BIAS OF SELECTED COEFFICIENTS IN STEPWISE REGRESSION

Leonard J. Lane, Agricultural Research Service
Duane L. Dietrich, University of Arizona

Purchased By
U. S. Department of Agriculture
For Official Use

ABSTRACT

The object of this experiment is to determine the sampling properties of the statistics associated with a multiple linear regression equation as they are influenced by the stepwise selection procedure. The experiment involved generating a large number of observations using known relationships and then performing regression analyses on a large number of samples of varying sizes. Since all of the assumptions underlying regression theory were satisfied in the generating process, any anomalies in the derived equations not directly related to sampling error can be ascribed to the stepwise procedure. Results of the analyses are quite surprising in terms of the accuracy with which the regression coefficients were estimated: some of the estimated coefficients were more than twice the magnitude of their population values. The existence of a heretofore unknown bias in the "significant" estimated coefficients is demonstrated and explained in terms of distribution theory resulting from the normality assumption and the application of the stepwise procedure.

INTRODUCTION

Selection of the best multiple linear regression equation from a specified set of variables has been a problem of interest for some time. Draper and Smith [1] list six methods of selecting the best regression equation: (1) all possible regressions, (2) backward elimination, (3) forward selection, (4) stepwise regression, (5) variations on the previous methods, and (6) stagewise regression. Draper and Smith then recommend No. 4, the stepwise regression procedure, as the best of the procedures discussed. In all six methods the central F-distribution is used as a selection criterion.

Pope and Webster [4] considered the use of the F-statistic in determining the independent variables to be included and in making statements of significance. They conclude that the major problem with using the central F-distribution is that since the assumptions are not met, traditional probability statements and significance levels are incorrect. In their words, "The one major disadvantage is that no probability is available that gives any measure of goodness to this procedure of selecting the subset of independent variables." (p. 339)

This paper lists additional problems associated with the stepwise procedure and discusses their origin. The primary result of this research is the discovery of a heretofore unknown bias in the estimated coefficients determined by using the stepwise regression procedure.

THE STEPWISE REGRESSION PROCEDURE

The multiple linear regression program [Huszar (3)] selects entering variables in the order in which the unexplained variance is reduced the most with the condition that the entering variable must have a computed F-level larger than a critical value. After each step the program then tests to see if all variables in the equation have F-levels larger than some critical value, and if not, then that variable with the F-level below the critical value is removed. The algorithm continues until a significant (based upon the critical F-values) equation is determined by considering all independent variables. The stepwise procedure is explained in more detail by Efroymson (2) and Draper and Smith (1).

PROCEDURE

The experiment reported herein requires the generation of large numbers of observations from a known multiple linear relation. The generating equation is in the form of Equation 1:

$$Y = X\beta + \varepsilon \qquad (1)$$

where

- Y is an (nxl) vector of observations on the dependent variable,
- X is an (nxp) matrix of the independent variables,
- β is a (pxl) vector of coefficients,
- ε is an (nxl) vector of errors,
- n is the number of observations, and
- p is the number of independent variables plus one, $k + 1$.

Assumptions for Equation 1 are that the independent variables are independent and known without error, that observations on the dependent variable are uncorrelated random events, that the residuals (actual minus predicted Y) are not serially correlated, and that random components or errors are normally distributed with zero mean and constant variance.

In this experiment there were six independent variables simulated as six uniform variables using the pseudo-random numbers generator on the computer. The normally distributed random component was generated by a subroutine using Newton's method for finding the roots of a differentiable function defined as the difference between a uniform variable and the cumulative distribution function for a standard normal variable.

An F-value of 4.0 was chosen for entering and removing variables. This corresponds to about a 5% significance level for large ($n > 30$) samples and is a liberal criterion for small samples. Small subsamples from the large number of observations on Equation 1 were taken and multiple linear regression analyses were performed on the subsamples. This allowed determination of the sampling properties of the regression coefficients with the F-level entrance and removal

Reprinted from the 1976 Statistical Computing Section
Proceedings of the American Statistical Association©

criteria.

The next step was to compute, using distribution theory based upon the normality assumption, the theoretical values of the statistics associated with the specified regression equations. These values are then compared to the sample or derived values and an attempt was made to account for any "serious" differences in terms of the stepwise procedure, since all assumptions have been satisfied. In this manner the effectiveness of the stepwise procedure in screening independent variables was determined.

RESULTS AND DATA

The experiment is divided into four main parts depending upon the size of the variance of the random component, ε, in Equation 1. Values of the variance of ε Var(ε) or $\sigma^2_{y \cdot x}$ were 8/3, 2/3, 1/6, and 1/24, a range of 64:1. In each part of the experiment 100 samples of size 20, 50, and 100 were selected for each of the variance levels. In all cases, there were six independent variables. Specified values used in generating the 40,000 observations on Equation 1 are shown in Table 1. The experiment is designed so that the order of entry of the variable should be X_3 & X_6, X_2 & X_5, and then X_1 & X_4, where either variable within each pair is equally likely.

Data from the three sampling schemes and the four variance levels are summarized in Tables 2, 3, 4, and 5. These tables show how often each variable was judged significant and the mean and standard deviations of the selected coefficients. Each of the tables corresponds to a different variance level of the random component. Notice that in some cases there are relatively large differences between the mean values of the selected coefficients and their population values.

The results of the selection procedure are summarized in Table 6 which gives the frequency of selection for each of the seven possible outcomes. The selection procedure could judge from zero to six of the six independent variables as significant. Notice that for the smallest sample size, when the error level is high few of the "significant" equations contain all six variables. However, when the error level is low ($\sigma^2_{y \cdot x}$ = 1/24) virtually all of the equations contained all six variables.

BIAS IN ESTIMATED COEFFICIENTS

The estimated coefficients, b_j, have expected value β_j and variance estimated as

$$S^2_{b_j} = S^2_{y \cdot x} / \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2 \qquad (2)$$

where $S^2_{y \cdot x}$ is an estimate of $\sigma^2_{y \cdot x}$, calculated as the sum of squares of actual minus predicted y, divided by the n-k-1 residual degrees of freedom. The denominator, $\sum_{i=1}^{n} (X_{ij} - X_j)^2$, is the sum of squares of the j<u>th</u> independent variable about its mean. Hypothesis tests concerning b_j are then made using:

$$t = (b_j - \beta_j) / S_{b_j} \qquad (3)$$

with n-k-1 degrees of freedom. A variable was selected if its computed F-value was greater than or equal to 4.0. This is equivalent to $|t|$ = 2.0 in Equation 3. Under the null hypothesis β_j = 0 and the variable is selected for inclusion in the regression equation if $|t|$ > 2.0. Thus, only the relatively larger (absolute) values of b_j are judged "significant" and included in the equation. For this reason, average values of the selected coefficients are biased toward the larger (absolute) values, and do not have expected values of β_j. As calculated, the individual estimated values are not biased, however, as some of the estimated coefficients are judged non-significant those remaining as significant are the "larger" values.

Table 1. Specified values of the parameters and statistics used in generating synthetic data.

Item	$\sigma^2_{y \cdot x}$ = 8/3	$\sigma^2_{y \cdot x}$ = 2/3	$\sigma^2_{y \cdot x}$ = 1/6	$\sigma^2_{y \cdot x}$ = 1/24
β_0	100.0	100.0	100.0	100.0
β_1	-1.00	-1.00	-1.00	-1.00
β_2	-2.00	-2.00	-2.00	-2.00
β_3	-3.00	-3.00	-3.00	-3.00
β_4	1.00	1.00	1.00	1.00
β_5	2.00	2.00	2.00	2.00
β_6	3.00	3.00	3.00	3.00
E(X_j)	0.50	0.50	0.50	0.50
Var(X_j)	0.083	0.083	0.083	0.083
E(Y)	100.0	100.0	100.0	100.0
Var(Y)	5.00	3.00	2.50	2.375
R^2	0.467	0.778	0.933	0.982

Table 2. Frequency of selection, mean observed values, and specified values of the regression coefficients. Based on 100 samples of various sizes, $\sigma^2_{y \cdot x} = 8/3$.

100 samples of size 20	b_1	b_2	b_3	b_4	b_5	b_6
Frequency of Selection						
Observed	0.11	0.26	0.45	0.07	0.22	0.52
Specified	0.12	0.33	0.62	0.12	0.33	0.62
Mean Values						
Selected	-2.76	-3.51	-3.97	2.26	3.66	4.10
Specified	-1.00	-2.00	-3.00	1.00	2.00	3.00
Standard Deviation						
Selected	2.14	0.85	1.26	1.97	0.89	1.60
Specified	1.30	1.30	1.30	1.30	1.30	1.30

100 samples of size 50	b_1	b_2	b_3	b_4	b_5	b_6
Frequency of Selection						
Observed	0.18	0.64	0.90	0.23	0.56	0.89
Specified	0.23	0.68	0.96	0.23	0.68	0.96
Mean Values						
Selected	-2.08	-2.40	-3.16	1.94	2.60	3.17
Specified	-1.00	-2.00	-3.00	1.00	2.00	3.00
Standard Deviation						
Selected	0.50	0.60	0.78	1.08	0.57	0.81
Specified	0.81	0.81	0.81	0.81	0.81	0.81

100 samples of size 100	b_1	b_2	b_3	b_4	b_5	b_6
Frequency of Selection						
Observed	0.38	0.96	1.00	0.38	0.89	1.00
Specified	0.40	0.93	1.00	0.40	0.93	1.00
Mean Values						
Selected	-1.67	-2.03	-2.99	1.65	2.07	3.05
Specified	-1.00	-2.00	-3.00	1.00	2.00	3.00
Standard Deviation						
Selected	0.41	0.57	0.61	0.33	0.48	0.63
Specified	0.57	0.57	0.57	0.57	0.57	0.57

Table 3. Frequency of selection, mean observed values, and specified values of the regression coefficients. Based on 100 samples of various sizes, $\sigma^2_{y \cdot x} = 2/3$.

100 samples of size 20	b_1	b_2	b_3	b_4	b_5	b_6
Frequency of Selection						
Observed	0.26	0.59	0.82	0.19	0.48	0.75
Specified	0.31	0.86	0.99	0.31	0.86	0.99
Mean Values						
Selected	-1.93	-2.58	-3.03	1.70	2.33	3.23
Specified	-1.00	-2.00	-3.00	1.00	2.00	3.00
Standard Deviation						
Selected	1.16	0.72	0.93	1.35	0.72	0.93
Specified	0.65	0.65	0.65	0.65	0.65	0.65

100 samples of size 50	b_1	b_2	b_3	b_4	b_5	b_6
Frequency of Selection						
Observed	0.64	1.00	1.00	0.67	0.98	1.00
Specified	0.68	0.99	1.00	0.68	0.99	1.00
Mean Values						
Selected	-1.30	-2.02	-3.03	1.33	1.95	3.02
Specified	-1.00	-2.00	-3.00	1.00	2.00	3.00
Standard Deviation						
Selected	0.37	0.46	0.44	0.37	0.42	0.46
Specified	0.40	0.40	0.40	0.40	0.40	0.40

100 samples of size 100	b_1	b_2	b_3	b_4	b_5	b_6
Frequency of Selection						
Observed	0.91	1.00	1.00	0.94	1.00	1.00
Specified	0.98	1.00	1.00	0.98	1.00	1.00
Mean Values						
Selected	-1.06	-2.01	-2.99	1.07	1.98	3.02
Specified	-1.00	-2.00	-3.00	1.00	2.00	3.00
Standard Deviation						
Selected	0.27	0.30	0.28	0.26	0.29	0.29
Specified	0.28	0.28	0.28	0.28	0.28	0.28

Table 4. Frequency of selection, mean observed values, and specified values of the regression coefficients. Based on 100 samples of various sizes, $\sigma^2_{y \cdot x} = 1/6$.

100 samples of size 20	b_1	b_2	b_3	b_4	b_5	b_6
Frequency of Selection						
Observed	0.60	0.90	0.93	0.59	0.85	0.95
Specified	0.86	1.00	1.00	0.86	1.00	1.00
Mean Values						
Selected	-1.29	-2.05	-3.04	1.21	1.97	3.04
Specified	-1.00	-2.00	-3.00	1.00	2.00	3.00
Standard Deviation						
Selected	0.38	0.46	0.47	0.34	0.44	0.54
Specified	0.32	0.32	0.32	0.32	0.32	0.32

100 samples of size 50	b_1	b_2	b_3	b_4	b_5	b_6
Frequency of Selection						
Observed	.99	1.00	1.00	1.00	1.00	1.00
Specified	.998	1.00	1.00	.998	1.00	1.00
Mean Values						
Selected	-1.06	-2.02	-3.01	1.02	1.96	3.02
Specified	-1.00	-2.00	-3.00	1.00	2.00	3.00
Standard Deviation						
Selected	0.22	0.23	0.21	0.23	0.23	0.23
Specified	0.20	0.20	0.20	0.20	0.20	0.20

100 samples of size 100	b_1	b_2	b_3	b_4	b_5	b_6
Frequency of Selection						
Observed	1.00	1.00	1.00	1.00	1.00	1.00
Specified	1.00	1.00	1.00	1.00	1.00	1.00
Mean Values						
Selected	-1.02	-2.01	-3.01	1.02	1.97	3.03
Specified	-1.00	-2.00	-3.00	1.00	2.00	3.00
Standard Deviation						
Selected	0.15	0.16	0.16	0.14	0.14	0.15
Specified	0.14	0.14	0.14	0.14	0.14	0.14

Table 5. Frequency of selection, mean observed values, and specified values of the regression coefficients, Based on 100 samples of various sizes, $\sigma^2_{y \cdot x} = 1/24$.

100 samples of size 20	b_1	b_2	b_3	b_4	b_5	b_6
Frequency of Selection						
Observed	0.97	0.98	0.98	0.96	0.97	0.98
Specified	0.999	1.00	1.00	0.999	1.00	1.00
Mean Values						
Selected	-1.03	-1.99	-3.02	0.99	2.02	2.99
Specified	-1.00	-2.00	-3.00	1.00	2.00	3.00
Standard Deviation						
Selected	0.27	0.21	0.21	0.19	0.20	0.26
Specified	0.16	0.16	0.16	0.16	0.16	0.16

100 samples of size 50	b_1	b_2	b_3	b_4	b_5	b_6
Frequency of Selection						
Observed	1.00	1.00	1.00	1.00	1.00	1.00
Specified	1.00	1.00	1.00	1.00	1.00	1.00
Mean Values						
Selected	-1.00	-2.00	-2.99	0.99	2.00	3.00
Specified	-1.00	-2.00	-3.00	1.00	2.00	3.00
Standard Deviation						
Selected	0.11	0.10	0.10	0.10	0.11	0.12
Specified	0.10	0.10	0.10	0.10	0.10	0.10

100 samples of size 100	b_1	b_2	b_3	b_4	b_5	b_6
Frequency of Selection						
Observed	1.00	1.00	1.00	1.00	1.00	1.00
Specified	1.00	1.00	1.00	1.00	1.00	1.00
Mean Values						
Selected	-1.00	-2.00	-2.99	1.00	2.00	2.99
Specified	-1.00	-2.00	-3.00	1.00	2.00	3.00
Standard Deviation						
Selected	0.08	0.07	0.07	0.07	0.08	0.12
Specified	0.07	0.07	0.07	0.07	0.07	0.07

Table 6. Number of times a given number of variables were included in equation, 100 samples of varying sizes.

Sample Size	$\sigma^2_{y \cdot x}$	Given Number of Variables 0	1	2	3	4	5	6
20	8/3	13	40	26	15	5	1	0
50	8/3	0	3	16	33	34	12	2
100	8/3	0	0	0	9	37	38	16
20	2/3	1	13	16	31	26	11	2
50	2/3	0	0	0	0	15	41	44
100	2/3	0	0	0	0	0	15	85
20	1/6	1	5	3	5	14	32	40
50	1/6	0	0	0	0	0	1	99
100	1/6	0	0	0	0	0	0	100
20	1/24	1	1	0	1	1	0	96
50	1/24	0	0	0	0	0	0	100
100	1/24	0	0	0	0	0	0	100

EXAMPLE

As stated in section 3, the standard assumptions were satisfied in generating a large number of observations on

$$Y = 100.0 - 1.00X_1 - 2.00X_2 - 3.00X_3 - 1.00X_4 + 2.00X_5 + 3.00X_6 + \varepsilon \qquad (4)$$

where ε is normally distributed with zero expectation and variance $\sigma^2_{y \cdot x}$. The data were broken into 4 parts of 10,000 observations each, depending on the variance level of the random component. Each X_j has variance 1/12 and the 4 variances of ε were 8/3, 2/3, 1/6, and 1/24. The experiment consisted of performing 100 regression analyses with samples of size 20, 50, and 100 for each of the 4 variance levels.

The estimated coefficients were biased as described earlier with the bias more pronounced for the smaller sample sizes with higher variances for the random components. For example, the average values of the 6 estimated coefficients for 100 samples of size 20 with $\sigma^2_{y \cdot x} = 8/3$ were:

$$Y = 99.9 - 2.76X_1 - 3.51X_2 - 3.97X_3 + 2.26X_4 + 3.66X_5 + 4.10X_6 \qquad (5)$$

The degree of bias is found by comparing this equation with Equation 4. In contrast, for 100 samples of size 20 with $\sigma^2_{y \cdot x} = 1/24$ the "average" values were:

$$Y = 100.0 - 1.03X_1 - 1.99X_2 - 3.02X_3 + 0.99X_4 + 2.02X_5 - 2.99X_6 \qquad (6)$$

Again, the bias is due to including the larger absolute valued estimates of the coefficients and excluding the smaller absolute valued estimates. As an example, the "average" value of b_1 in Equation 5 is based on only 12 of the largest absolute values while corresponding value in Equation 6 is based on 97 of the largest absolute values of the 100 estimates.

SUMMARY

The stepwise regression procedure can be used to select the "best" regression equation from a specified set of independent variables. However, the selection criteria used in the stepwise procedure produces bias in the selected coefficients.

Data were generated according to a pre-specified multiple linear equation, and regression analyses were performed on subsets of those data. Comparison of the parameters and their estimators allows an evaluation of the stepwise procedure. For small sample sizes and large error levels, few of the derived equations contained all of the six independent variables. Furthermore, mean values of the estimated regression coefficients were biased.

Bias in the estimated coefficients is due to the selection criteria used in the stepwise procedure. Examples using synthetic data from known linear relations show that there can be a high degree of bias in the estimated coefficients.

ACKNOWLEDGMENTS

The authors would like to thank the late Dr. Henry Tucker for his initial interest, for his assistance, and for his advice which made this research possible.

Contribution of the Southwest Watershed Research Center, Agricultural Research Service, USDA, in cooperation with the Arizona Agricultural Experiment Station, Tucson, Arizona.

REFERENCES

1. Draper, N. R. and H. Smith (1966). Applied Regression Analysis. John Wiley and Sons, Inc., New York.

2. Efroymson, M. A. (1960). Mathematical Methods for Digital Computers. Ed. by A. Ralston and H. S. Wile. John Wiley and Sons, New York.

3. Huszar, C. K. (1969). "A Generalized Stepwise Regression Analysis," Mimeographed notes. University Computer Center, University of Arizona, Tucson, Arizona.

4. Pope, P. T. and J. T. Webster (1972). "The Use of an F-Statistic in Stepwise Regression Procedures." Technometrics, 14, pp. 327-340.