Frequency of Selecting Noise Variables in Subset Regression Analysis: A Simulation Study

VIRGINIA F. FLACK and POTTER C. CHANG*

This article presents the results of a simulation study of variable selection in a multiple regression context that evaluates the frequency of selecting noise variables and the bias of the adjusted R^2 of the selected variables when some of the candidate variables are authentic. It is demonstrated that for most samples a large percentage of the selected variables is noise, particularly when the number of candidate variables is large relative to the number of observations. The adjusted R^2 of the selected variables is highly inflated.

KEY WORDS: Variable selection; Exploratory regression analysis.

1. INTRODUCTION

Because of widespread availability of statistical software, automated variable selection in regression analysis is now a simple task. It is often performed when there are a large number of candidate variables and a priori knowledge of their relevance is not clear. Strictly speaking, variables should not be selected solely on the basis of statistical data analysis. They should be viewed as only preliminary suggestions and should not be used for any purpose until they have been confirmed by principles in the area of the subject matter. This basic doctrine of research methodology is well known but is often ignored. Selected variables are often referred to as "significant" variables on statistical grounds without substantive confirmation based on principles in the subject matter.

Rencher and Pun (1980) and Freedman (1983) demonstrated that variables are selected by regression algorithms when data were generated from null models having uncorrelated noise variables. Lovell (1983) illustrated that noise variables also are often selected from data generated from models having one and two authentic variables. In spite of these valuable results, variables are still often selected by subset regression algorithms.

In this article, results of simulation experiments assessing frequencies of selecting noise variables in the presence of authentic variables are presented. The purpose is to illustrate the effects of sample size and the number of candidate variables on the frequency of selecting noise variables.

2. DESIGN OF THE SIMULATION STUDY

The simulation study was designed so that it reflects practical applications of variable selection analysis.

*Virginia F. Flack is Assistant Professor of Biostatistics, and Potter C. Chang is Professor of Biostatistics, Division of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90024.

2.1 Definition of Authentic and Noise Variables

Let Y be a response variable and $X_1, X_2, \ldots, X_p$ be P candidate variables. For this study, we define authentic and noise variables as follows: A candidate variable X_i is an authentic predictor variable if its coefficient β_i in the regression equation $E(Y|\mathbf{X}) = \mathbf{X}'\boldsymbol{\beta}$ is nonzero and X_i is a noise variable otherwise.

2.2 Underlying Distribution

We assume that Y and $X_1, \ldots, X_p$ are random variables and that their joint distribution follows a $(P + 1)$-variate Gaussian distribution with mean $\mathbf{0}$. We have two variables, X_1 and X_2, with $\rho_{YX_1} = \rho_{YX_2} = .5$, where ρ_{YX_i} is the simple correlation coefficient between Y and X_i. For all other candidate variables, $\rho_{YX_i} = .0$. In many areas of application, especially the health and social sciences, the correlations between the response and the candidate variables are usually not large. In this sense, the value $\rho_{YX_1} = \rho_{YX_2} = .5$ is rather optimistic.

We specify the correlation matrix of the P candidate variables by the autocorrelation pattern among its elements: $\rho_{ij} = \rho^{j-i+1}$ for $j > i = 1, \ldots, P - 1$, where ρ_{ij} $(i \neq j)$ is the correlation coefficient between X_i and X_j. Three values, $\rho = 0, .3$, and $.5$, were selected for the autocorrelation, specifying three positive-definite covariance matrices. For these three covariance matrices, the true regression equation can be written as follows:

$$E(Y|\mathbf{X}) = [(1 - \rho)X_1 + (1 - \rho + \rho^2)X_2 - \rho X_3]/2(1 - \rho^2).$$

Thus there are two authentic variables when $\rho = 0$ and three when ρ is nonzero. Let $\rho^2_{Y\mathbf{X}}$ be the true squared multiple correlation coefficient between the response and all P candidate variables. Parameters for the models specified by the three different covariance matrices are shown in Table 1.

2.3 Variable Selection Procedures

Two variable selection procedures were used. The first was the procedure that selects that subset of K variables that maximizes the sample R^2 value (among all subsets of size K). The value of K was prespecified; because we have two authentic variables in this study when $\rho = 0$, we used $K = 2$. The RSQUARE procedure of SAS (SAS Institute Inc. 1985) was used to select the two variables.

The second procedure is stepwise regression (Draper and Smith 1981). The number of variables selected is not prespecified. The SAS STEPWISE procedure with the default significance levels of 15% was used.

Table 1. Parameters for the Three Specified Models

	Regression coefficients				Multiple correlation coefficient,
ρ	β_1	β_2	β_3	β_i $(i > 3)$	ρ^2_{Yx}
0	.50	.50	0	0	.5000
.3	.38	.43	−.16	0	.3725
.5	.33	.50	−.33	0	.3125

2.4 Sample Size (N) and Number of Candidate Variables (P)

In many applications, the number of candidate variables (P) is large and the sample size (N) is small. In the more extreme cases, the number of candidate variables exceeds the sample size. For instance, an example reported by Rencher and Pun (1980) had 54 candidate variables and only 26 observations.

For this study, a factorial arrangement with $N = 10$, 20, and 40 and $P = 10$, 20, and 40 was used. These values were selected because they closely resemble many practical applications. Some of the cases use a larger number of candidate variables than sample size.

2.5 Sampling Procedures

For each combination of N, P, specific covariance structure, and fixed or varying K, 50 samples were generated. This is a large enough number of replications to demonstrate the problems of variable selection in the presence of authentic variables and to allow some patterns to emerge as N and P change.

The RANNOR random number generator in SAS was used to generate standard normal pseudodeviates (SAS Institute Inc. 1985). Standard linear transformations of vectors of standard normal deviates were used to create vectors having the prespecified covariance matrices.

3. RESULTS OF THE SIMULATION STUDY

3.1 Fixed K ($K = 2$)

Results of the simulations when the number of selected variables was fixed are presented in Table 2. Let P_A denote the percent of the repeated samples from which the two selected variables were authentic. For most of the situations examined in the simulation study, P_A was low. The only exceptions occurred when $N = 40$ and the candidate variables were uncorrelated (an unrealistic assumption).

As expected, P_A decreases as the number of candidate variables increases, and increases as the sample size increases. The effect of sample size on P_A seems to be more important than that of the number of candidate variables.

In addition, noise variables were selected more frequently when the correlation among the candidate variables was higher. For the most favorable case of $N = 40$ and $P = 10$, when candidate variables are not correlated, $P_A = 96\%$. An introduction of a weak autocorrelation ($\rho = .3$) among the candidate variables decreased P_A substantially, even though at the same time the number of authentic variables was increased from two to three.

3.2 Varying K

The stepwise simulation results (Table 3) are presented only for the case with the correlation matrix with $\rho = .3$ because it best approximates the intercorrelation among candidate variables in usual applications. Percentages of the samples with 0, 1, 2, or 3 authentic variables selected are given. In addition, median and interquartile ranges are presented for K (the number of variables selected from each sample), P_N (the proportion of selected variables that are noise), and the adjusted R^2 of the selected variables.

As expected, the number of selected variables varied widely. It was limited by N when $N < P$, and its median increased when P was larger. The percentage of samples from which all of the authentic variables were selected was low. Even for the best result (when $N = 40$ and $P = 10$), it was still only 34%.

P_N, the percentage of selected variables that are noise, varied widely. For most models, the median of P_N was at least as large as 50%, indicating that for most of the samples at least half of the selected variables were noise. The median of P_N increased with P, and the rate of increase was sharp even when N was 40.

The adjusted R^2 of the selected variables was highly inflated. The 25th percentiles of the adjusted R^2 were smaller than the model value (.3725) only when $P < N$. Otherwise, the 25th percentiles of the adjusted R^2 were much greater than the model value.

Table 2. Selection Percentage of the Two Authentic Variables (50 repeated samples)

Model	% of simulation samples with:	$N = 10$			$N = 20$			$N = 40$		
		$P = 10$	$P = 20$	$P = 40$	$P = 10$	$P = 20$	$P = 40$	$P = 10$	$P = 20$	$P = 40$
$\rho = 0$	Two authentic	20	14	6	72	58	44	96	90	86
($\rho^2_{Yx} = .5$)	One noise	56	42	30	24	34	38	4	10	14
	All noise	24	44	64	4	8	18	0	0	0
$\rho = .30$	Two authentic	16	8	2	48	28	20	76	68	54
($\rho^2_{Yx} = .3725$)	One noise	58	48	32	46	66	66	24	30	40
	All noise	26	44	66	6	6	14	0	2	6
$\rho = .50$	Two authentic	24	14	4	44	28	22	74	56	50
($\rho^2_{Yx} = .3125$)	One noise	56	46	36	50	62	56	26	44	42
	All noise	20	40	60	6	10	22	0	0	8

NOTE: The model parameters are $\rho_{YX_1} = \rho_{YX_2} = .5$, $\rho_{YX_i} = 0$ $(i = 3, \ldots, p)$, and $\rho_{X_iX_j} = \rho^{j-i+1}$.

Table 3. *Summary of Results When Variables Are Selected by STEPWISE (Varying K, 50 repeated samples)*

	$N = 10$			$N = 20$			$N = 40$		
	$P = 10$	$P = 20$	$P = 40$	$P = 10$	$P = 20$	$P = 40$	$P = 10$	$P = 20$	$P = 40$
% of samples									
Three authentic	4	2	0	14	10	22	34	28	28
Two authentic	24	18	16	40	40	40	50	56	46
One authentic	42	54	48	44	38	36	16	14	26
None authentic	30	26	36	2	12	2	0	2	0
K = No. of variables									
25–75th %	1–3	3–9	8–9	2–4	3–7	10–18	2–4	4–6	7–12
Median	2	6	9	3	4	18	3	5	9
P_N = % Noise									
25–75th %	33–100	78–100	88–100	0–50	50–75	83–90	0–50	50–67	71–83
Median	67	85	89	33	67	89	33	59	78
Adjusted R^2									
25–75th %	.48–.90	.75–1.0	1.0–1.0	.36–.62	.46–.83	.95–1.0	.32–.57	.42–.66	.61–.81
Median	.72	.96	1.0	.48	.71	1.0	.49	.56	.72

NOTE: The model parameters are $\rho_{YX_1} = \rho_{YX_2} = .5$, $\rho_{YX_i} = 0$ $(i = 3, \ldots, P)$, $\rho_{X_iX_{i+j}} = .3^j$, and $\rho^2_{Y\mathbf{X}} = .3725$.

4. SUMMARY

The results of this study further confirm that noise variables are often selected by regression algorithms. The adjusted R^2 of the selected variables is highly inflated. The frequency of selecting noise variables and the biasness of the adjusted R^2 are particularly high when the sample size is not large relative to the number of candidate variables.

Clearly methods for the evaluation of the authenticity of variables selected in regression analysis and for the determination of less biased estimates of $\rho^2_{Y\mathbf{X}}$ are needed. Before such methods are available, variables selected by subset regression algorithms should be examined critically. The variables selected from a given data set should not be used as the basis of any conclusion unless they are firmly supported by considerations in the subject matter and are validated by other data sets. Such confirmation and validation are especially important when the number of candidate variables is large and a priori knowledge about their relationships to the response variables are not clear.

[*Received November 1985. Revised January 1986.*]

REFERENCES

Draper, H., and Smith, H. (1981), *Applied Regression Analysis* (2nd ed.), New York: John Wiley.

Freedman, D. A. (1983), "A Note on Screening Regression Equations," *The American Statistician*, 37, 152–155.

Lovell, M. C. (1983), "Data Mining," *The Review of Economics and Statistics*, 65, 1–12.

Rencher, A. C., and Pun, F. C. (1980), "Inflation of R^2 in Best Subset Regression," *Technometrics*, 22, 49–53.

SAS Institute Inc. (1985), *SAS User's Guide: Statistics, Version 5 Edition*, Cary, NC: Author.