W. "Federal Milk Marketing Order Provisions, Effects on Producer Prices and Intermarket Price Relationships," *Journal of Farm Economics*. 51: 625-642, August 1969.

[5] Sahota, Gian S. "Efficiency of Resource Allocation in Indian Agriculture," *Journal of Farm Economics*. 50: 584-605, August 1968.

[6] Schutzer, W. A., and M. C. Hallberg. "Impact of Water Recreational Development on Rural Property Values," *Journal of Farm Economics*. 50: 584-605, August 1968.

[7] Scott, John T., and Earl O. Heady. "Regional Demand for Farm Buildings in the United States," *Journal of Farm Economics*. 49: 184-198, February 1967.

[8] Selvin, Hanan C., and Alan Stuart. "Data Dredging Procedures in Survey Analysis," *The American Statistician* 20: 20-23, June 1966.

[9] Wallace, T. D. and V. G. Ashar. "Sequential Methods in Model Construction," *Review of Economics and Statistics*. 54: 172-178, May 1972.