Automated Subset Selection Algorithms and the Selection of Authentic Predictor Variables in the Presence of Noise

by

Shelley Derksen

A thesis
presented to the University of Manitoba
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Interdisciplinary Studies
Computer Science
Statistics and
Psychology

Winnipeg, Manitoba

ISBN 0-315-76677-8

AUTOMATED SUBSET SELECTION ALGORITHMS AND THE SELECTION OF AUTHENTIC PREDICTOR VARIABLES IN THE PRESENCE OF NOISE

BY

SHELLEY DERKSEN

A thesis submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

I hereby declare that I am the sole author of this thesis.

I authorize the University of Manitoba to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Shelley Derksen

I further authorize the University of Manitoba to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Shelley Derksen

ACKNOWLEDGEMENTS

The author would like to thank a number of people whose assistance in the writing of this thesis proved invaluable. First and foremost, I would like to thank Dr. Harvey Keselman for his encouragement and guidance throughout the process. Next, I would like to thank the committee members, Dr. Neil Arnason and Dr. John Brewster for their participation and helpful comments. I would also like to thank my friends and fellow students Ms. Linda Neden, Mr. George McClure, Mr. Bill Kiss, and Ms. Ina Vincent for their help in matters both technical and personal. Finally, I want to thank my husband, Mr. Kane Gin, without whom none of this would have been possible.

Abstract

Flack and Chang (1987) studied the effects of sample size and the number of candidate variables on the frequency that noise variables are selected by the STEPWISE algorithm. Additionally the bias of the adjusted R^2 of the selected variables was examined. They demonstrated that, often, a large percentage of the selected variables are noise, especially when the number of candidate variables exceeds the sample size. Also, the adjusted R^2 of the selected variables is highly inflated. However, these findings may not be relevant to behavioral scientists as the conditions Flack and Chang (1987) investigated did not typify behavioural science phenomena.

The present study used Monte Carlo simulation techniques to investigate the frequency with which authentic and noise variables are selected by subset selection algorithms under conditions characteristic of behavioural science investigations. In particular, the effects of the correlation between predictor variables, the number of candidate predictor variables, the size of the sample, and the level of significance for inclusion and deletion of variables were studied for the three subset algorithms implemented by SAS: STEPWISE, BACKWARD, and FORWARD. The results of this study were shown to largely agree with those of Flack and Chang in that even under favourable parametric conditions a significant portion of the final subset could be noise. It was further found that the trends of the BACKWARD procedure could differ both in magnitude and direction from those of the STEPWISE and FORWARD procedures. The BACKWARD procedure tended to select both more authentic and more noise variables on average than either the STEPWISE or FORWARD procedures. Similarly the BACKWARD procedure produced more inflated values of R^2 and R_k^2 [an estimate of the population coefficient of multiple determination that is

adjusted by the final subset size]. However, for each of the subset selection algorithms under optimal conditions, about half of the available authentic variables were selected on average and the average number of noise variables selected was less than one. Similarly, R^2 and R_k^2 accurately estimated the population multiple coefficient of determination when a conservative inclusion/deletion level was used and the predictor variables were uncorrelated and the sample size was large compared to the number of predictor variables. In addition, the population multiple coefficient of determination was never over-estimated in the correlated case by adopting an estimate that is adjusted by the total number of candidate predictor variables rather than the number of variables in the final model.

Table of Contents

INTRODUCTION 1
Introduction to Multiple Linear Regression 2
Correlation in Multiple Regression. 9
Definition of Effect Size in Multiple Regression. 11
Collinearity. 13
Detection of Collinearity. 19
Responding to Collinear Data. 22
'Best' Subset Selection Algorithms. 23
Forward Selection Method. 23
Backward Elimination Method. 23
Stepwise Method. 24
Problems Associated with Subset Selection Algorithms. 24
Inflation of R^2 in 'Best' Subset Selection. 25
Controlling Stepwise Algorithms. 27
Stopping Rules in Subset Selection Algorithms. 28
Collinearity and 'Best' Subset Selection Algorithms. 30
METHOD. 34
Design. 34
Data Generation. 34
Statistics 42
RESULTS. 44
Selection of Authentic Predictor Variables in the Presence of Noise . 44
Trend Analysis of $\rho_{x_i x_j}$, P and N within Method and
Inclusion/Deletion Level. 49
Main Effects. 49
Collinearity. 49
Number of Candidate Variables. 50
Sample Size. 57
Two-way Interactions 58
Three-way Interactions. 60

Table of Contents

Estimates of the Population Coefficient of Multiple Determination. . . . 61
Trend Analysis of $\rho_{x_i x_j}$, P and N within Algorithm and
Inclusion/Deletion Level. 65
Main Effects. 65
Collinearity. 65
Number of Candidate Variables. 71
Sample Size. 71
Two-way Interactions. 72
Three-way Interactions. 75
DISCUSSION AND CONCLUSIONS. 76
References. 84
Appendix A
Appendix B

List of Tables

1. Schematic for a Multiple Linear Regression Model. 3
2. ANOVA for Multiple Regression. 8
3. Effect of Increasing Correlation Among Independent Variables . . .18
4. Data Generation Parameters. 37
5. Level of Significance Conditions. 41
6. Mean Frequency of Authentic and Noise Variables and % Noise Contained in the Final Subset in the STEPWISE Procedure. . . .45
7. Mean Frequency of Authentic and Noise Variables and % Noise Contained in the Final Subset in the BACKWARD Procedure. . . .46
8. Mean Frequency of Authentic and Noise Variables and % Noise Contained in the Final Subset in the FORWARD Procedure. . . .47
9. Proportion of Model Sum of Squares Accounted for by Contrast Dependent Variable:C_A 51
10. Effect of Collinearity, Number of Candidate Variables and Sample Size on the Mean Values of C_A 52
11. Proportion of Model Sum of Squares Accounted for by Contrast Dependent Variable:C_N 53
12. Effect of Collinearity, Number of Candidate Variables and Sample Size on the Mean Values of C_N 54
13. Proportion of Model Sum of Squares Accounted for by Contrast Dependent Variable:P_N 55
14. Effect of Collinearity, Number of Candidate Variables and Sample Size on the Mean Values of P_N 56
15. Effect of Collinearity, Number of Candidate Variables and Sample Size on the Mean Values of R^2 62
16. Effect of Collinearity, Number of Candidate Variables and Sample Size on the Mean Values of R_k^2 63
17. Effect of Collinearity, Number of Candidate Variables and Sample Size on the Mean Values of R_p^2 64
18. Proportion of Model Sum of Squares Accounted for by Contrast Dependent Variable:R^2 66
19. Proportion of Model Sum of Squares Accounted for by Contrast Dependent Variable:R_k^2 67

List of Tables

20. Proportion of Model Sum of Squares Accounted for by Contrast Dependent Variable:R_p^2 68
21. Effect of Method and Inclusion/Deletion Level on the Mean Values of C_A, C_N, and P_N 77
22. A Comparison of Flack and Chang's Results to Those of the Present Study 80

List of Figures

1. The Effect of Collinearity on Estimation 15
2. The Effect of Sample Size, Number of Candidate Variables, and Collinearity on the Mean Value of Two Measures of the Population Squared Coefficient of Determination within the STEPWISE Procedure. 69

INTRODUCTION

A common goal of behavioural and social scientists is to quantify relationships between a response variable and one or more predictor variables using multiple linear regression analysis. Therefore, it is important to researchers that the relevant predictor variables of the response variable be known. When there are many candidate predictors to choose from and prior knowledge does not dictate their relevance, a researcher may use automated 'best' subset selection algorithms to choose the 'best' predictor variables from the larger set. An examination of some textbooks on multiple linear regression (e.g. Cohen & Cohen, 1983, pp. 123-125; Neter, Wasserman & Kutner, 1983, pp. 417-443; Pedhazur, 1982, pp. 150-171; Younger, 1985, pp. 488-489) indicates that the search algorithms most commonly used are the forward, backward elimination and stepwise algorithms [these algorithms often are referred to collectively as stepwise methods, see SAS' (1985) STEPWISE PROCEDURE for example]. By using these algorithms, it is hoped that the most effective predictors which adequately explain the behaviour of the response variable may be found.

Recently, Flack and Chang (1987) compared the all-subsets and stepwise algorithms for the frequency with which they specified 'best' models containing authentic versus noise predictor variables. For the parametric conditions they investigated, it was found that both algorithms typically selected a large percentage of noise variables. Additionally, they found that the 25th percentile of their adjusted estimate of the multiple coefficient of determination often exceeded the model value. These findings are most interesting but may have limited generalizablity to behavioural scientists since the conditions Flack and Chang (1987) investigated did not typify behavioural science phenomena.

Consequently, the goal of this study was to extend the research on the selection of predictor variables and to compare 'best' subset selection algorithms under parametric conditions more characteristic of behavioural science investigations. Moreover, subset algorithms, levels of significance for inclusion and deletion of variables, and an estimate of the population multiple coefficient of determination not examined by Flack and Chang (1987) were investigated.

Introduction to Multiple Linear Regression

Multiple linear regression theory holds that given a response variable, y, and k predictor variables, $X_1, X_2, \ldots, X_k$, there is a linear relationship between the response variable and the predictor variables (see Table 1 for schematic). The general multiple regression model statement expressing the relationship between the response variable and the predictor variables is given by

$$y_i = \beta_0 + \beta_1 X_{i\,1} + \beta_2 X_{i\,2} + , \ldots , + \beta_k X_{i\,k} + \varepsilon_i \qquad (1)$$

where y_i is the response variable; $X_{i1}, X_{i2}, \ldots, X_{ik}$ are the ith observations on the k predictor variables, measured without error; β_0 is the y intercept; $\beta_1, \beta_2, \ldots, \beta_k$ are the k regression parameter constants where β_j (j = 1, ... , k) measures the change in y per unit change in X_j when all other predictor variables are held constant; and ε_i is the random error term of the ith observation.

Given a set of values for $X_1, X_2, \ldots, X_k$ and for $\beta_1, \beta_2, \ldots, \beta_k$ a researcher could estimate y (with some error represented by ε). If a particular predictor variable, X_j, was unrelated to the response variable, that is if a unit change in X_j produced no corresponding change in y, then its corresponding regression coefficient, β_j, would be zero.

Under the multiple linear regression model it is assumed that the errors are independent, identically distributed random variables with mean zero and common variance. Notationally these assumptions are expressed in the following way:

$$E(\varepsilon_i) = 0$$
$$Var(\varepsilon_i) = \sigma^2$$
$$Cov(\varepsilon_i , \varepsilon_j) = 0 \qquad (i \neq j)$$

where σ^2 is the common population variance of the error term. Thus, the expected value of y_i, known as the regression function, is given by

$$E(y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} \qquad (2)$$

Table 1

Schematic for a Multiple Linear Regression Model [a]

Y_1	(X_{11}	X_{12}	X_{13}	...	X_{1k})
Y_2	(X_{21}	X_{22}	X_{23}	...	X_{2k})
Y_3	(X_{31}	X_{32}	X_{33}	...	X_{3k})
.					
.					
.					
Y_N	(X_{N1}	X_{N2}	X_{N3}	...	X_{Nk})

[a] Note: X_{i1} , ... , X_{ik} are the ith observations on the k independent variables, Y_i is the dependent variable.

and the variance of y_i is given by

$$Var(y_i | X_1, X_2, \ldots, X_k) = Var(\varepsilon_i) = \sigma^2.$$

For a sample of N observations the model and associated assumptions may be expressed as in the following matrix equations. The model is given by:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & & X_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{N1} & X_{N2} & \cdots & X_{Nk} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

or

$$\mathbf{y} = \mathbf{X} \quad \boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

where **y** is a N X 1 vector of observations on the response variable; **X** is a N X k+1 matrix of observations on the k predictor variables; **β** is a k+1 X 1 vector of regression coefficients; and **ε** is a N X 1 vector of random errors.

The associated assumptions are

$$E(\varepsilon) = \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ . \\ 0 \end{bmatrix} = \mathbf{0}$$

and

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \begin{bmatrix} \mathrm{Var}(\varepsilon_1) & \mathrm{Cov}(\varepsilon_1, \varepsilon_2) & \cdots & \mathrm{Cov}(\varepsilon_1, \varepsilon_N) \\ \mathrm{Cov}(\varepsilon_2, \varepsilon_1) & \mathrm{Var}(\varepsilon_2) & \cdots & \mathrm{Cov}(\varepsilon_2, \varepsilon_N) \\ . & . & . & . \\ . & . & & . \\ . & . & & . \\ \mathrm{Cov}(\varepsilon_N, \varepsilon_1) & \mathrm{Cov}(\varepsilon_N, \varepsilon_2) & \cdots & \mathrm{Var}(\varepsilon_N) \end{bmatrix} = \sigma^2 \mathbf{I}_N .$$

where I_N is an N X N identity matrix, since $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$. The regression function expressed in matrix notation is

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}. \tag{4}$$

When a random sample of size N is taken from the population of experimental units, one can only estimate the population regression coefficients. The method of least-squares provides estimates having certain desirable properties. If **b** denotes the vector of estimated regression coefficients then the regression equation may be written as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

where **e** is the N X 1 vector of residuals. The method of least-squares minimizes the residual sum of squares,

$$\mathbf{e}^T\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b})$$

(where ^{T} is the transpose operator) to yield the normal equations,

$$\mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X})\,\mathbf{b}.$$

If $\mathbf{X}^T\mathbf{X}$ is nonsingular, a unique solution to the normal equations is given by

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y}. \tag{5}$$

This equation provides the least-squares estimates of the population regression coefficients. The requirement that $\mathbf{X}^T\mathbf{X}$ be nonsingular implies that no predictor variable can be a linear function of the others.

According to the Gauss-Markov Theorem, (See Fox, 1984, p. 42) the least-squares estimates are unbiased and have minimum variance among the class of all linear unbiased estimators. Thus

$$E(\mathbf{b}) = \boldsymbol{\beta}. \tag{6}$$

The variances of the regression coefficients are the diagonal elements of

$$\sigma^2 \ (\mathbf{X}^T\mathbf{X})^{-1}. \tag{7}$$

The off-diagonal elements are the covariances between the regression coefficients. Thus, $\sigma^2 \ (\mathbf{X}^T\mathbf{X})^{-1}$ is known as the variance-covariance matrix. Using the estimated regression coefficients, the estimated regression function is expressed as

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} \tag{8}$$

where $\hat{\mathbf{y}}$ is an N X 1 vector of predicted values.

If the response variable and the predictor variables are standardized to have zero intercept and unit length by the correlation transformation,

$$y_i = \frac{(y_i - \bar{y})}{\sqrt{S_{yy}}}$$

and

$$X_{ij} = \frac{(X_{ij} - \bar{X}_j)}{\sqrt{S_{jj}}}$$

where S_{yy} and S_{jj} are the corrected sample sums of squares and $\bar{y}$ and $\bar{X}_j$ are the sample means of y and X_j respectively, then $\mathbf{X}^T\mathbf{X}$ is a (k X k) matrix of correlations among the predictor variables known as the correlation matrix, $\mathbf{R}_{xx}$, and $\mathbf{X}^T\mathbf{y}$ is a (k X 1) vector of correlations between the response variable and

the predictor variables. When the predictor variables are uncorrelated, $\mathbf{R}_{xx}$ is an identity matrix and each estimated regression coefficient will be equal to the simple correlation between the predictor variable with which it is associated and the response variable (Pedhazur, 1973, p. 234). Whether the predictor variables are uncorrelated or not, the standardized regression coefficients may be obtained from the unstandardized coefficients from the relations

$$b^*_0 = 0$$

$$b^*_j = b_j \frac{S_j}{S_y}$$

where S_y and S_j are the standard deviations of y and X_j before they had been transformed. From equation 7, it is clear that the variance covariance matrix of the standardized coefficients is

$$\sigma^{*2} \mathbf{R}_{XX}^{-1}$$

where $\sigma^{*2} = \sigma^2 / S_{yy}$, is the variance of the residuals of the transformed model.

Conducting statistical tests in regression analysis requires the further assumption that the random error term of the model be normally distributed [with mean zero and variance σ^2]. The regression model with this added assumption is known as the normal error model. The addition of the normal error assumption implies that the response variable, the estimated regression coefficients, the predicted values and the residuals are also normally distributed, thus allowing the usual analysis of variance (ANOVA) of the regression model to proceed (See Table 2). The ANOVA tests the global null hypothesis that all the regression coefficients are zero, $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$. Under the null hypothesis, F = MSR / MSE is distributed as an F statistic with k and N - k - 1 degrees of freedom.

Table 2

ANOVA for Multiple Regression

Source	df	Sum of Squares	Mean Square	F
Regression	k	$\mathbf{b}^T \mathbf{X}^T\mathbf{X}\, \mathbf{b} - \frac{(\mathbf{y}^T\mathbf{1}\mathbf{1}^T\mathbf{y})}{N}$	$MSR = \frac{SSR}{k}$	$\frac{MSR}{MSE}$
Error	N - (k+1)	$(\mathbf{y} - \mathbf{Xb})^T(\mathbf{y} - \mathbf{Xb})$	$MSE = \frac{SSE}{N - (k+1)}$	
Total	N - 1	$\mathbf{y}^T\mathbf{y} - \frac{(\mathbf{y}^T\mathbf{1}\mathbf{1}^T\mathbf{y})}{N}$		

[a]NOTE: **1** is a N X 1 vector of 1's

Correlation in Multiple Regression

Measures of correlation are important in multiple regression because, unlike the unstandardized regression coefficients for example, they supply a unitless measure of the strength of a relationship. In multiple regression several such measures may be calculated.

The multiple coefficient of determination (R^2) can be calculated from Table 2 as

$$R^2 = \frac{SSR}{SST} . \tag{9}$$

R^2 may be thought of as the squared correlation coefficient of y with $\hat{y}$ or as the proportion of the total variation that is explained by the model. In simple linear regression, where only one predictor variable is considered, the positive or negative square root of R^2 is the simple correlation between the response variable, y, and the single predictor variable, X. In the multiple regression case, correlation among the different variables is more complex.

The simple linear correlation between X_j and y can be calculated as follows

$$r_{ij} = \frac{S_{jy}}{\sqrt{S_{jj}\, S_{yy}}} \tag{10}$$

where $S_{jy} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(y_i - \bar{y})$, and S_{jj} and S_{yy} are as previously defined.

The simple linear correlation between X_j and y measures the strength of the relationship between X_j and y ignoring the effect of all other predictor variables. Because multiple regression involves more than one predictor variable two other measures of correlation of a single predictor with the response variable may be calculated. Cohen and Cohen (1983) describe these as the semi-partial correlation coefficient (sr_i) and partial correlation coefficient (pr_i) (pp. 88-91).

The semi-partial correlation coefficient measures the correlation of y with X_i where the effect of all of the other predictor variables has been partialled out from X_i. The semi-partial coefficient between y and X_i is calculated as

$$sr_i = \frac{\mathbf{R}^{-1}(i,k+1)}{[\mathbf{R}^{-1}(k+1,k+1)]^{1/2}}$$

where $\mathbf{R}^{-1}$ is defined to be the inverse of the matrix of simple correlations among the set of k + 1 variables: $X_1, X_2, \ldots, X_k$, y, and $\mathbf{R}^{-1}(p,q)$ is the element in the pth row and qth column of that matrix. The squared semi-partial coefficient represents the proportion of variance in y uniquely associated with X_i and may be more simply calculated as

$$sr_i^2 = \frac{SSR_k - SSR_{k-i}}{TSS} = R_k^2 - R_{k-i}^2 \qquad (11)$$

where SSR_k and SSR_{k-i} and R_k^2 and R_{k-i}^2 are the sums of squares for regression and the coefficients of determination for the k - term model and the k minus variable i - term model, respectively, and TSS is the total sums of squares. From this formula it is clear that sr_i^2 is the unique contribution of X_i to the coefficient of determination of the full model.

The partial correlation coefficient is the correlation between that portion of X_i not linearly associated with the remaining k - 1 variables with that portion of y not linearly associated with the remaining k - 1 variables. The partial correlation between y and X_i correcting for $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k$ can be calculated by

$$pr_i = \frac{\mathbf{R}^{-1}(i,k+1)}{[\mathbf{R}^{-1}(k+1,k+1)\ \mathbf{R}^{-1}(i,i)]^{1/2}}$$

where $\mathbf{R}^{-1}$ is as defined previously.

The squared partial correlation coefficient represents the proportion of y variance not accounted for by the k - 1 remaining predictor variables that is accounted for by X_i. Like the sr_i^2, pr_i^2 may be more simply calculated as

$$pr_i^2 = \frac{SSR_k - SSR_{k-i}}{SSE_{k-i}}$$

where SSE_{k-i} is the sums of squares for error in the k - variable i - term model.

If the predictor variables are uncorrelated, the semi-partial correlation coefficient is just the simple correlation coefficient. However, when the predictor variables are correlated, the simple correlation coefficient, the semi-partial and the partial correlation coefficients each measure a unique kind of correlation between the response variable and a given predictor variable.

Definition of Effect Size in Multiple Regression

Cohen (1969), defined " 'effect size' to mean 'the degree to which the phenomenon is present in the population,' or 'the degree to which the null hypothesis is false.' " (p. 9). Since measures of correlation provide a unitless measure of a relationships strength, they are a natural choice for describing the effect size in multiple regression.

Cohen and Cohen (1983) give three general guidelines for determining the value of the population effect size under study:

> 1. Previous experience or study may indicate the effect size to expect in the population.
> 2. Some minimum effect size may be advanced that would have either practical or theoretical significance.
> 3. It may be possible to use some conventional definitions of a small, medium or large effect size for the phenomenon under study. (pp. 59-60)

The size of the phenomenon present in the population is important for it will determine, for a given sample size and significance level, the probability that a false null hypothesis is rejected. Therefore, Cohen and Cohen (1983) describe a method for using the expected effect size to determine the sample

size necessary to reject a false null hypothesis (with some given probability) (pp. 116-118).

In multiple regression, two different kinds of null hypotheses may be of interest. Firstly, the overall F statistic given in Table 2 tests the null hypothesis, $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$. When the overall significance of the model is the hypothesis of interest Cohen and Cohen (1983) define the effect size as a function of the expected population coefficient of multiple determination

$$f^2 = \frac{R^2}{1 - R^2} . \qquad (12)$$

As described, researchers may posit a value of R^2 to determine the expected effect size. If no value from previous research is available Cohen and Cohen (1983, p. 161) offer the following conventional values of f^2: "small", $f^2 = 0.02$; "medium", $f^2 = 0.15$; and "large", $f^2 = 0.35$. Knowing f^2, α, and k, the number of independent variables, the required sample size is calculated as

$$n = \frac{L}{f^2} + k + 1 \qquad (13)$$

where L is a tabled value dependent upon k_B, the number of degrees of freedom associated with the source of y variation being tested [generally equal to k] and the required degree of power (Cohen & Cohen, 1983, p. 117).

The second kind of hypothesis of interest in multiple regression may be the null hypothesis that any partial correlation or regression coefficient for a given X_i (among k independent variables) is zero. When this is of interest, f^2 is defined as

$$f^2 = \frac{sr_i^2}{1 - R^2} . \qquad (14)$$

As before, researchers may posit values for sr_i^2 and R^2 or use Cohen and Cohen's (1983, p. 61) conventional values of f^2. When the value of α and the desired power are set, L can be determined from the tables (k_B is set to 1 since the source of variation is a single X_i) and the sample size may be calculated from equation 13 as described above.

Collinearity

Recall from equation 5 that the least-squares solution to the normal equations requires that the $\mathbf{R}_{xx}$ matrix be invertible. When linear dependencies exist among the independent variables, the determinant of the correlation matrix is zero. The correlation matrix is then deemed singular, and its inverse does not exist. Thus, when linear dependencies exist among the predictor variables, the regression coefficients cannot be estimated by least-squares techniques.

However, the correlation matrix need only approach singularity in order for the regression analysis to be affected.

Harmful collinearity may be introduced in several ways. Gordon (1968) and Pedhazur (1982, p. 242) point out the common practice of using multiple or repetitive measures of the same variable in regression analysis. These variables will be highly correlated, thus introducing collinearity into the correlation matrix. Such highly correlated independent variables are called redundant since they each supply essentially the same information to the model.

Simple pairwise collinearities and more complex near linear dependencies involving three or more independent variables can be introduced into the data either because such a relationship exists in the population or because the relationship was created through sampling error. Gunst and Mason (1977) use the simple example of a sample of spinal cord injury cases containing only females under 30 years of age and males over 30 years of age. Sex and age would then be correlated in the sample even though it is clear that in the population of spinal cord injuries there are young and old patients of both sexes. Though the example illustrates only a pairwise collinearity detectable by examination of pairwise correlations, more complex near linear dependencies may also exist in the sample. Such dependencies are not necessarily detectable by examination of pairwise correlations.

The presence of severe collinearity in the correlation matrix is damaging to least-squares analysis. When the independent variables are highly or perfectly correlated, the regression plane becomes unstable. Figure 1(a) shows the case where two independent variables, X_1 and X_2, are only slightly correlated. The regression plane is well supported by a broad scattering of

points defining the plane. Since the residuals are small, the estimates of the regression parameters are precise. In Figure 1(b), however, the correlation between X_1 and X_2 is perfect and the observations now fall on a straight line in the X_1 X_2 plane. Any number of planes will pass through the line because the numerical solution is undefined. Lastly, consider Figure 1(c) where X_1 and X_2 are highly correlated though imperfectly. The spread in the points on the regression plane is narrow, ill-defining the plane. In such a case the plane lacks support, and is thus fitted very poorly. A slight shift in any of the data points will drastically alter the plane. This is illustrated by the large variability of the estimated regression coefficients.

Recall that the variances of the regression coefficients are calculated as the diagonal elements of equation 7. In the two predictor case, when X_1 and X_2 are transformed by the correlation transformation, the variance-covariance matrix is

$$\sigma^{*2}\begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}^2}{1-r_{12}^2} \\ \frac{-r_{12}^2}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix}.$$

Thus $V(b_1) = V(b_2) = \sigma^{*2}\,[1/(1 - r_{12}^2)]$, where r_{12}^2 is the squared correlation between X_1 and X_2. As the absolute value of the correlation between X_1 and X_2 approaches 1, the term $[1/(1 - r_{12}^2)]$, known as the variance inflation factor (VIF), will approach infinity. Thus, high correlation among the predictor variables inflates the variance of the regression coefficients (See Gordon, 1968; Keselman, 1988; Rockwell, 1975). This in turn leads to imprecise estimation of the regression coefficients and hence, irreproducible results, since a small fluctuation in the correlations (due to sampling or random errors) can lead to large fluctuations in the estimated regression coefficients when collinearity is present (Pedhazur, 1982, p. 235).

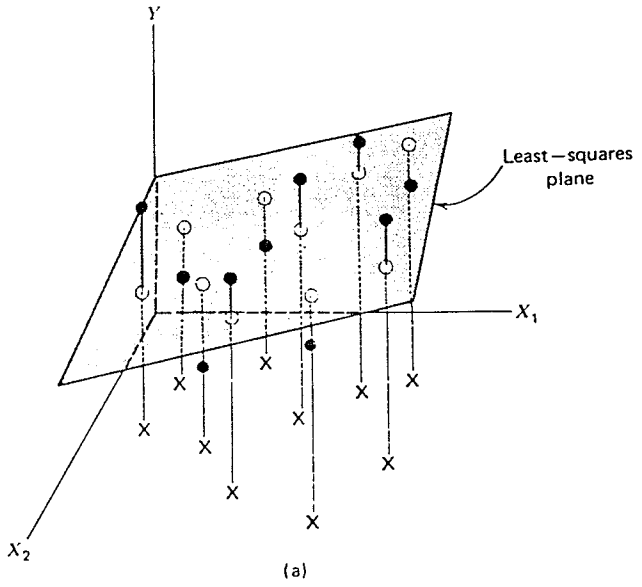

Figure 1 The effect of collinearity on estimation. (a) Small correlation between X_1 and X_2: regression plane well supported. (b) Perfect correlation between X_1 and X_2: regression plane not uniquely defined. (c) Strong correlation between X_1 and X_2: regression plane defined but not well supported.

Note: From Linear statistical models and related methods (pp. 139-140) by J. Fox, 1984, New York: John Wiley. Copyright by John Wiley. Reprinted by permission.

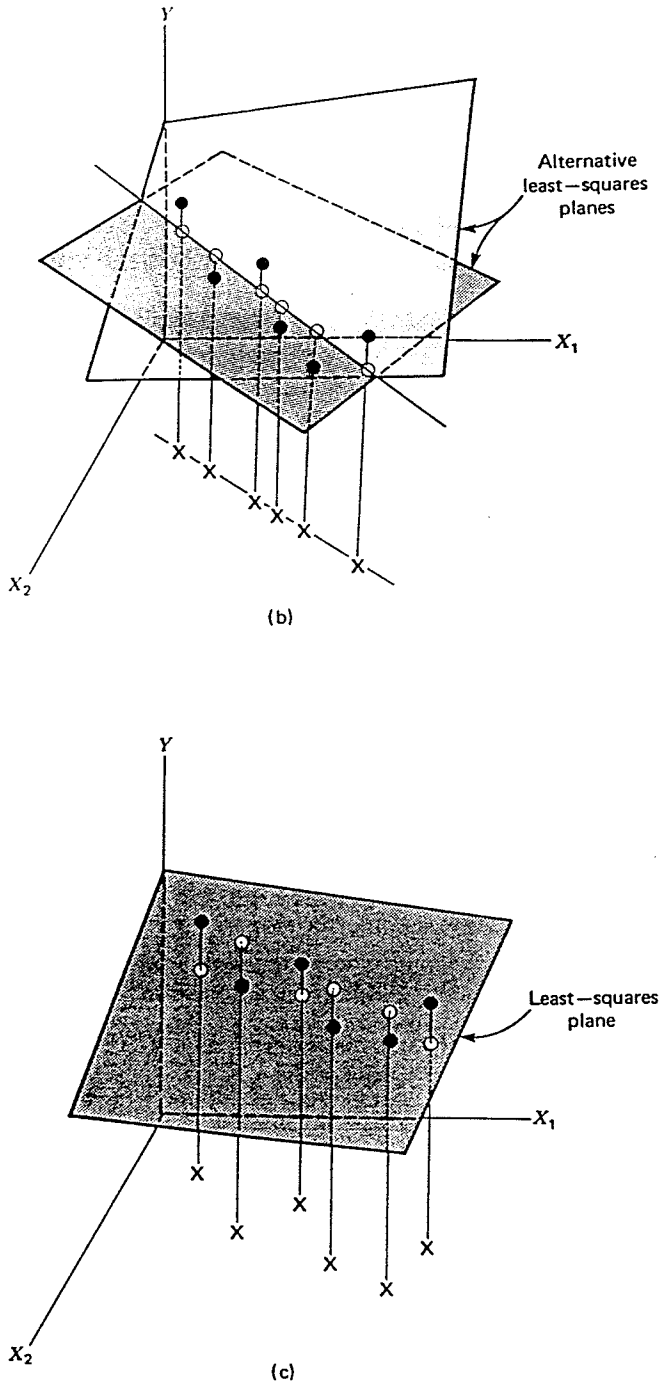

Figure 1 (continued).

In addition to the increase of the variances of the regression coefficients, collinearity also reduces the magnitude of the regression coefficients. Recall from equation 5 that the regression coefficients are calculated as

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y}.$$

In the two predictor case where the independent variables and the dependent variable have been transformed by the correlation transformation, the regression coefficients are calculated as follows:

$$\mathbf{b}^* = \begin{bmatrix} \frac{1}{1 - r_{12}^2} & \frac{-r_{12}^2}{1 - r_{12}^2} \\ \frac{-r_{12}^2}{1 - r_{12}^2} & \frac{1}{1 - r_{12}^2} \end{bmatrix} \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

where r_{12} is the correlation between X_1 and X_2 and r_{1y} and r_{2y} are the correlations between X_1 and y and X_2 and y, respectively. Therefore

$$b^*_1 = \frac{r_{1y} - r_{12}\, r_{2y}}{1 - r_{12}^2} \quad \text{and}$$

$$b^*_2 = \frac{r_{2y} - r_{12}\, r_{1y}}{1 - r_{12}^2} \,. \tag{15}$$

The values that r_{12} may take on are constrained by the values of r_{1y} and r_{2y}. Specifically, the mathematically possible upper and lower bounds for r_{12} are given by (Cohen & Cohen, 1983, p. 280)

$$r_{1y}\, r_{2y} \pm (1 - r_{1y}^2)\,(1 - r_{2y}^2)\,.$$

Table 3 examines what happens to the regression coefficients, and the VIF as the correlation between X_1 and X_2 approaches its maximum (positive) limit.

Table 3 [a]

Effect of Increasing Correlation Among Independent Variables

r_{12}	VIF	b_1	b_2
0.00	1.00	0.30	0.40
0.05	1.00	0.28	0.39
0.10	1.01	0.26	0.37
0.15	1.02	0.25	0.36
0.20	1.04	0.23	0.35
0.25	1.07	0.21	0.35
0.30	1.10	0.20	0.34
0.35	1.14	0.18	0.34
0.40	1.19	0.17	0.33
0.45	1.25	0.15	0.33
0.50	1.33	0.13	0.33
0.55	1.43	0.11	0.34
0.60	1.56	0.09	0.34
0.65	1.73	0.07	0.35
0.70	1.96	0.04	0.37
0.75	2.29	0.00	0.40
0.80	2.78	0.06	0.44
0.85	3.60	-0.14	0.52

[a] Note: The model parameters are $r_{1y} = 0.3$, and $r_{2y} = 0.4$.

Clearly, as collinearity becomes more severe, the estimated regression coefficients are reduced in magnitude, and the VIF is increased. As a consequence of this, the variances of regression coefficients would be increased and, the t statistics testing the hypothesis $H_0: \beta_j = 0$, would become less and less significant. Gordon (1968) empirically illustrated this effect of redundancy and further showed that the problem is compounded as the repetitiveness, or the number of correlated variables, is increased.

Some researchers in the behavioural sciences use the regression coefficients (or standardized beta weights) as a measure of the relative strengths of their associated predictors. However, it is clear from the above that high collinearity adversely affects the magnitudes and standard errors of the regression coefficients and hence their tests of significance and confidence intervals (Pedhazur, 1982, p. 235). Thus the regression coefficients obtained when collinearity is present may be unreliable measures of the predictors influence on the response.[1] Because of these consequences of collinearity, considerable effort has gone into the detection of collinearity.

Detection of Collinearity

Recall that when the determinant of the $(\mathbf{X}^T\mathbf{X})$ matrix is zero, an exact linear dependency exists in the data matrix. Normally one does not encounter exact linear dependencies in nonexperimental data, but only near dependencies of greater or lesser magnitude. The more severe the dependency the closer the determinant will be to zero. Thus, the determinant of $(\mathbf{X}^T\mathbf{X})$ has become a natural indicator of the severity of collinearity. Farrar and Glauber (1967) and Rockwell (1975) both approached the detection of collinearity in this manner using a chi- squared test to determine whether the determinant of the correlation matrix differs significantly from zero. However, the validity of these tests has been questioned (see Kumar, 1975). In addition, no test of the determinant will reveal where the linear dependencies lie, thus the determinant is only of limited usefulness.

---

[1]When the regression equation is used for purposes of prediction, collinearity among the predictor variables is not a problem. In this case, it is the accuracy of the predictions of the model that is important and not the standard error of the regression coefficients.

A more fruitful approach to the detection of collinearity is via the eigenvalues (λ_i) of the ($\mathbf{X}^T\mathbf{X}$) matrix. Recall that a singular matrix is one with at least one linear dependency. Such a matrix will also have one or more zero eigenvalues (See Chatterjee & Price, 1977, p. 162; Tatsuoka, 1971). In particular, it has been recommended that one compute

$$tr(\mathbf{X}^T\mathbf{X})^{-1}$$

(see Hoerl, Schuenemeyer & Hoerl, 1986), where tr is the trace operator. This is equivalent to the sum of the reciprocals of the eigenvalues as recommended by Chatterjee and Price (1977, p. 200). As the eigenvalues of the correlation matrix approach zero indicating increasingly severe collinearity, this number will increase in value. Various values have been advocated as indicating severe collinearity (See Chatterjee & Price, 1977, p. 200; Hoerl et al, 1986), however, like $|\mathbf{X}^T\mathbf{X}|$, this method can only indicate that collinearity is present. That is, it cannot pin-point the number and location of the linear dependencies in the data.

In an effort to find the location of collinearity in the data, some authors (See Cronbach, 1987; Kendall, 1957; Silvey, 1969) have examined the number of small eigenvalues in the data. This method presents a problem in that it is not clear how to judge whether an eigenvalue is too small. Belsley, Kuh and Welsch (1980, pp. 104-105) show that even well-conditioned data matrices may have arbitrarily small eigenvalues.

Belsley, Kuh, and Welsh (1980, pp. 112-113) have therefore come up with a comprehensive strategy for diagnosing collinearity. Their eigenvalue - eigenvector analysis can be used to (1) determine when least-squares analysis is severely degraded by collinearity, (2) identify the number of dependencies in the data, and the variables involved in them and, (3) identify which regression coefficients are affected by the collinearities. Largely for reasons of computational accuracy, Belsley, Kuh and Welsch (1980) base their analysis on the singular-value decomposition of the regressor matrix. Fox (1984, pp. 147-149) employs an equivalent technique based on the eigenvalues and principal components of $\mathbf{R}_{XX}$ where it is assumed that the response and predictor variables have been standardized and the eigenvectors of $\mathbf{R}_{XX}$ have been normalized. For the sake of simplicity, the analysis of Belsley, Kuh and Welsch (1980) is presented here based on Fox's (1984) assumptions.

Assuming that the response and predictor variables have been standardized by subtracting their mean and dividing each by their standard deviation, the variance of b_j is given by the jth diagonal value of the variance - covariance matrix:

$$V(b_j) = \frac{\sigma^2}{n-1} \mathbf{R}_{XX}^{-1}(j,j).$$

It can be shown that the diagonal elements of the inverse of the correlation matrix ($\mathbf{R}_{XX}^{-1}$) are equal to

$$VIF_j = \frac{1}{1 - R_j^2} \qquad j = 1, \ldots, k$$

where VIF_j is the variance inflation factor of the jth independent variable and R_j^2 is the squared multiple correlation coefficient of the jth independent variable with the remaining k-1 independent variables. Because the VIF_j of a particular variable will increase rapidly as the squared multiple correlation of X_j with the remaining variables approaches one (indicating a perfect linear dependency), it can be used to identify which regression coefficients are affected by the collinearities. That is, a large VIF_j value indicates that b_j is adversely affected by collinearity.

To determine the number of dependencies present in the data, Belsley, Kuh and Welsch (1980) and Fox (1984) define the condition index, η_j:

$$\eta_j = \sqrt{\frac{\lambda_L}{\lambda_j}} \qquad j = 1, 2, \ldots, k \qquad (16)$$

where λ_L is the largest eigenvalue and λ_j is the jth eigenvalue of $\mathbf{R}_{XX}$. The number of large values of η_j (>30) will identify the number of dependencies present in the data. To indentify which predictor variables are involved in harmful collinear relations Belsley, Kuh and Welsch (1980) suggest examining each principal components contribution to the variance inflation factor of each regression coefficient. The proportional contribution of the mth principal component to the variance inflation factor of b_j is given by

$$P_{jm} = \frac{\left(\frac{A_{jm}^2}{\lambda_m}\right)}{VIF_j} \qquad (17)$$

where A_{jm} is the jth coefficient of the mth principal component and P_{jm} is known as the variance decomposition proportion of the jth variable on the mth principal component. According to Belsley et al (1980) " a high proportion of the variance of two or more coefficients concentrated in components associated with the same small ... eigenvalue ... is evidence that the corresponding near dependency is causing problems." (Belsley, Kuh & Welsch, 1980, p.106). Thus a large value of P_{jm} (>0.5) associated with a large condition index, η_m (>30), indicates that the data is seriously ill-conditioned due to the dependency represented by the mth eigenvector, and that the jth predictor is involved in that dependency.

Responding to Collinear Data

Once data has been diagnosed as being severely collinear, it is not clear how to respond. Many procedures have been suggested. They include:

(1) achieving a well conditioned matrix of predictor scores (Farrar & Glauber, 1967).
(2) biased estimation techniques (Chatterjee & Price, 1977, ch. 8; Hoerl & Kennard, 1970).
(3) 'best' subset selection algorithms (Hoerl, Schuenemeyer & Hoerl, 1986).

Of importance to this research is 'best' subset algorithms. 'Best' subset selection algorithms can be used to select a set of nonredundant variables from a larger collinear set (Hoerl, Schuenemeyer & Hoerl, 1986). However, it is not clear how well 'best' subset selection algorithms perform in the presence of collinearity.

'Best' Subset Selection Algorithms

The most thorough subset selection technique is called "all possible regressions" or " all subsets". This technique fits all $\binom{p}{c}$ c = 1, ... , p regression models where p is the total number of predictor variables in the set and c is the subset size. However if p is large then the $2^p - 1$ different models produced by this technique rapidly becomes too cumbersome to manage (Younger, 1985, p. 487). Thus several algorithms have been developed which build a 'best' subset of predictor variables in a stepwise manner.

Forward Selection

This method begins with no predictor variables in the model. At each step one variable is added to the model provided that it meets the criterion to enter. The 'best' subset is reached when either no more variables meet the criterion to enter or all the variables have been entered into the model. The criterion to enter is usually stated as an F statistic, so that variable i is added to the c-term equation if

$$F_i = \max\left[\frac{RSS_c - RSS_{c+i}}{MSE_{c+i}}\right] > F_{in} \quad i = c+1, c+2, \cdots, p \qquad (18)$$

where the candidate variables are ordered such that the first c variables are the variables already entered in the model, RSS_c and RSS_{c+i} are the residual sum of squares of the c-term model and c + variable i-term model, respectively, and MSE_{c+i} is the mean square error of the c+variable i-term model (Hocking, 1976).

It is felt that forward selection may miss groups of variables that perform poorly individually, but very well as a group (e.g. Mantel, 1970). In response to this criticism the backward elimination method was developed.

Backward Elimination Procedure

This technique begins with all the candidate predictor variables in the model. At each step the variable with the smallest F-ratio is eliminated if it does not meet the pre-specified criterion to remain in the model. The 'best' subset is reached when all the remaining variables either have been eliminated or else

meet the criterion to remain. The criterion to remain is usually stated as an F statistic. That is, variable i is deleted from the c-term model if

$$F_i = \min \left[\frac{RSS_{c-i} - RSS_c}{MSE_c} \right] < F_{out} \quad i = 1, 2, \cdots, c \qquad (19)$$

where the first c candidate variables are those remaining in the model, and RSS_{c-i}, RSS_c and MSE_c are defined in a manner similar to equation 18 (Hocking, 1976)

Stepwise Method

In forward selection, it is possible that a variable selected at an early stage may become superfluous at a later stage as other variables enter the model. Similarly, in backward elimination a variable deleted at an early stage cannot be re-entered into the model should it become a significant predictor again as other variables are deleted. In response to this a combination of the two methods was developed by Efroymson (1960). The method is basically forward selection but at each step the model is examined for the possibility of deleting a variable as in backward elimination. [1] Both a criterion to enter and a criterion to remain must be specified in this method (Hocking, 1976). The 'best' subset is reached when either no new variables meet the criterion for inclusion or when the variable to be entered was the one deleted at the previous step (Younger, 1985, p. 489).

Problems Associated with Subset Selection Algorithms

These methods have been criticized for many reasons. One of the most troublesome aspects of these methods is that they may not agree on the 'best' subset of predictor variables. None of these methods guarantee that the subset with the lowest residual sum of squares value will be found for each subset size (Berk, 1978; Hocking, 1976).

Illustrating this, Berk (1978) compared the residual mean squares for the

---

[1] A backward stepwise algorithm can also be implemented (See BMDP, 1988, p.373; Younger, 1985, pp. 489, 501-502).

backward and forward procedures to all subsets. The three methods were employed on nine data sets. In three of these the subset selection algorithms agreed, but in three others there were increases of 20% to 30% in residual variance for the backward and forward procedures over that of the all subsets procedure. However, when comparing their performance based on known populations, the differences among the three procedures was much smaller, 7% or less in eight of the nine data sets. While only the forward, backward and all subsets procedures were studied here, Berk (1978) notes that the stepwise procedure is likely to be an improvement.

A further criticism of the backward and forward procedures is that they imply an order of importance to the order in which variables are added to or deleted from the model. These procedures were never claimed to have this property by their original proponents (Hocking, 1976). All those who comment on the usefulness of 'best' subset selection algorithms, caution that the user must use his or her own knowledge of the subject under investigation in examining the results of these procedures (e.g. Flack & Chang, 1987; Hoerl, Schuenemeyer & Hoerl, 1986).

Inflation of R^2 in 'Best' Subset Selection

A number of criteria for assessing the "appropriateness" of a subset have been proposed. Among them is the squared multiple correlation coefficient, R^2.

Recall from equation 9 that the squared multiple correlation coefficient is calculated as

$$R^2 = \frac{SSR}{SST}.$$

R^2, then, is a measure of the proportion of the variance in y that is accounted for by the model. When no subset selection has taken place, a significance test of R^2 is

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \qquad (20)$$

with k and n - k - 1 degrees of freedom, where k is the number of predictor variables and n is the sample size.

There is one problem, however, associated with the use of R^2 as a measure of the variance explained by a model. Even when the number of predictors in the model is fixed, the sample R^2 is a positively biased estimate of the population coefficient of determination. That is, even when the predictor variables are uncorrelated with the response variable in the population, nonzero correlational values would be present in the sample simply due to random sampling variation (Cohen & Cohen, 1983, pp. 105-106). Since further capitalization on chance occurs when a subset selection algorithm is used to choose k predictors from p candidate variables, the bias in R^2 and hence the F-statistic is increased (see Berk, 1978). Thus, under cross-validation, a significant multiple correlation coefficient from a stepwise analysis may shrink drastically (Wilkinson, 1979).

A number of researchers have used Monte Carlo methods to determine the distribution of the sample R^2 statistic under subset selection. Diehr and Hoflin (1974) developed approximate percentage points for R^2. Their results are restricted to independence among the k predictor variables. When the predictor variables are collinear their results are conservatively biased.

Rencher and Pun (1980) extended Diehr and Hoflin's (1974) results to include the average inflation of R^2 under subset selection, upper percentage points of R^2, correlated predictor variables, and the situation where p, the total number of predictor variables, exceeds n, the sample size. Using stepwise regression, Rencher and Pun (1980) showed large increases in the average value of R^2 under selection, especially when p is greater than n. When the predictor variables were intercorrelated, the inflation of R^2 was somewhat less.

Wilkinson (1979) constructed tables of the upper 95th and 99th percentage points of the sample R^2 distribution in forward selection using Monte Carlo simulation and least-squares smoothing techniques. Like the Diehr and Hoflin (1974) study, the results are applicable to uncorrelated predictor variables. The tabled results are likely to be conservative when the predictor variables are correlated.

Since R^2 is commonly used to evaluate subets chosen by these algorithms, its inflation may be misleading to researchers. Therefore, Wilkinson (1979) recommends that researchers use his tables to evaluate the significance of the final equation selected through the a stepwise procedure.

An alternative approach is to develop an estimate of the population R^2 that is not positively biased. One such estimate is given by the shrunken $\tilde{R}^2$ where

$$\tilde{R}^2 = 1 - (1 - R^2)\frac{n - 1}{n - k - 1}. \qquad (21)$$

Cohen and Cohen (1983, pp. 106-107) suggest this estimate is appropriate when k, the number of predictor variables, is fixed. The degree of shrinkage will be larger for small values of R^2 and for large values of the ratio k/n. $\tilde{R}^2$ may take on negative values which, by convention, are reported as zero.

Whenever the k predictor variable have been selected by a subset selection algorithm, Cohen and Cohen (1983, pp. 106-107) indicate that $\tilde{R}^2$ will still be too large. In such a case, they recommend that p, the total number of candidate predictor variables, be used in place of k in the calculation of $\tilde{R}^2$.

Controlling Stepwise Algorithms

Controlling subset selection algorithms involves two concepts. Firstly, the sample size, n, affects the power of multiple regression to detect effects in the data. Secondly, the criterion to enter or remain in the model controls the number of variables that remain in the final model.

Like any other statistical test, stepwise algorithms require protection against incorrect results (i.e., false positives) without lowering the power to detect correct results. This introduces the concepts of α, the probability of rejecting a true null hypothesis (a Type I error), and β, the probability of failing to reject a false null hypothesis (a Type II error). The probability of correctly rejecting a false null hypothesis, $1-\beta$, is called the power of the test.

Any statistical test of a null hypothesis can be seen as a function of these four parameters:

1. The power of the test ($1-\beta$).
2. The probability of Type I error (α). As α increases power increases.
3. The sample size (n). As n increases power increases.
4. The magnitude of the effect under study in the population. The larger the effect the greater the power.

These four parameters are interrelated. For a given sample size and population effect size, setting the value of α determines β and vice-versa. Therefore, the usual method of controlling both α and β at acceptable levels is to set the value of α and then calculate the sample size necessary to control β for a given effect size. This technique can be used in multiple regression.

Stopping Rules in Subset Selection Algorithms

Setting the criterion to enter or remain in the model is usually done by setting the significance level of the F to enter in forward selection, the F to delete in backward elimination and both in the stepwise procedure. Because of the sequential nature of the computations, the number of variables can be controlled by making the F to enter sufficiently large so that not all the candidate variables enter the model or the F to delete sufficiently small so that not all the candidate variables are deleted from the model. Consequently, schemes for selecting F to enter and F to delete are known as 'stopping rules' (Hocking, 1976).

Bendel and Afifi (1977) used the unconditional mean square error of prediction (UMSE) to establish optimum levels of significance in forward selection. They compared the mean values of the normalized prediction error for values of α = 0.05 to .4 in increments of .05 stratified by the number of degrees of freedom. Their findings suggest that a significance level between 0.15 and 0.25 yields an F to enter that is large enough to keep nonauthentic candidate variables out of the model yet small enough so that authentic candidate variables could be detected. The best overall results occurred with α = 0.15. Hoerl, Schuenemeyer and Hoerl (1986) confirmed these results using the stepwise procedure.

Other literature (Lovell, 1983; Wilkinson, 1979) states that when subset selection occurs the level of significance is inflated. Wilkinson (1979) states that when k is fixed the usual test of the null hypothesis, H_0: $\beta_0 = \beta_1 = \ldots = \beta_k = 0$, given by

$$F = \frac{R^2 (n - k - 1)}{(1 - R^2) k} = \frac{MSR}{MSE}$$

has an F distribution with k and n-k-1 degrees of freedom under the null hypothesis. When k predictors are chosen from p candidate predictor variables on the basis of sample data, this statistic is not distributed as a central F variable (Pope & Webster, 1972). No exact distributions are known except for the two cases where k = 1 and k = p. In the case where k = 1 and the predictor that maximizes the sample R^2 is chosen, the F statistic may be used with the critical value

$$\alpha = 1 - (1 - \tilde{\alpha})^{1/p}, \qquad (22)$$

where $\tilde{\alpha}$ is the probability of making at least one Type I error in the set (family) of tests or the maximum familywise rate of Type I error.

Lovell (1983) compared the familywise levels of significance for various nominal values of α when choosing k=2 from p=2, 5, 10, 20, 100, and 500 orthogonal candidate predictor variables. Lovell's (1983) results show that the claimed nominal level of significance is increasingly inflated as the number of candidate predictor variables increases. So, for example, searching for the best two predictor variables out of ten candidate variables at a claimed nominal level of significance of 5% actually yields a familywise error rate of 22.6%. Therefore, Lovell suggests conducting the test at a more conservative nominal level of significance to counteract the effect of searching. Consequently, Lovell (1983) suggests that when choosing the best k out of p candidate explanatory variables the familywise level of significance can be calculated (approximately) as

$$\tilde{\alpha} = 1 - (1 - \alpha)^{p/k}, \qquad (23)$$

where $\tilde{\alpha}$ and α are the familywise and nominal levels of significance, respectively.

According to the SAS USER'S GUIDE: STATISTICS (SAS Institute,1985) the choice of the significance level is dependent upon the goal of the investigation. If it is necessary to guard against any variables that do not contribute to the predictive power of the model in the population entering the model a small significance level is warranted. If a model that provides the best

prediction using sample estimates is required a more moderate significance level is warranted (SAS Institute, 1985, p. 765).

Collinearity and 'Best' Subset Selection Algorithms

It is clear that collinearity is damaging to least-squares analysis. Its effect on the size and stability of regression coefficients is well documented (See Farrar & Glauber, 1967; Gordon, 1968; Rockwell, 1975).

Until recently, however, little work has focused on the effect of collinearity under subset selection. Citing the serious distortions that are introduced in standard analysis by collinear data, Chatterjee and Price (1977, pp. 203, 206) simply state that they do not recommend the use of stepwise procedures in a collinear situation. They go on to say that with a small number of collinear variables it is possible to evaluate all-possible equations to select an equation. They also quote Mantel (1970) in saying that backward elimination is better able to handle collinearity than the forward procedure.

Beale (1970) disputes Mantel's (1970) assertion regarding backward elimination. When variables are linearly dependent, it is a pure matter of chance which variable gets eliminated by backward elimination. Once eliminated, a variable is irretrievably excluded from the model even if the other variables involved in the dependency are subsequently eliminated (Beale, 1970). Gunst and Mason (1977) corroborate this finding. Because all collinear variables tend to have small t statistics, backward elimination may delete collinear predictor variables somewhat randomly, i.e., not on the basis of the true magnitude of the β_i in the population (Gunst & Mason, 1977).

Hoerl, Schuenemeyer and Hoerl (1986) note that subset selection has in fact been used to overcome the problems of least-squares estimation with collinear data. Their simulation, however, led them to recommend that subset selection not be used as a general strategy to combat collinearity.

Lovell (1983) investigated the performance of forward selection under various model conditions. Using twenty candidate explanatory variables Lovell (1983) artificially generated dependent variables from nine different models. To observe whether forward selection would be likely to select those candidate variables which participated in the generation of the dependent variable (authentic predictor variables) from a larger set in the presence of collinearity, two sets of quite closely related time series were included among the twenty

candidate variables. Fifty samples of 23 observations were generated for each model. When the null hypothesis was true and the level of significance set to 5%, forward selection correctly specified "none significant" 64% of the time. In the non-null case, forward selection chose variables that participated in the generation of the dependent variable 70% of the time.

Lovell (1983) does not quantify the severity of the collinearity present in the data, stating only that " the candidate explanatory variables used in the simulations were highly collinear rather than orthogonal" (Lovell, 1983). Through fear of inflating the Type I error rate, Lovell (1983) kept the level of significance at 5%, well below the level of 15% to 25% recommended by Bendel and Afifi (1977). Lovell (1983) also reminds the reader that a sample size of 23 may be less than required to choose the authentic variables from 20 highly collinear candidate series.

Flack and Chang (1987) used simulation experiments to assess the effects of sample size (n = 10, 20, 40) and the number of candidate variables (p = 10, 20, 40) on the frequency of selecting noise variables in the presence of authentic variables. Flack and Chang (1987) define a candidate variable, X_i, to be an authentic variable if its corresponding regression coefficient in the full regression model, β_i, is nonzero. X_i is defined to be a noise variable otherwise. Collinearity was introduced among the p candidate variables by the autocorrelation pattern

$$\rho_{ij} = \rho^{(j-i)} \qquad \text{for } j > i=1, 2, \ldots, p-1,$$

where $\rho_{i\,j}$ $(i \neq j)$ is the correlation coefficient between X_i and X_j. Three values of autocorrelation were selected (ρ= 0, 0.3, 0.5). The simple correlation between y and X_i was set to 0.5 for X_1 and X_2. For all other candidate variables $\rho_{yx_i} = 0$. Flack and Chang (1987) state that a design such as this will yield a regression equation with two authentic variables when $\rho = 0$ and three authentic variables when $\rho > 0$.

Two variable selection procedures were compared. An all-subsets, procedure (SAS RSQUARE; SAS Institute Inc. 1985) was used to find a subset of a prespecified size (k = 2). The second procedure used was a stepwise procedure (SAS STEPWISE, SAS Institute Inc. 1985) with the default level of significance of $\alpha = .15$.

The all-subsets procedure performed very well when p was small compared to n, and when the candidate variables were uncorrelated. The frequency with which authentic variables were chosen decreased as the number of candidate variables increased, and increased as the sample size increased.

As the autocorrelation coefficient increased, the frequency with which noise variables were selected increased, even though with a non-zero autocorrelation coefficient the number of authentic candidate variables was increased from 2 to 3.

When the stepwise procedure was used, k, the subset size was not prespecified. In general the subset size increased with p, the number of candidate variables, but always remained less than the sample size, n.

Flack and Chang (1987) only present the case where $\rho = 0.30$, so it is not known how the stepwise procedure fared under optimal conditions (p uncorrelated candidate variables where $p << n$) . The stepwise procedure performed best when $n = 40$, $p = 10$ and $\rho = 0.30$, where 34% of the samples correctly found three authentic variables, 50% found two authentic variables, and 16% found one authentic variable. The 25th and 75th percentiles of k, the subset size, were two and four, respectively. The median value of k was three variables. The 25th and 75th percentiles of P_n, the percentage of variables selected that are noise, were 0% and 50%, respectively. The median value of P_n was 33% noise variables. When $n = 10$, $p = 40$ and $\rho = .30$ stepwise performed very poorly. None of the samples correctly specified three authentic variables, 16% specified two authentic, 48% specified one authentic, and 36% specified noise variables only. k had a 25th percentile of eight variables, with a median value of nine. The 25th percentile of P_n was 88% noise variables, with a median value of 89%.

Clearly, the ability of both subset selection algorithms to select authentic variables from noise is affected by the sample size, the number of candidate variables, especially in relation to n, and the degree of collinearity present. However, the parametric conditions Flack and Chang (1987) investigated are not typical of psychological research. In particular, the simple correlation between the response variable and the authentic variables was set to 0.5, a value typically higher than those that characterize psychological relationships (Cohen, 1977). Secondly, Flack and Chang (1987) created intercorrelations

between candidate variables through serial correlation. This form of collinearity would have limited generalizability to psychological research.

Therefore the purpose of this research was to extend the research on the selection of candidate variables using best subset selection algorithms under parametric conditions characteristic of psychological research.

The investigation varied four factors:

1) the number of candidate variables (p);
2) the degree of intercorrelation between the candidate variables ($\rho_{x_i x_j}$);
3) the significance levels (α) for inclusions and/or deletion of candidate variables; and
4) the sample size (n) to simulate different levels of power.

Three subset selection algorithms were compared. Following Flack and Chang (1987) the proportion of authentic to noise variables were collected under FORWARD, BACKWARD, and STEPWISE selection (SAS Institute, 1987).

METHOD

Design

This simulation models the multiple regression design given in (1) where the k predictor variables in the model are chosen from P candidate variables by a subset selection algorithm. Among the P candidate variables, 6 authentic variables and P - 6 noise variables are defined. An authentic variable is specified to be a predictor variable whose corresponding population regression coefficient in the full model is nonzero. All predictor variables with corresponding zero full model population regression coefficients are defined as noise variables.

Data Generation

It was assumed that Y and $X_1, \ldots, X_P$ are randomly distributed and that their joint distribution follows a P + 1 multivariate normal distribution with mean **0** and covariance (correlation) structure defined below. The N observations for each of P candidate explanatory variables were generated by the algorithm employed by Galarneau-Gibbons (1981), McDonald and Galarneau (1975), and Wichern and Churchill (1978)

$$X_{ij} = (1- \delta^2)^{1/2} \; Z_{ij} + \delta \, Z_{i(P+1)} \qquad i = 1, \ldots, N; \; j = 1, \ldots, P, \tag{24}$$

where Z_{ij} and $Z_{i(P+1)}$ are independent identically distributed standard normal pseudo-random variables and δ is prespecified. The resulting candidate explanatory variables have a pairwise correlation of δ^2. Of the P candidate explanatory variables k = 6 were defined to be authentic predictor variables. These variables were generated using a value of δ which reflects a collinearity condition. The remaining P - 6 "noise" candidate variables were uncorrelated among themselves and with the authentic predictor variables.

Pseudo-random unit normal deviates were generated by the procedure due to Marsaglia, MacLaren and Bray (1964).

The number of candidate predictor variables were P = 12, 18, and 24. This represented the case where 50% (50%), 33.3% (66.7%) and 25% (75%) of the available predictor variables were authentic (noise), respectively.

Three different sets of correlation were considered corresponding to $\rho_{x_i x_j}$ = 0.00, 0.40, and 0.80. The range of these correlations typify the size of intercorrelations found in psychological test batteries (See Cronbach, 1987; Sax, 1989; Thorndike & Hagen, 1977). A value of $\delta = \sqrt{\rho_{x_i x_j}}$ corresponding to each value of $\rho_{x_i x_j}$ was used to generate three sets of explanatory variables having the specified correlation structure.

Observations on the dependent variable were generated by

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + e_i \quad i = 1, \ldots, N, \qquad (25)$$

where the e_i are independent identically distributed standard normal pseudorandom numbers and the X_{ij} are the predictor variables previously generated. The first six candidate predictor variables participated in the generation of the dependent variable and were therefore authentic predictor variables. The remaining P - 6 candidate predictor variables did not participate in the generation of the dependent variable and were thus "noise" variables.

The explanatory variables and the response variable were then standardized so that $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{y}$ were in correlation form; hence β_0 will be equal to zero due to the standardization process (Galarneau-Gibbons,1981; McDonald & Galarneau, 1975; Wichern & Churchill, 1978).

It was decided that the squared population coefficent of determination ($\rho_{yx_i}^2$) of the full model should be chosen to reflect a medium effect size in the noncollinear case (i.e. $\rho_{x_i x_j}$ = 0.0). Using equation 12, the conventionally medium squared coefficient of determination value was calculated to be 0.130435.

Recall from equation 11 that the squared semipartial correlation coefficient, sr_i^2, of a predictor variable, X_i, is the proportion of the total variance in y accounted for by X_i when the other predictor variables are already in the model. If there are k authentic predictor variables in the model then

$$sr_i^2 = \frac{SSR_k - SSR_{k-i}}{TSS},$$

where SSR_k, and SSR_{k-i} are the sums of squares for regression for the k-term model and the k minus variable i - term model, respectively and TSS is the total sums of squares. If the k predictor variables in the model are uncorrelated, R^2 is

equal to the sum of their k squared semipartial correlation coefficients. Assuming that these coefficients are equal their value is given by

$$sr_i^2 = \frac{R^2}{k}.$$

When the k variables in the model are uncorrelated, the semipartial correlation coefficient is equal to the simple correlation coefficient of the predictor variable with the response variable. Thus the value of the simple correlation between a authentic predictor variable and the response variable reflecting a medium effect size in the noncollinear case may be calculated as

$$r_{yj} = \left(\frac{R^2}{k}\right)^{1/2}$$

where the calculated value of $\rho_{yx_i}^2$ representing a medium effect size ($\rho_{yx_i}^2$ = .130435) is substituted for R^2 and k, the number of authentic predictor variables, is equal to six.

The regression coefficients are calculated from the relationship

$$\boldsymbol{\beta} = (\mathbf{R}_{XX})^{-1}\mathbf{r}_{xy} \qquad (26)$$

where $\mathbf{R}_{XX}$ is a predictor variable correlation matrix and $\mathbf{r}_{xy}$ is a vector of correlations between the response variable and the authentic variables. Table 4 shows the values of $\mathbf{R}_{XX}$, $\mathbf{r}_{xy}$, β, $\rho_{yx_i}^2$, and N used in this simulation. It may be noted that when the reponse and predictor variables have been transformed by the correlation transformation, the squared coefficient of determination may be calculated as

$$\rho_{yx_i}^2 = \boldsymbol{\beta}^T\mathbf{r}_{xy}$$

which, according to Equation 26, is dependent on $\mathbf{R}_{XX}$. In fact, Table 4 shows that the squared population coefficient of determination is reduced from a value representing a medium effect size when $\rho_{x_ix_j}$ = 0.0 to a value representing a conventionally small effect size when $\rho_{x_ix_j}$ = 0.8 according to the criterion of Cohen and Cohen (1983, p. 161).

Table 4

Data Generation Parameters

$\mathbf{R}_{XX}$	$\mathbf{r}_{XY}$	β	$\rho_{yx_i}^2$	N
$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & & & \\ . & . & . & . & . & . & & & \\ . & . & . & . & . & . & & \mathbf{I} & \\ . & . & . & . & . & . & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & & \end{bmatrix}$	$\begin{bmatrix} .147442 \\ .147442 \\ .147442 \\ .147442 \\ .147442 \\ .147442 \\ .0 \\ . \\ . \\ . \\ .0 \end{bmatrix}$	$\begin{bmatrix} .147442 \\ .147442 \\ .147442 \\ .147442 \\ .147442 \\ .147442 \\ .0 \\ . \\ . \\ . \\ .0 \end{bmatrix}$	.130435	30 60 90
$\begin{bmatrix} 1 & .4 & .4 & .4 & .4 & .4 & 0 & \dots & 0 \\ .4 & 1 & .4 & .4 & .4 & .4 & 0 & \dots & 0 \\ .4 & .4 & 1 & .4 & .4 & .4 & 0 & \dots & 0 \\ .4 & .4 & .4 & 1 & .4 & .4 & 0 & \dots & 0 \\ .4 & .4 & .4 & .4 & 1 & .4 & 0 & \dots & 0 \\ .4 & .4 & .4 & .4 & .4 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & & & \\ . & . & . & . & . & . & & & \\ . & . & . & . & . & . & & \mathbf{I} & \\ . & . & . & . & . & . & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & & \end{bmatrix}$	$\begin{bmatrix} .147442 \\ .147442 \\ .147442 \\ .147442 \\ .147442 \\ .147442 \\ .0 \\ . \\ . \\ . \\ .0 \end{bmatrix}$	$\begin{bmatrix} .049147 \\ .049147 \\ .049147 \\ .049147 \\ .049147 \\ .049147 \\ .0 \\ . \\ . \\ . \\ .0 \end{bmatrix}$	.043478	30 60 90

Table 4

Data Generation Parameters

$\mathbf{R}_{XX}$	$\mathbf{r}_{XY}$	β	$\rho_{yx_i}^2$	N
$\begin{bmatrix} 1 & .8 & .8 & .8 & .8 & .8 & 0 & \dots & 0 \\ .8 & 1 & .8 & .8 & .8 & .8 & 0 & \dots & 0 \\ .8 & .8 & 1 & .8 & .8 & .8 & 0 & \dots & 0 \\ .8 & .8 & .8 & 1 & .8 & .8 & 0 & \dots & 0 \\ .8 & .8 & .8 & .8 & 1 & .8 & 0 & \dots & 0 \\ .8 & .8 & .8 & .8 & .8 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & & & \\ . & . & . & . & . & . & & & \\ . & . & . & . & . & . & & \mathbf{I} & \\ . & . & . & . & . & . & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & & \end{bmatrix}$	$\begin{bmatrix} .147442 \\ .147442 \\ .147442 \\ .147442 \\ .147442 \\ .147442 \\ .147442 \\ .0 \\ . \\ . \\ . \\ .0 \end{bmatrix}$	$\begin{bmatrix} .029488 \\ .029488 \\ .029488 \\ .029488 \\ .029488 \\ .029488 \\ .029488 \\ .0 \\ . \\ . \\ . \\ .0 \end{bmatrix}$	.026087	30 60 90

For each combination of the above parameters, the ideal sample size, N, was defined as the sample size that would yield 80% power to detect the squared semipartial correlation coefficient. This is in accordance with Cohen and Cohen (1983, pp. 116-119) who proposed that a behavioural science researcher should adopt a power value of .80. To determine the sample size necessary to test the null hypothesis that any semipartial correlation or regression coefficient for a given X_i is zero equations (18) and (19) were employed. To investigate the effect of sample size in the detection of authentic variables, samples sizes that were 50% and 150% of the ideal sample size were also generated (See Table 4).

Three subset selection procedures were compared. The SAS STEPWISE procedure allows for the three different stepwise techniques: FORWARD, BACKWARD, and STEPWISE. Within a subset selection algorithm, it is possible to vary the level of significance for inclusion and/or deletion in the model. Therefore, level of significance values of 0.15, 0.05 and $\alpha_p = 1 - (1 - \tilde{\alpha})^{1/p}$ (where $\tilde{\alpha} = 0.15$ represents the familywise level of significance) formed a within algorithm condition.

The 0.15 value was chosen as it reflected the recommendations of Bendel and Afifi (1977) and corresponded to the value used by Flack and Chang (1987). The α_p value was chosen to reflect a concern for the issue of multiplicity of testing. Lovell (1983) and Wilkinson (1979) documented how the maximum familywise Type I error rate (MFWER) was inflated when k predictor variables were chosen from P candidate predictor variables. For P = 12, 18, and 24 candidate predictor variables, the MFWER equals .858, .946, and .980, respectively, when $\alpha = 0.15$. The value of α_p therefore was chosen to limit the MFWER to 0.15. For P = 12, 18, and 24, the protected inclusion and deletion values (α_p) were 0.0134519, 0.0089882, and 0.0067481, respectively.

Finally, since many statistical software packages use 0.05 as a default level of significance [See for example BMDP (Dixon et al, 1988 p. 381), SPSSX (1985, p. 57), and MINITAB (Ryan et al, 1981)], this value was also investigated. One should note however, that, for $\alpha = 0.05$, the MFWERs are .460, .603 and .708 for P = 12, 18, and 24, respectively.

The nominal significance levels (α) used in the simulation compared to the MFWER ($\tilde{\alpha}$) for choosing k = 6 from P = 12, 18, and 24 candidate variables are given in Table 5.

A complete crossing of all the levels of the three data conditions ($\rho_{x_i x_j}$, P, and N) and the single within algorithm condition (α) yields a total of 81 sets of conditions.

250 replications of each of the 27 possible combinations of the three data conditions were generated according to the algorithms given in equations 22 and 23 and stored on disk using FORTRAN. The data was then reread from disk by SAS and processed by each of the three subset selection algorithms in combination with each level of the within algorithm condition. Results from each of these algorithms were rerouted to disk so that SAS could reread this information and strip the values of R^2, k and number of authentic candidate variables in the final subset for further processing (see Appendix A for FORTRANand SAS programs).

Table 5

Level of Significance Conditions

Number of Candidate Variables (P)	Nominal Type I Error Rate (α)	Maximum Familywise Type I Error Rate MFWER($\tilde{\alpha}$)
12	.05	.460
18	.05	.603
24	.05	.708
12	.15	.858
18	.15	.946
24	.15	.980
12	.0134519	.15
18	.0089882	.15
24	.0067481	.15

Statistics

Each combination of the three data conditions and one within algorithm condition were tested on each of the three subset selection algorithms: (1) STEPWISE, (2) BACKWARD, and (3) FORWARD.

In any stepwise subset selection algorithm, the number of variables selected is usually not prespecified. Therefore k, the final subset size, was a random variable determined by the 'stopping rule' used by the subset selection algorithm. In general, a variable is entered into the equation when a test of its partial correlation is significant at some prespecified level of significance and deleted when it is not significant. The procedure is terminated when either no more significant variables are left to be entered into the model, no more insignificant variables are left to be deleted from the model or, in the case of the STEPWISE algorithm, the only significant variable to be entered into the model is the one deleted from the model in the last step. The final subset chosen when the procedure terminates is deemed the 'best' subset based upon the stopping rule used.

Upon termination of the procedure four characteristics of the final subset were noted:

1) k, the subset size,
2) R^2, the multiple coefficient of determination,
3) C_N, the number of noise variables in the final subset, and
4) C_A, the number of authentic variables in the final subset.

From these variables, three additional variables were calculated:

1) $R_k^2 = 1 - (1 - R^2)\frac{N - 1}{N - k - 1}$, the shrunken value of R^2 (See Cohen & Cohen, 1983, p. 106),
2) $R_p^2 = 1 - (1 - R^2)\frac{N - 1}{N - P - 1}$, the shrunken value of R^2 appropriate when subset selection has taken place (See Cohen & Cohen, 1983, pp. 106-107), and
3) $P_N = \frac{C_N}{k}$, the proportion of the final subset that is noise.

Values for each of the seven dependent variables above were calculated and summarized over the 250 replications by their mean and standard deviation.

When it occurred that no variables remained in the final subset, R^2, R_k^2 and R_p^2 were set to a missing value rather than zero, since R_k^2 and R_p^2 may legitimately take on the value zero when their calculation leads to a negative number. This was deemed the best response. In such a case, C_A and C_N were set to zero and hence $k = C_A + C_N = 0$ and $P_N = C_N / k$ leads to a missing value.

RESULTS

Selection of Authentic Predictor Variables in the Presence of Noise

In the generation of each sample of a single response variable and P candidate predictor variables, the first six (X_1, X_2 , ... , X_6) predictor variables were defined to be authentic predictor variables. That is, these predictor variables participated in the generation of the response variable by having nonzero regression coefficients. The remaining P - 6 predictor variables, which did not participate in the generation of the response variable, were defined to be noise. Of interest, then, was the effect that (1) the three study factors (degree of collinearity, number of candidate variables, and sample size), (2) the subset selection algorithms (STEPWISE, FORWARD, and BACKWARD) and (3) the α inclusion/deletion levels (α_p, .05 and .15) had on the selection of authentic predictor variables in the presence of noise. Thus, a comparison of the effect that the three study factors within each inclusion/deletion level of significance had on (1) the mean number of authentic variables (C_A), (2) the mean number of noise variables (C_N) and (3) the mean percentage of the final subset that is noise (P_N), may be found in Tables 6, 7, and 8 for the STEPWISE, FORWARD and BACKWARD procedures, respectively (see Appendix B for associated standard errors).

An examination of the results obtained within the subset selection algorithms shows that when $\alpha = \alpha_p$ the mean value of C_A was consistently less than one for all three algorithms. That is, often no authentic variables remained in the final subset at this level of significance. As the level was increased to 0.15, the mean value of C_A was increased to a value generally greater than one. However, even at this level, when $\rho_{x_i x_j} = 0.8$ the mean value of C_A remained generally less than one in the STEPWISE and FORWARD algorithms (See Tables 6 and 7). The mean values of C_N were similarly affected by a change in the inclusion/deletion levels.

Table 6

Mean Frequency of Authentic and Noise Variables and %Noise Contained in the Final Subset in the STEPWISE Procedure[a]

			I/D Level of Significance (α)								
			α_p			0.05			0.15		
			Variable								
$\rho_{x_i x_j}$	P	N	C_A	C_N	P_N	C_A	C_N	P_N	C_A	C_N	P_N
0.0	12	30	0.228	0.084	25.7	0.680	0.368	32.2	1.516	1.100	41.9
		60	0.476	0.096	15.8	1.076	0.316	20.0	2.192	0.928	27.9
		90	0.728	0.068	8.1	1.580	0.300	12.9	2.708	0.848	22.1
	18	30	0.200	0.112	34.9	0.684	0.652	47.7	1.604	2.120	55.2
		60	0.360	0.100	22.3	1.076	0.696	37.2	2.000	2.028	48.8
		90	0.492	0.080	14.8	1.560	0.656	29.5	2.712	1.980	40.5
	24	30	0.124	0.112	47.3	0.672	1.132	59.0	1.548	3.588	68.9
		60	0.288	0.128	28.7	1.172	0.984	44.0	2.104	3.012	56.2
		90	0.460	0.132	21.2	1.580	0.964	35.8	2.760	2.952	50.5
0.4	12	30	0.180	0.108	35.9	0.460	0.288	34.7	1.032	0.948	45.1
		60	0.308	0.072	17.2	0.580	0.292	29.1	1.136	0.940	42.9
		90	0.444	0.088	12.8	0.744	0.312	23.0	1.292	0.968	37.4
	18	30	0.108	0.108	46.8	0.476	0.704	56.3	1.036	2.200	66.1
		60	0.256	0.128	30.5	0.600	0.756	52.1	1.208	2.100	61.1
		90	0.396	0.104	16.2	0.824	0.624	35.5	1.308	1.904	54.3
	24	30	0.140	0.088	36.8	0.456	1.104	69.4	1.216	3.708	74.5
		60	0.204	0.136	38.1	0.584	0.900	58.5	1.204	2.860	67.7
		90	0.332	0.132	24.7	0.708	0.896	50.1	1.316	2.804	64.9
0.8	12	30	0.108	0.068	36.8	0.324	0.324	48.0	0.788	1.000	56.0
		60	0.204	0.100	30.6	0.440	0.328	39.4	0.900	0.996	51.5
		90	0.308	0.076	18.5	0.592	0.240	24.7	1.052	0.852	42.3
	18	30	0.116	0.084	40.8	0.300	0.744	70.5	0.776	2.080	73.5
		60	0.168	0.132	40.0	0.400	0.604	57.2	0.884	1.884	67.2
		90	0.236	0.108	29.7	0.580	0.604	46.4	0.976	1.836	62.3
	24	30	0.108	0.092	43.3	0.404	1.052	69.6	0.956	3.532	79.1
		60	0.128	0.108	47.2	0.456	0.900	63.3	0.944	2.916	76.0
		90	0.256	0.148	33.9	0.560	0.968	59.3	1.024	3.012	74.3

[a]Note: $\rho_{x_i x_j}$ = Degree of collinearity, P = Number of candidate variables, N = Sample size; $\alpha_p = 1-(1-\tilde{\alpha})^{1/p}$ where $\tilde{\alpha}$=0.15; C_A is the frequency of authentic variables, C_N is the frequency of noise variables, and P_N is the percentage of noise variables in the final subset;

Table 7

Mean Frequency of Authentic and Noise Variables and %Noise Contained in the Final Subset in the BACKWARD Procedure [a]

			I/D Level of Significance (α)								
			α_p			0.05			0.15		
			Variable								
$\rho_{x_i x_j}$	P	N	C_A	C_N	P_N	C_A	C_N	P_N	C_A	C_N	P_N
0.0	12	30	0.312	0.140	30.5	0.952	0.588	35.2	1.856	1.368	41.9
		60	0.508	0.108	16.5	1.192	0.356	19.5	2.360	1.048	28.4
		90	0.808	0.092	9.2	1.648	0.332	13.5	2.808	0.972	23.9
	18	30	0.256	0.208	42.1	1.164	1.412	51.8	2.324	3.628	58.5
		60	0.448	0.184	26.4	1.216	0.896	40.0	2.248	2.500	51.4
		90	0.608	0.120	14.5	1.744	0.780	30.0	2.884	2.264	41.9
	24	30	0.192	0.316	51.6	1.504	3.676	65.5	2.816	7.804	72.7
		60	0.324	0.176	31.3	1.332	1.552	49.5	2.432	3.944	59.5
		90	0.512	0.168	22.6	1.792	1.160	37.0	3.000	3.472	51.5
0.4	12	30	0.196	0.112	36.7	0.632	0.488	38.1	1.460	1.388	45.4
		60	0.344	0.104	20.0	0.668	0.332	29.6	1.364	1.088	43.0
		90	0.452	0.112	13.9	0.824	0.348	23.6	1.492	1.068	37.5
	18	30	0.156	0.240	57.7	0.812	1.348	60.9	1.904	3.516	64.5
		60	0.276	0.168	33.3	0.744	0.972	52.4	1.472	2.632	62.1
		90	0.384	0.124	18.4	0.880	0.760	37.8	1.492	2.212	55.9
	24	30	0.204	0.256	48.8	1.228	3.264	70.3	2.828	8.104	74.3
		60	0.204	0.168	41.7	0.716	1.288	61.1	1.584	3.856	69.9
		90	0.344	0.144	25.7	0.800	1.136	53.0	1.612	3.320	64.4
0.8	12	30	0.224	0.116	33.9	0.616	0.540	45.4	1.516	1.328	43.9
		60	0.276	0.104	27.6	0.652	0.364	32.6	1.384	1.092	41.8
		90	0.376	0.088	18.4	0.776	0.276	23.4	1.456	0.972	35.9
	18	30	0.204	0.204	41.4	0.852	1.404	61.8	1.920	3.580	64.2
		60	0.244	0.148	36.3	0.760	0.768	48.1	1.532	2.392	58.9
		90	0.292	0.128	29.9	0.748	0.644	43.2	1.456	2.148	56.6
	24	30	0.264	0.416	48.9	1.480	3.964	68.0	2.856	8.076	73.5
		60	0.128	0.156	53.0	0.784	1.392	60.8	1.588	3.824	70.0
		90	0.316	0.176	33.0	0.752	1.216	57.1	1.404	3.456	70.0

[a]Note: See Table 6 note.

Table 8

Mean Frequency of Authentic and Noise Variables and %Noise Contained in the Final Subset in the FORWARD Procedure [a]

			I/D Level of Significance (α)								
			α_p			0.05			0.15		
			Variable								
$\rho_{x_i x_j}$	P	N	C_A	C_N	P_N	C_A	C_N	P_N	C_A	C_N	P_N
0.0	12	30	0.228	0.084	25.7	0.684	0.368	32.2	1.536	1.104	41.6
		60	0.476	0.096	15.8	1.076	0.316	20.0	2.204	0.936	27.9
		90	0.728	0.068	8.1	1.580	0.304	13.0	2.716	0.848	22.1
	18	30	0.200	0.112	34.9	0.692	0.652	47.5	1.640	2.180	55.2
		60	0.360	0.100	22.3	1.072	0.728	37.9	2.024	2.064	48.9
		90	0.492	0.080	14.8	1.560	0.660	29.6	2.728	2.020	40.7
	24	30	0.124	0.112	47.3	0.672	1.140	59.1	1.584	3.648	68.8
		60	0.288	0.128	28.7	1.184	0.988	43.8	2.152	3.056	56.0
		90	0.460	0.132	21.3	1.580	0.976	35.9	2.788	3.016	50.7
0.4	12	30	0.180	0.108	35.9	0.460	0.288	34.7	1.052	0.960	44.9
		60	0.308	0.072	17.2	0.580	0.292	29.1	1.140	0.948	43.0
		90	0.444	0.088	12.8	0.744	0.312	23.0	1.292	0.968	37.4
	18	30	0.108	0.108	46.8	0.480	0.708	56.2	1.068	2.228	65.7
		60	0.256	0.128	30.5	0.604	0.756	52.0	1.248	2.128	60.8
		90	0.396	0.104	16.2	0.824	0.628	35.5	1.344	1.912	53.9
	24	30	0.140	0.088	37.9	0.460	1.112	69.6	1.268	3.804	74.3
		60	0.204	0.136	38.1	0.588	0.892	58.3	1.240	2.852	67.3
		90	0.332	0.132	24.7	0.708	0.896	50.1	1.332	2.816	64.7
0.8	12	30	0.108	0.068	36.8	0.328	0.324	47.9	0.808	1.016	55.9
		60	0.204	0.100	30.6	0.440	0.328	39.4	0.900	1.000	51.6
		90	0.308	0.076	18.5	0.592	0.240	24.7	1.060	0.852	42.3
	18	30	0.116	0.084	40.8	0.300	0.744	70.5	0.784	2.108	73.6
		60	0.168	0.132	40.0	0.400	0.604	57.2	0.884	1.928	67.6
		90	0.236	0.108	29.7	0.580	0.608	46.5	0.992	1.840	62.2
	24	30	0.108	0.092	43.3	0.404	1.080	70.0	1.016	3.588	78.9
		60	0.128	0.108	47.2	0.456	0.900	63.3	0.972	2.948	75.8
		90	0.256	0.148	33.9	0.560	0.972	59.4	1.036	3.036	74.2

[a]Note: See Table 6 note.

Within each method, the maximum mean value of C_A is to be found when $\alpha = 0.15$, $\rho_{x_i x_j} = 0.0$ and N = 90. For these conditions and for each value of P (12, 18 and 24) the mean value of C_A equals 2.708, 2.712 and 2.760 for the STEPWISE procedure, 2.808, 2.884, and 3.000 for the BACKWARD procedure, and 2.716, 2.728, and 2.788 for the FORWARD procedure. Hence, none of the subset selection procedures consistently contained all six of the authentic variables in their final subset. In fact, the corresponding mean values of C_N indicated that the final model may contain a large proportion of noise variables, especially when P = 24. For P = 12, 18 and 24, the mean value of C_N equals 0.848, 1.980, and 2.952 for the STEPWISE procedure, 0.972, 2.264, and 3.472 for the BACKWARD procedure, and 0.848, 2.020 and 3.016 for the FORWARD procedure. The results for P_N show this to be true.

In addition, note that while results for the STEPWISE and FORWARD procedures were generally very consistent, the mean values of both C_A and C_N were increased in the BACKWARD procedure for each level of significance (See Table 8). This indicates that the BACKWARD procedure likely eliminates fewer variables (both authentic and noise) to obtain the final subset. The same is not always true of P_N. The mean values of P_N in the BACKWARD procedure generally exceeded those of the STEPWISE and FORWARD procedures when $\rho_{x_i x_j} = 0.0$ and 0.4. However, when $\rho_{x_i x_j} = 0.8$, the BACKWARD mean P_N results were generally less than those of the other algorithms. Thus, the BACKWARD procedure may obtain a final subset with both more authentic variables and a lower proportion of noise than either the STEPWISE or FORWARD procedures when collinearity is high. Still, when $\rho_{x_i x_j} = 0.8$, P = 24 and N = 30 the mean percentage of noise variables in the BACKWARD procedure was 73.5% (compared to 79.1% and 78.9% in the STEPWISE and FORWARD procedures, respectively).

From the above findings, it is clear that trends due to the degree of collinearity, the number of candidate variables and sample size do exist in the analysis of C_A, C_N, and P_N. Therefore, tests for trend, employing orthogonal polynomials, were used to examine the effect of each of the three data conditions ($\rho_{x_i x_j}$, P and N) on each of the three dependent variables (C_A, C_N, and P_N) within each combination of method (STEPWISE, BACKWARD, and FORWARD) and the within algorithm condition, α (α_p, 0.05 and 0.15). Two contrasts representing the linear and quadratic effects of each of the factors

were computed. Additionally, each of the possible two-way and three-way interaction contrasts were calculated for a total of 26 one degree of freedom contrasts.

The sums of squares for each contrast were generated using the Contrast statement in the SAS procedure, GLM. To evaluate the importance of each contrast, an r^2 value, calculated as the contrast sum of squares divided by the model or explained sum of squares, was defined. Tables 9, 11, and 13 contain an enumeration of the r^2 values associated with each of the trend components for each of the dependent measures.

Cohen (1969) gives conventional "small", "medium" and "large" values for squared correlation coefficients in behavioural science. Using these values as a reference, r^2 values of less than 0.01 were defined as negligible, values from 0.01 to less than 0.09 were defined as "small", values from 0.09 to less than 0.25 were defined as "medium" and values greater than or equal to 0.25 were defined as "large". Classification of the r^2 values in this way, provides a description of the strength of the relationship between the factor, as represented by the contrast, and the dependent variable. Only trend components accounting for at least one small effect size for any combination of algorithm and inclusion/deletion level were enumerated.

To determine the direction of any linear relationship between the factors and the dependent variables, multiple linear regression procedures were used. That is, each trend component represented a predictor variable in the regression equation. Each of the dependent variables were regressed on the full-rank data matrix thus produced. The sign of the regression coefficient corresponding to the linear trend component vector indicates the direction of the linear relationship. The sign of the linear relationships have been included in the tables of r^2 values so that the strength and direction of the linear relationships are immediately apparent.

Trend Analysis of $\rho_{x_i x_j}$, P and N within Method and Inclusion/Deletion Level

Main Effects

Collinearity. Table 9 contains the r^2 values associated with the trend analysis of C_A. The results show that increasing collinearity ($\rho_{x_i x_j}$) had a large negative linear effect on the number of authentic variables in the final subset.

Within each combination of algorithm and inclusion/deletion level, the linear contrast in $\rho_{x_i x_j}$ resulted in a large r^2 value. A small to moderate quadratic component was shown to exist in the BACKWARD procedure where the quadratic contrast in $\rho_{x_i x_j}$ resulted in a small r^2 value when $\alpha = \alpha_p$ and a medium r^2 value when $\alpha = 0.05$ and 0.15. As well, a small quadratic effect was shown to exist when $\alpha = 0.05$ and 0.15 within the other algorithms. The first part of Table 10 shows the effect of collinearity on C_A. The strong negative effect of increasing collinearity is plain, particularly as $\rho_{x_i x_j}$ was increased from the uncorrelated case ($\rho_{x_i x_j} = 0$) to the correlated cases ($\rho_{x_i x_j} = 0.4$ and 0.8).

In comparison, the trend analysis for C_N given in Table 11 showed that the degree of collinearity did not significantly affect the number of noise variables in the final subset. A small quadratic effect was evident only in the STEPWISE and FORWARD procedures and only when $\alpha = \alpha_p$. The remainder of the contrasts in $\rho_{x_i x_j}$ resulted in r^2 values below 0.01. The lack of a trend due to increasing collinearity is evident in the first part of Table 12.

On the other hand, a trend due to the degree of collinearity is again evident in the analysis of P_N (See Table 13). Within the STEPWISE and FORWARD procedures, a moderate to large positive linear relationship was shown to exist for each value of α. To a lesser extent, the same relationship was shown to exist within the BACKWARD procedure. As well, a small quadratic effect due to $\rho_{x_i x_j}$ was shown to exist when $\alpha = 0.05$ and 0.15 in the BACKWARD procedure.

The positive effect of increasing collinearity on the mean proportion of noise variables in the final subset is shown in the first part of Table 14. It appears that increasing the value of $\rho_{x_i x_j}$ from 0.0 to 0.4 had the greatest effect on P_N in the BACKWARD procedure. The mean percentage noise was increased from 38.1% to 48.0% when $\alpha = 0.05$ and from 47.8% to 57.7% when $\alpha = 0.15$ when collinearity was increased from 0.0 to 0.4. A further increase in collinearity did not measurably change the mean value of P_N at these levels of α. In the STEPWISE and FORWARD procedures each increase in $\rho_{x_i x_j}$ produced a corresponding increase in P_N at each level of α.

Number of Candidate Variables. An analysis of the effect of increasing the number of candidate variables (P) on C_A indicated the presence of a moderate positive linear trend within the BACKWARD procedure when $\alpha = 0.05$ and 0.15 (See Tables 9 and 10). The mean value of C_A was increased from 0.884 and 1.744 when P = 12 to 1.154 and 2.236 when P = 24 for $\alpha = 0.05$ and

Table 9

Proportion of Model Sum of Squares Accounted for by Contrast [a]

Dependent Variable: C_A

METHOD:	STEPWISE			BACKWARD			FORWARD		
I/D LEVEL (α):	α_p	.05	.15	α_p	.05	.15	α_p	.05	.15
CONTRAST									
$\rho_{x_i x_j}$ (L)	-L	-L	-L	-L	-L	-L	-L	-L	-L
$\rho_{x_i x_j}$ (Q)		S	S	S	M	M		S	S
P (L)	-S	+	+	-M	+M	+M	-S	+	+
N (L)	+L	+L	+M	+L	+	+S	+L	+L	+M
N (Q)				S	S	S			
$\rho_{x_i x_j}$(L) x P(L)	S			S			S		
$\rho_{x_i x_j}$(L) x N(L)	S	S	S	M	M	M	S	S	S
$\rho_{x_i x_j}$(Q) x N(L)		S	S		S	S		S	S
P(L) x N(L)	S			S	S	M	S		
P(L) x N(Q)					S	S			

[a]Note: I/D = Inclusion/Deletion;

+/- signs indicate the direction of the relationship;

$\alpha_p = 1-(1-\tilde{\alpha})^{1/p}$ where $\tilde{\alpha}=0.15$;

$r^2 < 0.01$ = Negligible effect (left blank except for possible sign of relationship),

$0.01 \leq r^2 < 0.09$ = Small effect (S),

$0.09 \leq r^2 < 0.25$ = Medium effect (M),

$0.25 \leq r^2$ = Large effect (L);

$\rho_{x_i x_j}$ = Degree of collinearity, P = Number of candidate predictor variables, N = sample size;

(L) = Linear trend, (Q) = Quadratic trend;

Table 10
Effect of Collinearity, Number of Candidate Variables and Sample Size on the Mean Values of C_A[a]

		Method								
		STEPWISE			BACKWARD			FORWARD		
		I/D Level (α)								
		α_p	.05	.15	α_p	.05	.15	α_p	.05	.15
	$\rho_{x_i x_j}$									
	0.0	0.373	1.120	2.127	0.441	1.394	2.525	0.373	1.122	2.152
	0.4	0.263	0.604	1.194	0.284	0.812	1.690	0.263	0.605	1.220
	0.8	0.181	0.451	0.922	0.258	0.824	1.679	0.181	0.451	0.939
	P									
	12	0.332	0.720	1.402	0.388	0.884	1.744	0.332	0.720	1.412
	18	0.259	0.722	1.389	0.319	0.991	1.915	0.259	0.724	1.412
	24	0.227	0.732	1.452	0.276	1.154	2.236	0.227	0.735	1.488
	N									
	30	0.146	0.495	1.164	0.223	1.027	2.164	0.146	0.498	1.195
	60	0.266	0.709	1.397	0.306	0.896	1.774	0.266	0.711	1.418
	90	0.406	0.970	1.683	0.455	1.107	1.956	0.406	0.970	1.699
$\rho_{x_i x_j}$	N									
0.0	30	0.184	0.679	1.556	0.253	1.207	2.332	0.184	0.682	1.587
	60	0.375	1.108	2.099	0.427	1.246	2.347	0.375	1.111	2.126
	90	0.560	1.573	2.727	0.643	1.728	2.897	0.560	1.573	2.744
0.4	30	0.143	0.464	1.095	0.185	0.891	2.064	0.143	0.467	1.129
	60	0.256	0.588	1.183	0.275	0.709	1.473	0.256	0.591	1.209
	90	0.391	0.759	1.305	0.393	0.835	1.532	0.391	0.759	1.323
0.8	30	0.111	0.342	0.840	0.231	0.983	2.097	0.111	0.344	0.869
	60	0.167	0.432	0.909	0.216	0.732	1.501	0.167	0.432	0.918
	90	0.267	0.577	1.017	0.328	0.759	1.439	0.267	0.577	1.029
P	N									
12	30	0.172	0.488	1.112	0.244	0.733	1.611	0.172	0.491	1.132
	60	0.329	0.699	1.409	0.376	0.837	1.703	0.329	0.699	1.415
	90	0.493	0.972	1.684	0.545	1.083	1.919	0.493	0.972	1.689
18	30	0.141	0.487	1.139	0.205	0.943	2.049	0.141	0.491	1.164
	60	0.261	0.692	1.364	0.323	0.907	1.751	0.261	0.692	1.385
	90	0.375	0.988	1.665	0.428	1.124	1.944	0.375	0.988	1.688
24	30	0.124	0.511	1.240	0.220	1.404	2.833	0.124	0.512	1.289
	60	0.207	0.737	1.417	0.219	0.944	1.868	0.207	0.743	1.455
	90	0.349	0.949	1.700	0.391	1.115	2.005	0.349	0.949	1.719

[a]Note: See Table 6 note.

Table 11

Proportion of Model Sum of Squares Accounted for by Contrast [a]

Dependent Variable: C_N

METHOD:	STEPWISE			BACKWARD			FORWARD		
I/D LEVEL (α):	α_p	.05	.15	α_p	.05	.15	α_p	.05	.15
CONTRAST									
$\rho_{x_i x_j}$ (L)	+	–	–	+	–	–	+	–	–
$\rho_{x_i x_j}$ (Q)	S						S		
P (L)	+L	+L	+L	+L	+L	+L	+L	+L	+L
P (Q)					S				
N (L)	+S	–S	–S	–L	–M	–M	+S	–S	–S
N (Q)	S			S	S	S	S		
$\rho_{x_i x_j}$(L) x P(Q)	S						S		
$\rho_{x_i x_j}$(Q) x P(L)				S					
$\rho_{x_i x_j}$(Q) x P(Q)				S					
$\rho_{x_i x_j}$(L) x N(L)	S						S		
P(L) x N(L)	M		S	M	M	M	M		S
P(L) x N(Q)				S	S	S			
P(Q) x N(L)	S				S	S	S		
P(Q) x N(Q)	S			S			S		
$\rho_{x_i x_j}$(L)xP(L)xN(L)				S					
$\rho_{x_i x_j}$(L)xP(Q)xN(Q)	S						S		
$\rho_{x_i x_j}$(Q)xP(L)xN(Q)	M			S			M		

[a]Note: See Table 9 note.

Table 12
Effect of Collinearity, Number of Candidate Variables and Sample Size on the Mean Values of C_N[a]

		Method								
		STEPWISE			BACKWARD			FORWARD		
		I/D Level (α)								
		α_p	.05	.15	α_p	.05	.15	α_p	.05	.15
	$\rho_{x_i x_j}$									
	0.0	0.101	0.674	2.062	0.168	1.195	3.000	0.101	0.681	2.097
	0.4	0.107	0.653	2.048	0.159	1.104	3.020	0.107	0.654	2.068
	0.8	0.102	0.640	2.012	0.171	1.174	2.985	0.102	0.645	2.035
	P									
	12	0.084	0.308	0.953	0.108	0.403	1.147	0.084	0.308	0.959
	18	0.106	0.671	2.015	0.169	0.998	2.764	0.106	0.676	2.045
	24	0.120	0.989	3.154	0.220	2.072	5.095	0.120	0.995	3.196
	N									
	30	0.095	0.708	2.253	0.223	1.854	4.310	0.095	0.713	2.293
	60	0.111	0.642	1.963	0.146	0.880	2.486	0.111	0.645	1.984
	90	0.104	0.618	1.906	0.128	0.739	2.209	0.104	0.622	1.923
P	N									
12	30	0.087	0.327	1.016	0.123	0.539	1.361	0.087	0.327	1.027
	60	0.089	0.312	0.955	0.105	0.351	1.076	0.089	0.312	0.961
	90	0.077	0.284	0.889	0.097	0.319	1.004	0.077	0.284	0.889
18	30	0.101	0.700	2.133	0.217	1.388	3.574	0.101	0.701	2.172
	60	0.120	0.685	2.004	0.167	0.879	2.508	0.120	0.696	2.040
	90	0.097	0.628	1.907	0.124	0.728	2.208	0.097	0.632	1.924
24	30	0.097	1.096	3.609	0.329	3.635	7.995	0.097	1.111	3.680
	60	0.124	0.928	2.929	0.167	1.411	3.874	0.124	0.927	2.952
	90	0.137	0.943	2.923	0.163	1.171	3.416	0.137	0.948	2.956

[a]Note: See Table 6 note.

Table 13

Proportion of Model Sum of Squares Accounted for by Contrast [a]

Dependent Variable: P_N

METHOD:	STEPWISE			BACKWARD			FORWARD		
I/D LEVEL (α):	α_p	.05	.15	α_p	.05	.15	α_p	.05	.15
CONTRAST									
$\rho_{x_i x_j}$ (L)	+M	+M	+L	+S	+S	+S	+M	+M	+L
$\rho_{x_i x_j}$ (Q)					S	S			
P (L)	+M	+L	+L	+L	+L	+L	+M	+L	+L
P (Q)		S	S		S	S		S	S
N (L)	-L	-M	-M	-L	-L	-M	-L	-M	-M
$\rho_{x_i x_j}$(L) x N(L)	S			S		S	S		
$\rho_{x_i x_j}$(L) x N(Q)	S			S			S		
$\rho_{x_i x_j}$(Q) x N(L)				S					
$\rho_{x_i x_j}$(Q)xP(L)xN(Q)	S						S		
$\rho_{x_i x_j}$(Q)xP(Q)xN(L)	S			S			S		

[a]Note: See Table 9 note.

Table 14

Effect of Collinearity, Number of Candidate Variables and Sample Size on the Mean Values of P_N[a]

	Method								
	STEPWISE			BACKWARD			FORWARD		
	I/D Level (α)								
	α_p	.05	.15	α_p	.05	.15	α_p	.05	.15
$\rho_{x_i x_j}$									
0.0	21.4	35.2	45.8	24.2	38.1	47.8	21.4	35.3	45.8
0.4	25.9	45.9	57.4	29.6	48.0	57.7	25.9	45.9	57.1
0.8	33.9	54.0	65.2	34.0	50.0	57.6	33.9	54.1	65.1
P									
12	19.1	27.9	40.3	20.3	27.8	37.8	19.1	27.6	40.2
18	27.8	46.7	58.6	30.6	46.6	57.1	27.8	46.8	58.5
24	33.1	55.7	68.0	36.7	57.6	67.3	33.1	55.7	67.8
N									
30	38.1	55.2	62.7	43.2	56.5	60.3	38.1	55.3	62.6
60	28.3	44.7	55.5	30.1	44.5	54.0	28.3	44.7	55.5
90	18.9	35.3	49.9	19.7	35.6	48.7	18.9	35.4	49.9

[a]Note: See Table 6 note.

0.15, respectively. This is contrary to what one would expect, since the number of authentic variables remains constant as P increases. This trend was reversed but remained moderate in size when $\alpha = \alpha_p$. (This is because the value of α representing a familywise level of significance of 0.15 is reduced as P increases.) The effect of P on C_A within the other algorithms was negligible to small.

In contrast, a strong positive linear relationship between C_N and P was evident for every combination of subset selection algorithm and α inclusion/deletion level. (See Tables 11 and 12) Clearly, as the number of candidate predictor variables was increased, the number of noise variables in the final subset was also increased. For example, when $\alpha = 0.15$, the mean value of C_N was increased from 1.147 when P = 12 to 5.095 when P = 24 in the BACKWARD procedure and correspondingly from 0.953 when P = 12 to 3.154 when P = 24 in the STEPWISE procedure. Unlike the case of C_A, this is exactly as one would expect since the ratio of noise to authentic variables was increased from 2 to 4 for P = 12 and 24, respectively.

A similar moderate to large positive relationship was shown to exist between P and P_N for all algorithm-α level conditions (See Table 13). Furthermore, a small quadratic relationship between P_N and P was also evident when $\alpha = 0.05$ and 0.15 within each of the algorithms. The increase in the percentage of the final subset that was noise was in the range of 10% to 20% as P was increased from 12 to 18 and from 5% to 10% as P was further increased to 24 (See Table 14).

Sample Size. With respect to the effect of sample size on the number of authentic variables in the final subset, Table 9 shows that sample size had a medium to large positive effect on C_A in the STEPWISE and FORWARD procedures for all α levels. On the other hand, the relationship between N and C_A was not stable over increasing α in the BACKWARD procedure. Here, the linear effect of N was large and positive when $\alpha = \alpha_p$, but was reduced to a negligible effect and a small effect, respectively, when $\alpha = 0.05$ and 0.15. As well, a small quadratic effect due to N was evident in the BACKWARD procedure for each value of α. Nonetheless, in all three subset selection procedures the relationship between N and C_A was shown to be positive.

The positive linear effect of N on C_A is evident in Table 10 within the STEPWISE and FORWARD procedures. It is clear that increasing the sample size enables these two subset selection procedures to detect more authentic

predictor variables. However, though the linear effect of N was shown to be moderate to large in size, the actual increase in C_A was shown to be small. For example even when $\alpha = 0.15$, the mean value of C_A was only increased from 1.164 when N = 30 to 1.683 when N = 90 in the STEPWISE procedure. In fact, in all cases, the mean value of C_A remained well below the actual number of authentic variables present. This may be partially due to a $\rho_{x_i x_j}$ X N interaction.

In contrast to the analysis of C_A, sample size was shown to have a small linear effect on C_N in the STEPWISE and FORWARD procedures, and a moderate to large linear effect on C_N in the BACKWARD procedure (See Table 11). For the former procedures, a small linear effect due to N was evident for all levels of α, but the sign of the relationship was changed from positive when $\alpha = \alpha_p$ to negative when $\alpha = 0.05$ and 0.15. In the BACKWARD procedure, a large to medium linear effect due to N was evident for each level of α and the sign remained negative. This effect is evident in Table 12 where the mean value of C_N was decreased from 4.310 when N = 30 to 2.209 when N = 90 at $\alpha = 0.15$. Furthermore, a small quadratic effect in N was also present for each α-level for this procedure. A similar though less dramatic effect was evident within the STEPWISE and FORWARD procedures. Thus, it is clear that an increase in sample size aids the reduction of the number of noise present in the final subset.

As expected, a similar effect was shown to exist for sample size when the dependent variable was the percentage of noise in the final subset. A moderate to strong negative linear effect was evident for all algorithm-α level conditions (See Tables 13 and 14). However, unlike in the analysis of C_A and C_N, the strength and direction of the effect was fairly consistent within all subset selection methods. Generally, each increase in N was able to reduce the percentage of noise variables by about 5% to 10%.

Two-way Interactions

In the analysis of two-way interactions, none of the variables displayed any evidence of a consistent $\rho_{x_i x_j}$ X P interaction (See Tables 9, 11 and 13) However, a small to moderate interaction effect between $\rho_{x_i x_j}$ and N on C_A was evident in Table 9. A small $\rho_{x_i x_j}$ linear X N linear effect was shown to exist in the STEPWISE and FORWARD procedures and a medium $\rho_{x_i x_j}$ linear X N linear effect was shown to exist in the BACKWARD procedure for each level of

α. As well, a small $\rho_{x_i x_j}$ quadratic X N linear effect was shown to be present when $\alpha = 0.05$ and 0.15 within each of the algorithms.

Table 10 shows that for the STEPWISE and FORWARD procedures, sample size maintained its positive effect on C_A within each value of $\rho_{x_i x_j}$. Increasing N seems to be most effective when $\rho_{x_i x_j} = 0.0$. Here, for example when $\alpha = 0.15$ and $\rho_{x_i x_j} = 0.0$, the mean value of C_A was increased from 1.556 when N = 30 to 2.727 when N = 90 in the STEPWISE procedure. Comparing these results with the corresponding values when $\rho_{x_i x_j} = 0.4$ [1.095 (N = 30) to 1.305 (N = 90)] and $\rho_{x_i x_j} = 0.8$ [0.840 (N = 30) to 1.017 (N = 90)] indicates that the effectiveness of increased sample size in detecting authentic variables was reduced by collinearity. This is not surprising, since the sample sizes investigated in this study were chosen in order to detect a medium effect size and, when the value of $\rho_{x_i x_j}$ was increased from 0.0 to 0.8, the effect size present in the data was correspondingly reduced to a small effect size.

For the BACKWARD procedure, the interaction between collinearity and sample size was complex (See Table 10). When $\rho_{x_i x_j} = 0.0$, increasing N positively affected C_A. This remained true as the value of $\rho_{x_i x_j}$ increased and $\alpha = \alpha_p$. However, when $\alpha = 0.05$ and 0.15, as $\rho_{x_i x_j}$ was increased the effect of N became negative. It must be recalled that the BACKWARD procedure is unlike either the STEPWISE and FORWARD procedures in that it begins with all of the candidate predictor variables in the model. It may be that this, combined with the reduced effect size (due to increased collinearity) renders increasing the sample size completely ineffective.

A second two-way interaction was present in the analysis of C_A. Within the BACKWARD procedure, a small to moderate interaction effect on C_A between P and N was evident in Table 9. An examination of Table 10 indicated that when P = 12 N seems to have a positive effect on C_A and as P increases the effect of N seems to become more quadratic and negative in nature when α = 0.05 and 0.15. It may be that like increasing collinearity, increasing the number of candidate variables along with a liberal α value interferes with the effectiveness of increasing N in positively influencing the number of authentic variables in the final subset.

Similarly, a P X N interaction effect was also evident in the analysis of C_N within the BACKWARD procedure (See Table 11). Here, a medium P linear X N linear interaction effect was shown to be present for each value of α. Additionally, a number of small effect values were associated with the other

interaction contrasts which comprise the P X N interaction component. Table 12 showed that as P was increased, the number of noise variables in the final subset was also increased. However, the negative effect of N also seems to have been increased such that when P = 24, increasing N from 30 to 60 reduced the mean number of noise variables by about half . While the decrease in the number of noise variables due to increased sample size was reduced when P = 12, the actual mean value of C_N was also much less at this value of P. For example, when $\alpha = 0.15$ and P = 12, the mean value of C_N was reduced from 1.361 when N = 30 to 1.004 when N = 90, and from 7.995 when N = 30 to 3.416 when N = 90 and P = 24.

Three-way Interactions

The analysis of C_N showed some small to medium r^2 values among the three-way contrasts for each of the subset selection methods at $\alpha = \alpha_p$ (See Table 11). However, these interaction effects were not stable over increasing α and were reduced to a negligible size when α = 0.05 and 0.15. A similar occurrence of three-way interactions is evident in the analysis of P_N (See Table 13).

Estimates of the Population Coefficient of Multiple Determination

Tables 15, 16 and 17 illustrate the effect of collinearity among the authentic predictor variables, number of candidate variables and sample size within each combination of algorithm and inclusion/deletion level on the mean values of R^2, R_k^2, and R_p^2, respectively (see Appendix B for standard errors).

A comparison of the population ρ_{yxi}^2 value with the mean values of R^2 and R_k^2 showed that these two estimates tend to be inflated (See Tables 15 and 16). However, when $\alpha = 0.15/p$, $\rho_{x_i x_j} = 0$ and the sample size is 90, both of these estimates did very well. Under these conditions, $\rho_{yxi}^2 = .130$, and R^2 took on the values .130, .131, and .135 for the STEPWISE and FORWARD procedures and the values .135, .144, and .143 for the BACKWARD procedure for P = 12, 18 and 24, respectively. Similarly, R_k^2 took on the values .117, .118 and .122 for the STEPWISE and FORWARD procedures and the values .121, .129, and .129 for the BACKWARD procedure for P = 12, 18 and 24, respectively. As would be expected, the results for R^2 were slightly higher than those for R_k^2. On the other hand, Table 17 showed R_p^2 to be an extremely conservative estimate of ρ_{yxi}^2. Only when ρ_{yxi}^2 reached .043 and .026 did the mean value of R_p^2 occasionally exceed ρ_{yxi}^2.

A comparison of the mean values of R^2, R_k^2 and R_p^2 for each subset selection algorithm indicated that the results for the STEPWISE and FORWARD procedures were very similar (See Tables 15, 16 and 17). However, the values for the FORWARD procedure were occasionally slightly larger than those of the STEPWISE procedure when α was large (0.05 or 0.15). On the other hand, the results from the BACKWARD procedure were generally greater than those of the STEPWISE and FORWARD procedures for each of the three dependent variables. As a result, the BACKWARD procedure tended to produce slightly more inflated values of R^2 and R_k^2 and slightly less conservative values of R_p^2. However, the differences tended to be ameliorated by increases in sample size.

As expected, within any of the algorithms, an increase in the inclusion/deletion level generally increased the mean value of R^2, R_k^2 and R_p^2. Consequently, the least inflated values of R^2 and R_k^2 were to be found when α was small and the least conservative values of R_p^2 were to be found when α was large. (See Tables 15, 16 and 17)

Table 15

Effect of Collinearity ($\rho_{x_i x_j}$), Number of Candidate Variables (P), and Sample Size (N) on R^2

				Method								
				STEPWISE			BACKWARD			FORWARD		
				I/D Level (α)								
$\rho_{x_i x_j}$	P	N	ρ_{yxi}^2	.15/p	.05	.15	.15/p	.05	.15	.15/p	.05	.15
0.0	12	30	.130	.280	.290	.346	.317	.328	.375	.280	.290	.347
		60	.130	.170	.176	.215	.178	.184	.223	.170	.176	.215
		90	.130	.130	.150	.181	.135	.153	.185	.130	.150	.181
	18	30	.130	.315	.317	.430	.352	.408	.526	.315	.318	.433
		60	.130	.192	.207	.268	.212	.222	.287	.192	.208	.269
		90	.130	.131	.158	.216	.144	.166	.225	.131	.158	.217
	24	30	.130	.300	.361	.531	.373	.550	.719	.300	.362	.530
		60	.130	.173	.222	.323	.188	.250	.354	.173	.222	.324
		90	.130	.135	.175	.255	.143	.189	.271	.135	.175	.257
0.4	12	30	.043	.286	.259	.286	.295	.287	.324	.286	.259	.288
		60	.043	.151	.143	.156	.158	.146	.166	.151	.143	.156
		90	.043	.109	.100	.120	.111	.104	.127	.109	.100	.120
	18	30	.043	.300	.300	.387	.370	.380	.484	.300	.300	.389
		60	.043	.168	.177	.226	.176	.195	.248	.168	.177	.228
		90	.043	.123	.121	.159	.124	.127	.169	.123	.121	.159
	24	30	.043	.317	.338	.518	.383	.526	.725	.317	.339	.519
		60	.043	.174	.172	.254	.183	.197	.291	.174	.172	.254
		90	.043	.124	.130	.187	.126	.139	.202	.124	.130	.187
0.8	12	30	.026	.298	.242	.273	.327	.288	.318	.298	.242	.274
		60	.026	.163	.146	.144	.163	.155	.161	.163	.146	.144
		90	.026	.103	.100	.106	.107	.101	.115	.103	.100	.106
	18	30	.026	.289	.288	.363	.371	.391	.480	.289	.288	.364
		60	.026.	.164	.149	.189	.169	.172	.224	.164	.149	.190
		90	.026	.109	.109	.135	.112	.111	.151	.109	.110	.135
	24	30	.026	.326	.340	.483	.442	.569	.717	.326	.343	.485
		60	.026	.168	.172	.237	.182	.202	.283	.168	.172	.239
		90	.026	.123	.130	.180	.127	.144	.198	.123	.130	.180

Table 16

Effect of Collinearity ($\rho_{x_i x_j}$), Number of Candidate Variables (P), and Sample Size (N) on R_k^2

				Method								
				STEPWISE			BACKWARD			FORWARD		
				I/D Level (α)								
$\rho_{x_i x_j}$	P	N	ρ_{yxi}^2	.15/p	.05	.15	.15/p	.05	.15	.15/p	.05	.15
0.0	12	30	.130	.251	.250	.280	.283	.278	.298	.251	.250	.280
		60	.130	.152	.151	.173	.158	.157	.177	.152	.151	.173
		90	.130	.117	.129	.147	.121	.131	.149	.117	.129	.147
	18	30	.130	.287	.276	.356	.318	.347	.418	.287	.276	.357
		60	.130	.174	.179	.217	.192	.190	.228	.174	.180	.217
		90	.130	.118	.135	.173	.129	.141	.179	.118	.135	.174
	24	30	.130	.274	.315	.448	.340	.468	.588	.274	.315	.446
		60	.130	.157	.190	.260	.170	.211	.278	.157	.191	.261
		90	.130	.122	.150	.206	.129	.161	.214	.122	.151	.206
0.4	12	30	.043	.258	.224	.229	.265	.243	.249	.258	.224	.231
		60	.043	.136	.124	.124	.141	.125	.130	.136	.124	.124
		90	.043	.098	.087	.097	.100	.089	.100	.098	.087	.097
	18	30	.043	.272	.260	.317	.334	.322	.378	.272	.260	.317
		60	.043	.152	.152	.182	.159	.166	.194	.152	.152	.182
		90	.043	.111	.104	.127	.112	.108	.133	.111	.105	.128
	24	30	.043	.291	.294	.434	.346	.449	.584	.291	.295	.432
		60	.043	.158	.147	.201	.165	.166	.222	.158	.147	.201
		90	.043	.113	.112	.148	.114	.118	.157	.113	.112	.148
0.8	12	30	.026	.269	.207	.218	.289	.241	.241	.269	.207	.219
		60	.026	.146	.125	.114	.145	.131	.123	.146	.125	.114
		90	.026	.093	.086	.084	.095	.086	.089	.093	.086	.084
	18	30	.026	.263	.248	.295	.334	.330	.372	.263	.248	.296
		60	.026.	.147	.127	.149	.150	.144	.169	.147	.127	.149
		90	.026	.098	.093	.106	.100	.093	.115	.098	.093	.106
	24	30	.026	.300	.297	.403	.401	.485	.584	.300	.299	.403
		60	.026	.153	.146	.186	.164	.166	.214	.153	.146	.187
		90	.026	.111	.112	.142	.115	.121	.152	.111	.112	.142

Table 17

Effect of Collinearity ($\rho_{x_i x_j}$), Number of Candidate Variables (P), and Sample Size (N) on R_p^2

				Method								
				STEPWISE			BACKWARD			FORWARD		
				I/D Level (α)								
$\rho_{x_i x_j}$	P	N	ρ_{yxi}^2	.15/p	.05	.15	.15/p	.05	.15	.15/p	.05	.15
0.0	12	30	.130	.008	.033	.061	.031	.053	.084	.008	.033	.062
		60	.130	.023	.032	.059	.028	.037	.065	.023	.032	.060
		90	.130	.023	.044	.067	.027	.046	.071	.023	.044	.067
	18	30	.130	.000	.008	.037	.002	.032	.095	.000	.008	.038
		60	.130	.011	.024	.047	.020	.029	.055	.011	.024	.049
		90	.130	.009	.020	.051	.012	.024	.057	.009	.020	.052
	24	30	.130	.000	.001	.018	.019	.046	.125	.000	.001	.018
		60	.130	.000	.006	.027	.000	.011	.043	.000	.006	.027
		90	.130	.003	.010	.039	.003	.014	.047	.003	.010	.039
0.4	12	30	.043	.020	.017	.031	.019	.033	.046	.020	.017	.032
		60	.043	.009	.011	.018	.010	.012	.023	.009	.011	.019
		90	.043	.008	.010	.021	.010	.012	.025	.008	.010	.021
	18	30	.043	.000	.006	.024	.000	.026	.058	.000	.006	.023
		60	.043	.001	.008	.024	.001	.012	.032	.001	.008	.024
		90	.043	.004	.007	.018	.005	.009	.022	.004	.007	.019
	24	30	.043	.000	.000	.012	.011	.034	.108	.000	.000	.013
		60	.043	.000	.001	.008	.000	.002	.015	.000	.001	.009
		90	.043	.001	.003	.010	.001	.004	.014	.001	.003	.010
0.8	12	30	.026	.013	.011	.025	.028	.027	.043	.013	.011	.026
		60	.026	.017	.014	.018	.015	.017	.023	.017	.014	.018
		90	.026	.006	.010	.014	.007	.009	.018	.006	.010	.014
	18	30	.026	.000	.000	.013	.024	.036	.066	.000	.000	.013
		60	.026.	.001	.003	.009	.002	.005	.019	.001	.003	.009
		90	.026	.001	.002	.007	.002	.002	.011	.001	.002	.007
	24	30	.026	.000	.001	.010	.040	.067	.141	.000	.001	.006
		60	.026	.000	.001	.007	.000	.003	.015	.000	.001	.007
		90	.026	.000	.001	.007	.000	.002	.010	.000	.001	.007

Examination of Tables 15, 16, and 17 showed that trends due to collinearity, sample size and number of candidate variables were present in the data. As a result, trend analyses were performed.

Trend Analysis of $\rho_{x_i x_j}$, P and N within Algorithm and Inclusion/Deletion Level

Tables 18, 19 and 20 contain the proportion of the model sums of squares accounted for by each contrast (r^2) within each combination of method and α for the dependent variables R^2, R_k^2, and R_p^2, respectively. In general, the r^2 values indicated that the same pattern of trends were affecting both R^2 and R_k^2, whereas a different pattern of trends was affecting R_p^2. Figures 2 (a), 2 (b), 3 (a), and 3 (b) show the effect of increasing collinearity, number of candidate variables and sample size on the mean values of R_k^2 and R_p^2, for the STEPWISE and BACKWARD procedures, respectively, at $\alpha = 0.15$.

Main Effects

Collinearity. With regard to the effects of correlation among predictor variables, there was some evidence to show that collinearity negatively affects R^2, and R_k^2. A small negative linear effect due to collinearity was evident for each algorithm when $\alpha = 0.05$ and 0.15 (See Tables 18 and 19). Specifically, for a given combination of P and N, the mean values of R^2 and R_k^2 are reduced as the degree of collinearity is increased (See Figures 2 (a) and 3 (a)). However, it is important to note that the strength of the trend does not reflect the degree to which the population value of ρ_{yxi}^2 is reduced as collinearity is increased.

Further, the evidence for this negative effect was even stronger with regard to R_p^2 (See Table 20). In the STEPWISE and FORWARD procedures the negative linear effect of $\rho_{x_i x_j}$ was increased from a small value when $\alpha = 0.15/p$ to a large value when $\alpha = 0.05$ and 0.15. On the other hand, in the BACKWARD procedure the effect was of moderate value for $\alpha = 0.05$ and 0.15. Additionally, all three procedures displayed a small quadratic effect due to $\rho_{x_i x_j}$ when $\alpha = 0.05$ and 0.15. This effect was of moderate value when $\alpha = 0.15/p$ in the BACKWARD procedure. Figure 2 (b) shows that the mean value of R_p^2 is reduced within the STEPWISE algorithm as the degree of collinearity is increased and is most obvious as the value of $\rho_{x_i x_j}$ is increased from 0 to .4.

Table 18

Proportion of Model Sum of Squares Accounted for by Contrast [a]

Dependent Variable: R^2

METHOD:	STEPWISE			BACKWARD			FORWARD		
I/D LEVEL (a):	.15/p	.05	.15	.15/p	.05	.15	.15/p	.05	.15
CONTRAST									
$\rho_{x_i x_j}$ (L)	-	-S	-S	-	-S	-S	-	-S	-S
P (L)	+	+S	+M	+S	+M	+M	+	+S	+M
N (L)	-L	-L	-L	-L	-L	-L	-L	-L	-L
N (Q)	S	S	S	M	S	S	S	S	S
P(L) x N(L)		S	S	S	S	S		S	S
P(L) x N(Q)					S	S			

[a]Note: See Table 9 note.

Table 19

Proportion of Model Sum of Squares Accounted for by Contrast[a]

Dependent Variable: R_K^2

METHOD:	STEPWISE			BACKWARD			FORWARD		
I/D LEVEL (a):	.15/p	.05	.15	.15/p	.05	.15	.15/p	.05	.15
CONTRAST									
$\rho_{x_i x_j}$ (L)	-	-S	-S	-	-S	-S	-	-S	-S
P (L)	+S	+S	+M	+S	+M	+M	+S	+S	+M
N (L)	-L	-L	-L	-L	-L	-L	-L	-L	-L
N (Q)	S	S	S	M	S	S	S	S	S
P(L) x N(L)		S	S	S	S	S		S	S
P(L) x N(Q)					S	S			

[a]Note: See Table 9 note.

Table 20

Proportion of Model Sum of Squares Accounted for by Contrast[a]

Dependent Variable: R_p^2

METHOD:	STEPWISE			BACKWARD			FORWARD		
I/D LEVEL (a):	.15/p	.05	.15	.15/p	.05	.15	.15/p	.05	.15
CONTRAST									
$\rho_{x_i x_j}$ (L)	-S	-L	-L	-	-M	-M	-S	-L	-L
$\rho_{x_i x_j}$ (Q)		S	S	M	S	S		S	S
P (L)	-L	-L	-M	-M	-S	+S	-L	-L	-M
P (Q)	S	S		S	S		S	S	
N (L)	+	+S	+	-M	-L	-L	+	+S	+
N (Q)				S	M	M			
$\rho_{x_i x_j}$ (L) x P(L)		S	S	S	S	S		S	S
$\rho_{x_i x_j}$ (Q) x P(L)		S	S					S	S
$\rho_{x_i x_j}$ (L) x N(L)	S	S	S	M	S	S	S	S	S
$\rho_{x_i x_j}$ (L) x N(Q)				S					
$\rho_{x_i x_j}$ (Q) x N(L)	S	S		S			S	S	S
$\rho_{x_i x_j}$ (Q) x N(Q)	S						S		
P(L) x N(L)			S	S	S	M			S
P(L) x N(Q)				S	S	S			
P(Q) x N(L)	S			S	S		S		
P(Q) x N(Q)		S		S	S			S	
$\rho_{x_i x_j}$(L)xP(L)xN(L)	S						S		
$\rho_{x_i x_j}$(L)xP(L)xN(Q)					S				
$\rho_{x_i x_j}$(L)xP(Q)xN(Q)		S		S				S	
$\rho_{x_i x_j}$(Q)xP(L)xN(L)	S			S	S		S		
$\rho_{x_i x_j}$(Q)xP(L)xN(Q)	S			S			S		

[a]Note: See Table 9 note.

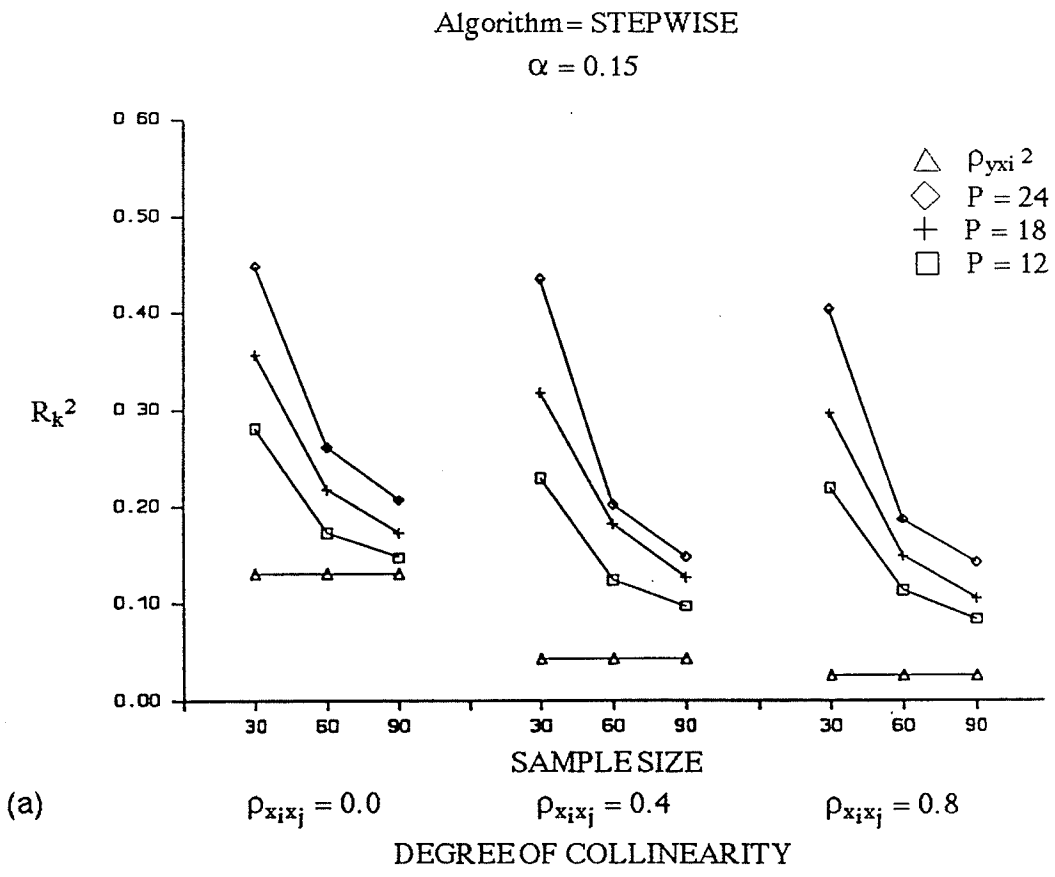

Figure 2. The effect of sample size (N), number of candidate variables (P) and collinearity ($\rho_{x_ix_j}$) on the mean value of three measures of the population squared coefficient of determination, ρ_{yx}^2, within the STEPWISE procedure. (a) The sample squared coefficient of determination adjusted by k, R_k^2. (b) The sample squared coefficient of determination adjusted by P, R_p^2.
Note: The inclusion/deletion level of significance was set to 0.15.

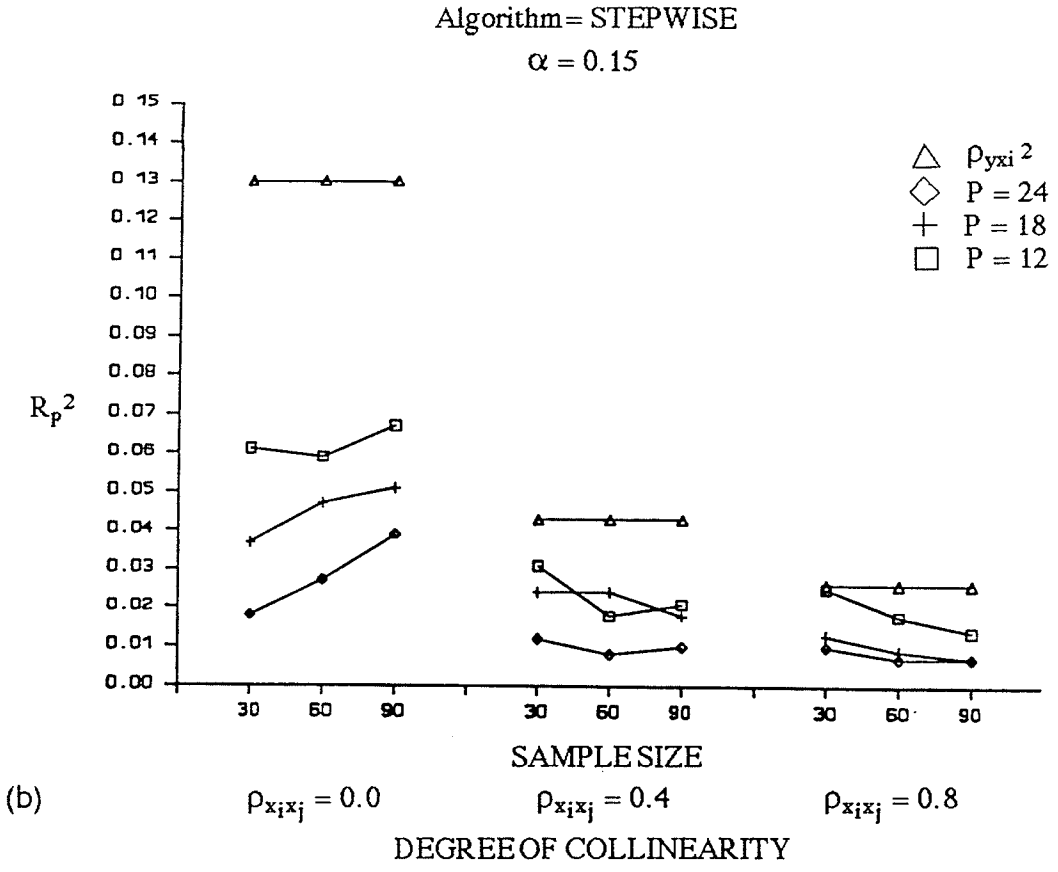

Figure 2. (continued).

Number of Candidate Variables. A small positive linear effect was present for the number of candidate variables (P) and R^2 and R_k^2 (See Tables 18 and 19). Within the STEPWISE and FORWARD procedures, the effect size rose from a negligible value for $\alpha = 0.15/p$ to a small value for $\alpha = 0.05$ and a medium value for $\alpha = 0.15$. The effect of P was slightly increased in the the BACKWARD procedure.

On the other hand, the linear effect of P on R_p^2 was strongly negative (See Table 20). In the STEPWISE and FORWARD procedures the linear contrast in P accounted for a large proportion of the model sums of squares for $\alpha = 0.15/p$ and $\alpha = 0.05$ and a medium proportion for $\alpha = 0.15$. The strength of this effect was reduced somewhat in the BACKWARD procedure and furthermore, when $\alpha = 0.15$, the sign of the effect was changed from negative to positive. Also, a small quadratic effect was present when $\alpha = 0.15/p$ and 0.05 in each of the algorithms.

For any value of $\rho_{x_i x_j}$ and N, an increase in the number of candidate variables results in a higher mean value of R_k^2 as surmised by Cohen and Cohen (1983, p. 107) (See Figures 2 (a) and 3 (a)). On the other hand, the negative effect of increasing P on R_p^2 within the STEPWISE procedure is most clear when $\rho_{x_i x_j} = 0$ (See Figure 2 (b)). However, as collinearity is increased, the effect of P becomes less evident.

Sample Size. For all combinations of algorithm and α, sample size (N) had a large effect on R^2 and R_k^2 (See Tables 18 and 19). Both the linear and quadratic contrasts accounted for a large and a small proportion of the model sums of squares, respectively. Since the sign of the linear contrast remained negative for each combination of algorithm and α, it is clear that increasing sample size reduces the inflation of these two estimates. However, the presence of a small quadratic component suggests that the effect of increasing sample size is not constant.

The analysis of trends for R_p^2 showed that sample size was not a significant factor in the STEPWISE and FORWARD procedures (See Table 20). That is, a non-neglible linear effect due to N was present only when $\alpha = 0.05$. However, in the BACKWARD procedure the effect of sample size on R_p^2 was like that of R^2 and R_k^2. Specifically, a medium to large negative linear effect was evident in the BACKWARD procedure for all levels of α. As well, a small to medium quadratic effect was shown to be present for each level of α.

While the values of R^2 and R_k^2 remain above the population value of ρ_{yxi}^2, there is a negative linear effect of increasing sample size (See Figures 2 (a) and 3 (a)). Also evident is that the effect of increasing sample size is not constant. The decrease in value of R_k^2 (and R^2) when the sample size is increased from 60 to 90 is not as great as the decrease when the sample size is increased from 30 to 60. This suggests that a point of diminishing returns may be reached in increasing the sample size. Figure 2 (b) shows that sample size has no consistent effect on R_p^2 in the STEPWISE procedure. However, the negative quadratic effect of sample size on R_p^2 in the BACKWARD procedure is evident in Figure 3 (b).

Two-Way Interactions

There was no evidence of a collinearity by number of candidate variables interaction ($\rho_{x_ix_j}$ X P) for the dependent variables R^2 and R_k^2 (see Tables 18 and 19); however, such an effect was present for R_p^2. (See Table 20) A small $\rho_{x_ix_j}$ linear X P linear effect was present in the STEPWISE and FORWARD procedures ($\alpha = 0.05$ and $\alpha = 0.15$) and for all levels of α in the BACKWARD procedure. Furthermore, a small $\rho_{x_ix_j}$ quadratic X P linear effect was present for $\alpha = 0.05$ and $\alpha = 0.15$ in the STEPWISE and FORWARD procedures. Figure 3 (b) illustrates this relationship in the BACKWARD procedure for $\alpha = 0.15$. The differences due to increasing P are more evident when $\rho_{x_ix_j} = 0$ than when $\rho_{x_ix_j} = .4$ and .8, suggesting that collinearity among the authentic predictor variables reduces the effect of P on R_p^2.

The effect of sample size also differed for the three levels of collinearity with respect to R_p^2. Table 20 shows that a small to moderate $\rho_{x_ix_j}$ linear X N linear effect was present for each algorithm-α level condition. Furthermore, several other small effects were present among the remaining $\rho_{x_ix_j}$ X N contrasts. This relationship may be deduced from Figure 3 (b) in that when $\rho_{x_ix_j} = 0.0$ the quadratic curve due to increasing N is most pronounced; however, as $\rho_{x_ix_j}$ is increased, the response to N becomes increasingly linear. Again, there was no evidence of a $\rho_{x_ix_j}$ X N interaction for R^2 and R_k^2 (See Tables 18 and 19).

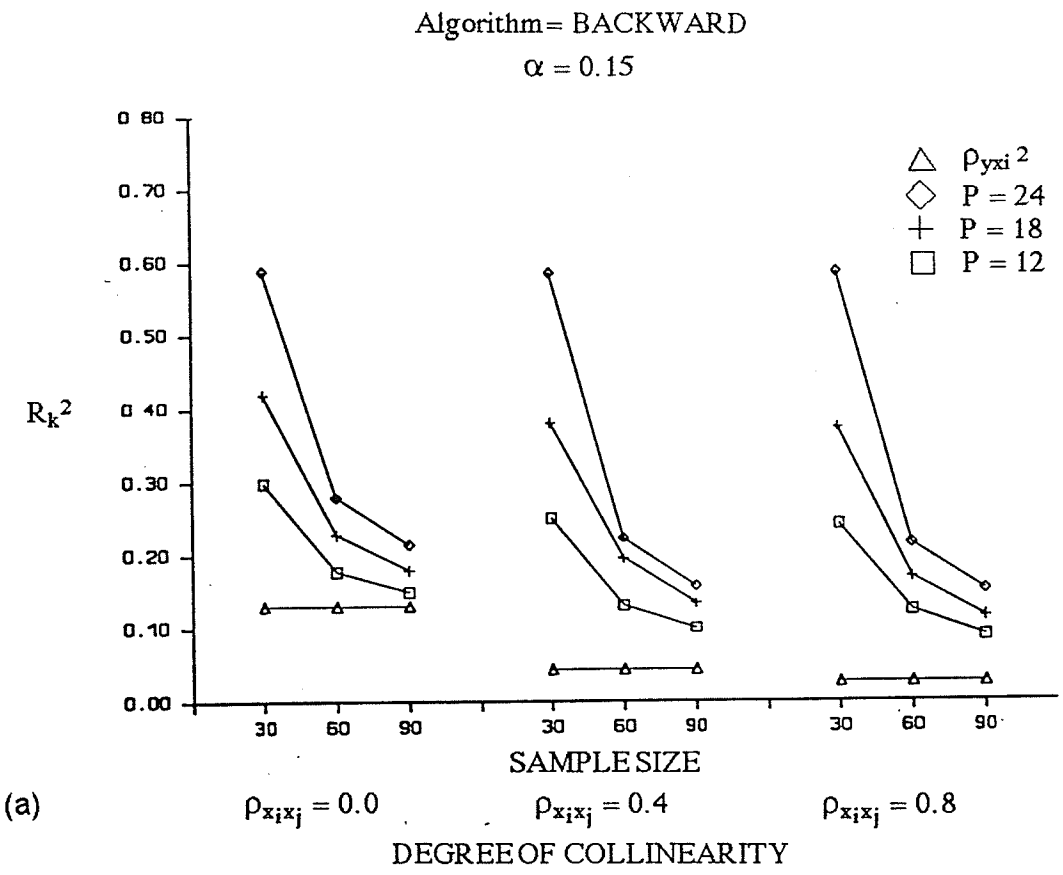

Figure 3. The effect of sample size (N), number of candidate variables (P) and collinearity ($\rho_{x_ix_j}$) on the mean value of three measures of the population squared coefficient of determination, ρ_{yx}^2, within the BACKWARD procedure. (a) The sample squared coefficient of determination adjusted by k, R_k^2. (b) The sample squared coefficient of determination adjusted by P, R_p^2.
Note: The inclusion/deletion level of significance was set to 0.15.

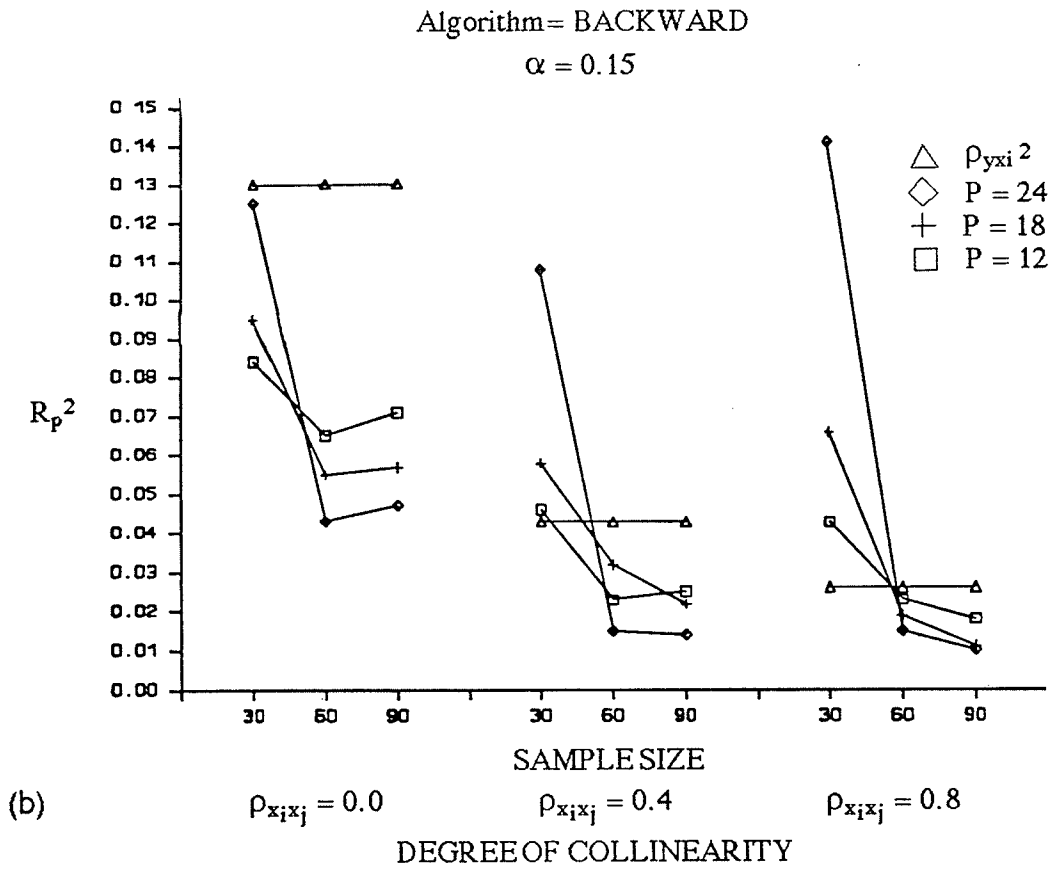

Figure 3. (continued).

Also evident were the P X N interaction effects for R^2 and R_k^2 (See Tables 18 and 19). In the STEPWISE and FORWARD procedures, a small P linear X N linear effect was present when α = 0.05 and 0.15, while in the BACKWARD procedure, a small effect was present for each level of α. Also, a small P linear X N quadratic effect was present for α = 0.05 and 0.15 in the BACKWARD procedure. This interaction is illustrated in Figure 3 (a) for the BACKWARD procedure. Note that increasing P has the greatest inflationary effect on R^2 and R_k^2 when N=30; as N increases the effect of P, though still evident, is decreased.

This interaction effect was also shown to exist for R_p^2 within the BACKWARD procedure (See Table 20). Here, a minimum small P linear X N linear effect and a P linear X N quadratic effect were consistently present for each level of α. As well, both a small P quadratic X N linear effect and a small P quadratic X N quadratic effect were present when α = 0.15/p and 0.05. Within the STEPWISE and FORWARD procedures a few small P X N interaction effects were present, but none were consistent over increasing α. Figure 3 (b) shows that when N = 30, the mean value of R_p^2 increases with increasing P. However, when N = 60 and 90, the relationship is reversed and the mean value of R_p^2 is decreased with increasing P. According to Cohen and Cohen (1983, p. 106), R_p^2 should decrease as the ratio of P/N increases. Thus, for a given N, R_p^2 should decrease as P increases. Why this is not the case when N = 30 is not clear.

Three-Way Interactions

The three-way interaction effects were negligible with regard to R^2 and R_k^2 (See Tables 18 and 19). However, a number of small three-way interaction effects were present for R_p^2 (See Table 20). However, these values were evident only when α = 0.15/p and 0.05; when α = 0.15, no non-negligible values remained.

Discussion and Conclusions

The results of this study show that collinearity, number of candidate variables and sample size affect the outcome of 'best' subset selection algorithms. In both the analysis of the selection of authentic variables in the presense of noise and in the comparison of estimates of $\rho_{yx_j}{}^2$, these effects were shown to be fairly consistent in magnitude and direction across the STEPWISE and FORWARD algorithms and level of significance for inclusion and deletion of variables. However, these effects often differed within the BACKWARD algorithm as compared to the others.

Comparing the analysis of C_A, C_N and P_N across algorithms it was shown that the pattern and strength of trends within the STEPWISE and FORWARD procedures were the same. Within the BACKWARD procedure, the pattern of trends was similar to the other procedures in the analysis of C_N and P_N. Any differences that existed here lay in the magnitude of the individual trends. However, with respect to C_A, the effect of sample size was greatly changed by both the level of collinearity and the number of candidate variables in the BACKWARD procedure.

Overall, the mean values of C_A, C_N, and P_N obtained by the STEPWISE and FORWARD procedures were extremely close (See Table 21). However, the average final subset obtained by the BACKWARD procedure contained both more authentic and noise variables than that of the STEPWISE and FORWARD procedures and, the proportion of noise variables was generally greater except when $\alpha = 0.15$ and $\rho_{x_i x_j}$ was nonzero.

Within an algorithm, increasing the level of α always increased the mean value of C_A, C_N and P_N. Unfortunately, the rate at which C_A was increased was less than the rate at which C_N was increased so that there were always a greater percentage of noise variables in the final subset when $\alpha = 0.15$ compared to when $\alpha = 0.15/p$.

Table 21

Effect of Method and Inclusion/Deletion Level on the Mean Values of C_A, C_N, and P_N

	Method								
	STEPWISE			BACKWARD			FORWARD		
	I/D Level (α)								
	.15/p	.05	.15	.15/p	.05	.15	.15/p	.05	.15
Vars									
C_A	0.272	0.725	1.415	0.328	1.010	1.965	0.272	0.726	1.437
C_N	0.103	0.656	2.041	0.166	1.158	3.002	0.103	0.660	2.067
P_N	26.3	44.4	56.0	28.7	45.1	54.3	26.3	44.4	55.9

With respect to the mean number of authentic variables in the final subset, it was shown that increased collinearity negatively affected C_A. Conversely, increased sample size was shown to have a positive effect, especially when collinearity was negligible. In the STEPWISE and FORWARD procedures, the effect of N remained positive even when collinearity was present, but in the BACKWARD procedure increased collinearity altered the effectiveness of sample size when α = 0.05 and 0.15. Interestingly, the mean value of C_A was shown to be relatively unaffected by the number of candidate predictor variables in the STEPWISE and FORWARD procedures. However, within the BACKWARD procedure it was shown that increasing P may positively affect C_A. However it is important to note that in the BACKWARD procedure, the number of candidate variables was shown to affect the ability of increased sample size to increase the mean value of C_A when α = 0.05 and 0.15. This would suggest that a liberal inclusion/deletion level combined with collinearity or many candidate predictor variables may render the BACKWARD procedures performance unreliable.

Analysis of C_N, showed that the mean number of noise variables in the final subset was strongly positively affected by the number of candidate predictor variables within all algorithms. Additionally, sample size was shown to have a small negative effect in the STEPWISE and FORWARD procedures and a moderate to large negative effect in the BACKWARD procedure. Also within the BACKWARD procedure, a P X N interaction effect was shown to be present. This suggests that while increasing P increases the mean value of C_N, the negative effect of increasing N is also increased somewhat thus allowing some measure of control over the number of noise variables present in the final subset.

In the STEPWISE and FORWARD procedures P_N was affected about equally by $\rho_{x_i x_j}$, P and N. Increases in collinearity and number of candidate variables positively affected P_N while increased sample size negatively affected P_N. In the BACKWARD procedure the direction of each of these effects was the same but the strength of the collinearity effect was reduced to small. Interactions were not a factor here.

Recall that Flack and Chang (1987) examined the behaviour of the STEPWISE algorithm and an all-possible subsets algorithm for three sample size conditions (N=10, 20, 40), three number of candidate variable conditions

(P=10, 20, 40) and for three values of a serial correlation coefficient (ρ=.0, .3, .5). The default level of significance for the SAS STEPWISE procedure (α = .15) was used and only the results for ρ = 0.3 were presented for the STEPWISE procedure. The effect of these parameters on three dependent variables was studied: 1) the frequency distribution of the number of authentic variables selected from a set containing both authentic and noise candidate variables, 2) the proportion of the selected variables that were noise, and 3) adjusted R^2 statistic [See Darlington, 1990, p. 121 for the definition of the adjusted R^2 used by SAS(1985)].

Flack and Chang (1987) showed that sample size had a strong positive effect on the number of authentic variables selected. When N = 40 the percentage of 'best' subsets with three authentic variables (the number of authentic variables in their study) ranged in value from 28% to 34% and the percentage with two authentic variables ranged from 46% to 56%. Conversely, when N = 10, 0% to 4% of the samples found three authentic and two authentic variables were found in only 16% to 24% of the 'best' subset models. Additionally, increasing the number of candidate variables was found to have a negative effect on the number of authentic variables selected. The mean number of authentic variables found found by Flack and Chang (1987) [based upon their percentage frequency table] is compared to the STEPWISE results from the presen study, when $\alpha = 0.15$ and $\rho_{x_i x_j} = 0.0$, 0.4 and 0.8 in Table 22. (Recall that the number of authentic predictor variables was six compared to three in the Flack and Chang (1987) study.) Considering that the value of $\rho_{yx_i}^2$ was considerably less in the present study, especially when $\rho_{x_i x_j}$ was nonzero, the results show that the increase in the N to P ratio was effective in increasing the number of authentic variables selected. However, at all times the average number of authentic variables remained at less than half of the available authentic variables.

Table 22

A Comparison of Flack and Chang's Results to Those of the Present Study[a]

Flack and Chang (1987)				Present Study					
							$\rho_{x_i x_j}$		
							0.0	0.4	0.8
P	N	N/P	C_A	P	N	N/P	C_A	C_A	C_A
10	10	1.00	1.02	12	30	2.50	1.516	1.032	0.788
	20	2.00	1.66		60	5.00	2.192	1.136	0.900
	40	4.00	2.48		90	7.50	2.708	1.292	1.052
20	10	0.50	0.96	18	30	1.67	1.604	1.036	0.776
	20	1.00	1.48		60	3.33	2.000	1.208	0.884
	40	2.00	2.10		90	5.00	2.712	1.308	0.976
40	10	0.25	0.80	24	30	1.25	1.548	1.216	0.956
	20	0.50	1.82		60	2.50	2.104	1.204	0.944
	40	1.00	2.02		90	3.75	2.760	1.316	1.024

[a] Note: These results are for the STEPWISE procedure at $\alpha = 0.15$.

The model parameters of the Flack and Chang study are $\rho_{yx1} = \rho_{yx2} = .5$, $\rho_{yxj} = 0$ (j = 3, ..., P), $\rho_{x_i x_{i+j}} = 0.3i$, and $\rho_{yx_i}^2 = .3725$.

The model parameters of the present study are $\rho_{yx1} = \rho_{yx2} = \ldots = \rho_{yx6} = .147442$, $\rho_{yxj} = 0$ (j = 6, ..., P), $\rho_{yx_i}^2 = .130435$ when $\rho_{x_i x_j} = 0.0$, $\rho_{yx_i}^2 = .043478$ when $\rho_{x_i x_j} = 0.4$, and $\rho_{yx_i}^2 = .026087$ when $\rho_{x_i x_j} = 0.8$.

Similar to the results of the present study, the median value of P_N was shown by Flack and Chang (1987) to be positively effected by P and negatively effected by N. The median value of P_N reached a minimum value of 33% when N = 20 and 40 and P =10 and a maximum value of 89% when N = 10 and 20 and P = 40. The results of the current study show the minimum mean value of P_N to be 22.1% ($\rho_{x_i x_j}$ = 0.0, P = 12, and N = 90) and the maximum value to be 79.1% ($\rho_{x_i x_j}$ = 0.8, P = 24 and N = 30) [STEPWISE algorithm and α = 0.15]. The slightly lower values found in the present study are likely due to the inclusion of the noncollinear case and more favourable N to P ratios. When excluding the noncollinear case, the minimum value of P_N is then equal to 37.4% ($\rho_{x_i x_j}$ = 0.4, P =12, and N = 90), a value comparable to Flack and Chang's (1987) value. The results of the present study show that, within each of the subset selection procedures and for every value of α, the mean number of authentic variables reaching the final subset (C_A) was low compared to the actual number of authentic variables available. Moreover, the average number of noise variables reaching the final subset (C_N) and hence, the mean proportion of the final subset that was noise (P_N) could be quite high. However, under optimal conditions (i.e. when α = 0.15, $\rho_{x_i x_j}$ = 0.0, P = 12 , and N = 90) nearly half of the authentic variables reached the final subset on average, the mean value of C_N was less than one and the mean percentage of noise variables ranged from 22% to 23%.

In the comparison of estimates of $\rho_{yx_j}^2$ of this study, the factors affecting R^2 and R_k^2 were shown to be generally consistent over 'best' subset selection algorithms and level of α. The most important factor was sample size, followed by number of candidate variables and then degree of collinearity. Specifically, both R^2 and R_k^2 were shown to be negatively influenced by increasing sample size and increasing collinearity while increasing the number of candidate variables had a positive influence. Furthermore, there was some evidence for a P x N interaction, suggesting that the rate of inflation due to the number of candidate variables may be negatively influenced by sample size.

The factors affecting R_p^2 were not as consistent over the three subset selection procedures. In the STEPWISE and FORWARD procedures, the degree of collinearity and the number of candidate variables mainly influenced R_p^2; the amount of influence depended upon on the level of α. In particular, the proportion of explained variation accounted for by collinearity increased with α

while the proportion of explained variation accounted for by the number of candidate variables decreased with increasing α. Both of these factors were negatively related to R_p^2. However, in the BACKWARD procedure, sample size was the major factor affecting R_p^2 followed by degree of collinearity and number of candidate variables. Specifically, increasing sample size was negatively related to R_p^2. Also, there was evidence for two-way interaction effects among degree of collinearity, sample size and number of candidate variables. In particular there was strong evidence of a P X N interaction. Generally, when N = 30, increasing the number of candidate variables had a positive effect on R_p^2. However, when N = 60 and N = 90, increasing the number of candidate variables had a negative effect on R_p^2.[1]

Cohen and Cohen (1983, p. 107) proposed R_p^2 as an alternative measure of ρ_{yxi}^2 when subset selection methods are used. However, the results of this study showed that R_p^2 over-compensated for inflation yielding a fairly conservative estimate. The only exception to this finding occurred when ρ_{yxi}^2 was small, α was large and N and P were minimized. Here R_p^2 approached the population value. Consequently, while neither R_k^2 nor R_p^2 can universally be recommended as unbiased estimates of ρ_{yxi}^2 when a 'best' subset selection algorithm is used, under favourable conditions, R_k^2 may not be as inflated an estimate as once thought and R_p^2 can provide an estimate that will not exceed the population value and on occasion not be very conservative.

In Flack and Chang's (1987) examination of R^2 inflation, they found the adjusted R^2 statistic to be highly inflated; thus concluding that methods that provide less biased estimates of ρ_{yxi}^2 are required. However, part of the stated purpose of Flack and Chang's (1987) study was to show how poorly variable selection procedures perform under harsh parametric conditions. To that end, the sample size (N) to number of candidate variables (P) ratio of the Flack and Chang (1987) study was allowed to range from 0.25 (10/40) to 4.0 (40/10) indicating that the number of candidate predictor variables often exceeded the sample size. Within the present study, the N to P ratios ranged from 1.25 (30/24) to 7.5 (90/12), indicating that the sample size always exceeded the number of candidate predictor variables. Hence, the results of the present study showed that when the squared population coefficient of multiple determination reached at least a medium conventional value and the collinearity was low,

---

[1]It is not clear why this should be so.

using an extremely conservative level of significance along with a large sample size produced favourable mean values of R^2 and R_k^2. It is important to note, however, that increasing N had a limited capacity to reduce inflation, as a point of diminishing returns was reached. However, R^2 and R_k^2 were generally inflated measures of ρ_{yxi}^2 and R_p^2 was generally a conservative measure. For all three algorithms, the minimum inflation in R^2 and R_k^2 occurred when α was set at the smallest investigated value ($\alpha = 0.15/p$) while R_p^2 attained its maximum value when α was set at the largest investigated value ($\alpha = 0.15$). The mean values of R^2, R_k^2 and R_p^2 indicated that for the 'best' subset algorithms the BACKWARD procedure revealed generally more inflated values of R^2 and R_k^2 and generally less conservative values of R_p^2. However, when sample size was large, the differences were minimal.

To conclude, this author must join Flack and Chang (1987) in cautioning the user about ascribing importance to variables based upon their appearance in 'best' subset models and on an over reliance in using the popular sample estimates of the coefficients of multiple determination as unbiased estimates of effect size. Indeed, even under favourable conditions, noise variables enter the final subset model, especially when collinearity is present, and R^2 and R_k^2 may substantially overestimate ρ_{yxi}^2. Certainly when N is small compared to P the number of noise variables selected will very likely outnumber the number of authentic variables and R^2 and R_k^2 will be inflated. Furthermore, the R_p^2 estimate can also be extremely inaccurate. In any case, the wise investigator would not use models nor estimates of ρ_{yxi}^2 obtained strictly by 'best' subset selection algorithms as a basis for deriving conclusions unless they are confirmed by theoretical considerations and subsequent validation.

References

Beal, E. M. L. (1970). Note on procedures for variable selection in multiple regression. Technometrics, 12, 909-914.

Bendel, R. B., & Afifi, A. A. (1977). Comparison of stopping rules in forward 'stepwise' regression. Journal of the American Statistical Association, 72, 46-53.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York: John Wiley.

Berk, K. N. (1978). Comparing subset regression procedures. Technometrics, 20, 1-6.

Breiman, L., & Freedman, D. (1983). How many variables should be entered in a regression equation. Journal of the American Statistical Association, 78, 131-136.

Chatterjee, S., & Price, B. (1977). Regression analysis by example. New York: John Wiley.

Cohen, J. (1969) Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cohen, J., & Cohen, P. (1975). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, New Jersey: Erblaum.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences (2nd ed). Hillsdale, New Jersey: Erlbaum.

Cronbach, L. J. (1987). Statistical tests for moderator variables: Flaws in analyses recently proposed. Psychological Bulletin, 102, 414-417.

Diehr, G., & Hoflin, D. R. (1974). Approximating the distribution of the sample R^2 in best subset regressions. Technometrics, 16, 317-320.

Dixon, W. J. (1981). BMDP statistical software 1981. Berkeley: University of California Press.

Efroymson, M. A. (1960). Mathematical methods for digital computers. New York: John Wiley.

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: The problem revisited. Review of Economics and Statistics, 49, 92-107.

Flack, V. F., & Chang, P. C. (1987). Frequency of selecting noise variables in subset regression. The American Statistician, 41, 84-86.

Fox, J. (1984). Linear statistical models and related methods. New York: John Wiley.

Galarneau Gibbons, D. I. (1981). A simulation study of some ridge estimators. Journal of the American Statistical Association, 76, 131-139.

Gordon, R. A. (1968). Issues in multiple regression. The American Journal of Sociology, 73, 592-616.

Gunst, R. F. & Mason, R. L. (1977). Advantages of examining multicollinearities in regression analysis. Biometrics, 33, 249-260.

Hocking, R. R. (1983) Developments in linear regression methodology: 1959-1982. Technometrics, 25, 219-230.

Hocking, R. R., & Leslie, R. N. (1967). Selection of the best subset in regression analysis. Technometrics, 9, 531-540.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. Biometrics, 32, 1-49.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. Technometrics, 12, 55-67.

Hoerl, R. W., Schuenemeyer, J. H., & Hoerl, A. E. (1986). A simulation of biased estimation and subset regression techniques. Technometrics, 28, 369-380.

Kendall, M. G. (1957). A course in multivariate analysis. London: Griffin.

Keselman, H. J. (1988). Collinearity among predictor variables: Consequences, detection and circumvention . Unpublished manuscript. University of Manitoba, Department of Psychology, Winnipeg.

Kumar, T. K. (1975). Multicollinearity in regression analysis. Review of Economics and Statistics, 57, 365-366.

Lovell, M. C. (1983). Data Mining. Review of Economics and Statistics, 65, 1-12.

Mantel, N. (1970). Why stepdown procedures in variable selection. Technometrics , 12, 621-625.

Marascuilo. L. A., & Serlin, R. C. (1988). Statistical methods for the social and behavioral sciences . New York: W. H. Freeman and Company.

Marsaglia, G., MacLaren, M. D., & Bray, T. A. (1964). A fast procedure for generating normal random variables. Communication of the ACM, 7, 4-10.

McDonald, G. C., & Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. Journal of the American Statistical Association , 70, 407-416.

Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and prediction . New York: Holt, Rinehart & Winston.

Rencher, A. C. & Pun, F. C. (1980). Inflation of R^2 in best subset regression. Technometrics , 22, 49-53.

Rockwell, R. C. (1975). Assessment of multicollinearity. Sociological Methods & Research , 3, 308-320.

Ryan, T. A., Jr., Joiner, B. L., and Ryan, F. (1981). Minitab reference manual , University Park, Pennsylvania: The Pennsylvania State University.

SAS Institute, (1985) SAS user's guide: Basics (5th ed.). Cary, N. C.: Author.

SAS Institute, (1985) SAS user's guide: Statistics (5th ed.). Cary, N. C.: Author.

Sax, G. (1989). Principles of educational and psychological measurement and evaluation . Belmont: Wadsworth.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association, 62, 626-633.

Silvey, S. D. (1969). Multicollinearity and imprecise estimation. Journal of the Royal Statistical Society, Series B, 31, 539-552.

SPSS, Inc. (1983). SPSS x user's guide. New York: Mcgraw-Hill.

Tatsuoka, M. M. (1971). Multivariate analysis. New York: John Wiley.

Thorndike, R. L., & Hagen, E. P. (1977). Measurement and evaluation in psychology and education (4th ed.). New York: John Wiley.

Wichern, D. W., & Churchill, G. A. (1978). A comparison of ridge estimators. Technometrics, 20, 301-311.

Wilkinson, L. (1979). Tests of significance in stepwise regression. Psychological Bulletin, 86, 168-174.

Younger, M. S. (1985). A first course in linear regression. Boston: Duxbury.

Appendix A

FORTRANand SAS computer programs

The following FORTRANprogram generates the 250 replications of the 27 combinations of the three data conditions ($\rho_{x_i x_j}$, P and N).

```
      INTEGERISEED1, ISEED2
      INTEGERK
      INTEGERKP1
      INTEGERDEBUG
      INTEGERNSIM
      REAL*8 RXX
      REAL*8 RXY
      INTEGERN
      INTEGERP
      REAL*8 B
      REAL*8 RXXSR1, RXXSR2
      INTEGERI, J, L
      REAL*8 Z1, Z2
      REAL*8 X(100,24)
      REAL*8 Y(100)
      REAL*8 XM(24)
      REAL*8 X2(24)
      REAL*8 YM
      REAL*8 Y2
      INTEGERCOND
C
C   VARIABLE DICTIONARY
C
C   ISEED1, ISEED2 -- SEEDS TO THE PSEUDO-RANDOM NUMBER
GENERATOR
C K          -- NUMBER OF AUTHENTIC PREDICTOR VARIABLES
C KP1        -- EQUAL TO K+1
C DEBUG      -- LOGICAL FLAG TO PRINT OUT DEBUGGING INFORMATION
C NSIM       -- NUMBER OF SIMULATION TRIALS
```

```
C  RXX       -- CORRELATION BETWEEN AUTHENTIC PREDICTOR VARS.
C  RXY       -- CORRELATION BETWEEN AUTHENTIC PREDICTORS AND
C             THE DEPENDENT VARIABLE.
C  N         -- SAMPLE SIZE
C  P         -- NUMBER OF CANDIDATE PREDICTOR VARIABLES.
C  B         -- REGRESSION COEFFICIENTS OF AUTHENTIC PREDICTOR
C             VARIABLES IN THE REGRESSION EQUATION.
C  RXXSR1    -- EQUALS THE SQUARE ROOT OF RXX
C  RXXSR2    -- EQUALS THE SQUARE ROOT OF (1-RXX)
C  I, J, L   -- LOOP COUNTERS
C  Z1, Z2    -- PSEUDO-RANDOM STANDARD NORMAL DEVIATES
C  X(90,24)  -- THE X MATRIX
C  Y(90)     -- THE DEPENDENT VARIABLE VECTOR
C  XM(24)    -- THE MEAN OF EACH OF THE P X VARIABLES
C  X2(24)    -- THE SUM OF SQUARES FOR THE P X VARIABLES
C  YM        -- THE MEAN OF THE DEPENDENT VARIABLE
C  Y2        -- THE  SUM OF SQUARES OF THE DEPENDENT VARIABLE
C  COND      -- COUNTS THE NUMBER OF SETS OF CONDITIONS
C
      READ(15,1) ISEED1, ISEED2, NSIM, DEBUG, K
1     FORMAT(2I11, 3I5)
      WRITE(6,2) ISEED1, ISEED2, NSIM, K
2     FORMAT('1',///,'SEED 1:',I11,/,' SEED 2:',I11,/,
     *' NUMBER OF SIMULATIONS: ',I5,/,'K=',I4)
      CALL RSTART(ISEED1,ISEED2)
      COND=0
      REWIND14
      REWIND16
      REWIND17
      REWIND18
C
C     READ IN SIMULATION DATA CONDITIONS
C
100   READ(5,3) RXX
3     FORMAT(F5.3)
      IF (RXX.EQ.1D0) GO TO 999
```

```
      COND=COND+1
      RXXSR1=DSQRT(RXX)
      RXXSR2=DSQRT(1-RXX)
      KP1=K+1
      READ(5,4) N, B, P, RXY
 4    FORMAT(I5,F15.10,I5,F10.6)
      WRITE(6,5) RXX, RXY, B, N, P
 5    FORMAT(///,'SIMULATION CONDITIONS: ',//,
     * 5X,'CORRELATION',T20,'CORRELATION',T40,'REGRESSION',T55,
     * 'SAMPLE',T65,'NUMBER OF',/,T5,'XVARS',T20,'X AND Y',T40,
     * 'COEFFICIENT',T55,'SIZE',T65,'PREDICTORS',//,T5,F5.3,T20,F10.6,
     * T40,F10.6,T55,I5,T65,I5)
      DO 20 I=1,NSIM
C
C     INITIALIZE STATISTICAL VARIABLES TO ZERO
C
          YM=0D0
          Y2=0D0
          DO 25 L=1,P
             XM(L)=0D0
             X2(L)=0D0
25        CONTINUE
C
C     GENERATE THE N * P X MATRIX
C
          DO 30 J=1,N
             Z1=RNOR(0)
             DO 40 L=1,K
                Z2=RNOR(0)
                X(J,L)=RXXSR2*Z2 + RXXSR1*Z1
                XM(L)=XM(L)+X(J,L)
                X2(L)=X2(L)+(X(J,L)*X(J,L))
                IF (DEBUG.EQ.1)WRITE(6,41) Z1,Z2,J,L,X(J,L),XM(L),X2(L)
41              FORMAT('0','Z1',8X,'Z2',8X,'J',4X,'L',4X,'X',14X,
     *          'XM',13X,'X2',/2F10.7,2I5,4F15.10)
40           CONTINUE
```

```
            DO 50 L=KP1,P
               Z2=RNOR(0)
               X(J,L)=Z2
               XM(L)=XM(L) + X(J,L)
               X2(L)=X2(L) + X(J,L)*X(J,L)
               IF(DEBUG.EQ.1)WRITE(6,51)Z2,J,L,X(J,L),XM(L),X2(L)
51             FORMAT('0','Z2',8X,'J',4X,'L',4X,'X',14X,
     *         'XM',13X,'X2',/,F10.7,2I5,3F15.10)
50          CONTINUE
30       CONTINUE
C
C     GENERATE THE VECTOR OF Y OBSERVATIONS
C
         DO 90 J=1,N
            Y(J)=0D0
            DO 95 L=1,K
               Y(J)=Y(J)+B*X(J,L)
               IF (DEBUG.EQ.1) WRITE(6,94) J, L, Y(J), X(J,L)
94             FORMAT('0','J= ',I5,'L= ',I5,'Y= ',F16.10,' X=',F16.10)
95          CONTINUE
            Z2=RNOR(0)
            Y(J)=Y(J)+Z2
            YM=YM+Y(J)
            Y2=Y2+(Y(J)*Y(J))
            IF (DEBUG.EQ.1) WRITE(6,42) Z2, Y(J), YM, Y2
42          FORMAT('0','Z2',18X,'Y',19X,'YM',18X,'Y2',/,4F20.8)
90       CONTINUE
C
C  CALC. THE SAMPLE MEAN AND SUM OF SQUARES OF EACH OF THE
C  X VARIABLES AND STANDARDIZE EACH X OBSERVATION.
C
         DO 60 L=1,P
            XM(L)=XM(L)/N
            X2(L)=(X2(L)-DFLOAT(N)*XM(L)*XM(L))
            IF (DEBUG.GE.1)WRITE(6,61)L, XM(L), X2(L)
61          FORMAT('0','L=',I5,'MEAN=',F15.10,' SS=',F15.10)
```

```
            DO 70 J=1,N
               X(J,L)=(X(J,L)-XM(L))/(DSQRT(X2(L)))
70          CONTINUE
60        CONTINUE
C
C  CALC. THE SAMPLE MEAN AND SUM OF SQUARES OF Y
C
          YM=YM/N
          Y2=(Y2 - DFLOAT(N)*YM*YM)
          IF (DEBUG.GE.1)WRITE(6,52) YM, Y2
52        FORMAT('0','Y MEAN:',F15.10,'Y SS:',F15.10)
C
C  STANDARDIZE Y TO HAVE THE CORRELATION TRANSFORMATION
C
          DO 80 J=1,N
             Y(J)= (Y(J)-YM)/(DSQRT(Y2))
             IF (COND.EQ.1)
     *       WRITE(14,71) COND,I,RXX,RXY,P,N,Y(J),(X(J,L),L=1,P)
             IF (COND.EQ.2)
     *       WRITE(16,71) COND,I,RXX,RXY,P,N,Y(J),(X(J,L),L=1,P)
             IF (COND.EQ.3)
     *       WRITE(17,71) COND,I,RXX,RXY,P,N,Y(J),(X(J,L),L=1,P)
71           FORMAT(2I5,2F10.6,2I5,F20.16,/,5(5F20.16,/))
80        CONTINUE
20      CONTINUE
        IF (DEBUG.GE.1) CALL  MATMLT(X, Y, P, N, DEBUG)
        GO TO 100
999     CONTINUE
        CALL RSTOP(ISEED1, ISEED2)
        WRITE(6,101)ISEED1, ISEED2
        REWIND15
        WRITE(15,1) ISEED1,ISEED2,NSIM,DEBUG,K
        101   FORMAT('1','FINAL SEEDS:',/,' SEED 1:',I11,/,' SEED 2:',I11)
        STOP
        END
C
```

```
C  SUBROUTINE MATMLT CALCULATES  THE TWO MATRICES:
C  A=X'X AND G=X'Y
C
      SUBROUTINE MATMLT(X, Y, P, N, DEBUG)
      REAL*8 X(1000,24)
      REAL*8 Y(1000)
      INTEGER P
      INTEGER N
      INTEGER DEBUG
      REAL*8 A(24,24)
      REAL*8 G(24)
      INTEGER J, L, M
      DO 10 J=1,P
         DO 20 L=1,J
            A(J,L)=0D0
            DO 30 M=1,N
               A(J,L)=A(J,L)+X(M,L)*X(M,J)
30          CONTINUE
            IF (J.NE.L) A(L,J)=A(J,L)
20       CONTINUE
         G(J)=0D0
         DO 40 M=1,N
            G(J)=G(J)+Y(M)*X(M,J)
40       CONTINUE
10    CONTINUE
      WRITE(6,50)
50    FORMAT('1',T40,'XTX',T100,'XTY')
      DO 60 J=1,P
         WRITE(6,70) (A(J,L),L=1,P), G(J)
70    FORMAT('0',6F16.10,' : ',F16.10)
60    CONTINUE
      WRITE(6,80)
      IF (DEBUG.NE.1)GO TO 100
80    FORMAT('',T40,'X MAT',T100,'Y')
      DO 90 J=1,N
         WRITE(6,70) (X(J,L),L=1,P), Y(J)
```

```
90    CONTINUE
100   CONTINUE
      RETURN
      END
```

The following program is a SAS program which applies a 'best' subset selection algorithm (STEPWISE, BACKWARD or FORWARD) to the simulation data using the specified inclusion/deletion level (ALPHA = .15/p, .05, or .15).

```
//RESULTS DD DSN=DERKSN.RESULTS,DISP=MOD
//SYSIN DD *
DATA ONE;  /* READ IN SIMULATION DATA */
  INFILE TD;
  ARRAY X{12} X1-X12;
  INPUT COND I RXX RXY P N Y #2 X1-X5 #3 X6-X10 #4 X11-X12;
DATA TWO;
  J=1;
  SET ONE NOBS=MAX POINT=J;
  ALPHA=.05;        /* INCLUSION/DELETION LEVEL */
  FORCE=0;          /* NUMBER OF PREDICTORS FORCED INTO MODEL*/
  METHOD=1;         /* 1=STEPWISE, 2=FORWARD, AND 3=BACKWARD */
  REPS=MAX/N;
  KEEP RXX RXY P N ALPHA FORCE METHOD REPS;
  OUTPUT;
  STOP;
PROC PRINTTO UNIT=17;          /* REROUTE OUTPUT TO FILE */
PROC STEPWISE DATA=ONE;  /* PERFORM SUBSET SELECTION */
  BY I;
  MODEL Y=X1-X12/ SLE=.05 SLS=.05 STEPWISE;
PROC PRINTTO;
```

```
 *  REREAD SUBSET SELECTION OUTPUT, COLLECTING DEPENDENT
VARIABLES;

DATA COLLECT;
  INFILE FT17F001 MISSOVER;
```

```
INPUT LABEL $ 2-6 LABEL3 $ 65-67 @;
LABEL3=SUBSTR(LEFT(LABEL3),1,2);
LABEL=SUBSTR(LABEL,1,4);
LABEL2=SUBSTR(LABEL,1,1);
ARRAY X{12} X1-X12;
IF (_N_=1) THEN DO;
  DO I=1 TO 12;
    X{I}=0;
  END;
  K=0;
  R2=.;
  PN=.;
END;
RETAIN R2 K X1-X12 SIM;
IF LABEL3='I=' THEN DO;
  INPUT @68 SIM;
END;
IF LABEL='STEP' THEN DO;
  DO I=1 TO 12;
    X{I}=0;
  END;
  K=0;
  R2=.;
  PN=.;
  INPUT R2 52-61;
END;
IF LABEL= 'REGR' THEN DO;
  INPUT K 22-23;
END;
IF LABEL2='X' THEN DO;
  INPUT I 3-4;
  X{I}=1;
END;
IF LABEL='NO O' AND SIM NE . THEN DO;
  KEEP R2 K X1-X12 SIM PN NAUTH NNOISE CL;
  NAUTH=X1+X2+X3+X4+X5+X6;
```

```
    NNOISE=K-NAUTH;
    PN=NNOISE/K;
    CL=1;
    OUTPUT;
    K=0;
    R2=.;
    PN=.;
    DO I=1 TO 12;
      X{I}=0;
    END;
  END;
  IF LABEL='NO V' AND SIM NE . THEN DO;
    CL=0;
    NAUTH=X1+X2+X3+X4+X5+X6;
    NNOISE=0;
    OUTPUT;
    K=0;
    R2=.;
    PN=.;
    DO I=1 TO 12;
      X{I}=0;
    END;
  END;
DATA COMBINE;   /* CALC. RT2 AND RA2 FROM R2 */
  IF _N_=1 THEN SET TWO;
  SET COLLECT;
  RT2=1-(1-R2)*((N-1)/(N-P-1));
  IF (RT2 LT 0) AND (RT2 NE .) THEN RT2=0;
  RA2=1-(1-R2)*((N-1)/(N-K-1));
  IF (RA2 LT 0) AND (RA2 NE .) THEN RA2=0;

* CALCULATE MEAN AND STD. DEV. OF DEPENDENT VARS ;

PROC SUMMARY DATA=COMBINE ;
  CLASS CL;
  VAR R2 K X1-X12 PN RT2 RA2;
```

```
  ID RXX RXY P N ALPHA FORCE REPS METHOD;
  OUTPUT OUT=FOUR
    MEAN(R2 K PN RT2 RA2 )=MR2 MK MPN MRT2 MRA2
    STD(R2 K RT2 RA2 PN)=SR2 SK SRT2 SRA2 SPN
    SUM(X1-X12)=C1-C12  N(PN)=NR2;
PROC PRINT;
  TITLE 'RESULTS FROM SUMMARY';
DATA RESULTS.RESA12;
  TITLE 'SUMMARY OF RESULTS';
  SET RESULTS.RESA12 FOUR;
PROC PRINT;

* FREQUENCEY TABLE OF NO. OF AUTHENTIC VARIABLES BY NO. OF
NOISE VARIABLES;

PROC FREQ DATA=COMBINE;
  TABLES NAUTH*NNOISE/ OUT=FREQ1;
DATA FREQ1;
  IF _N_=1 THEN SET TWO;
  SET FREQ1;
DATA RESULTS.FREQA12;
  SET RESULTS.FREQA12 FREQ1;
```

Appendix B

Appendix B

The Standard Error of the Mean Value of C_N by Method and I/D Level [a]

			Method								
			STEPWISE			BACKWARD			FORWARD		
			I/D Level (α)								
$\rho_{x_i x_j}$	P	N	α_p	.05	.15	α_p	.05	.15	α_p	.05	.15
0.0	12	30	.018	.040	.062	.025	.053	.073	.018	.040	.062
		60	.020	.034	.056	.021	.039	.058	.020	.034	.057
		90	.016	.036	.054	.019	.037	.059	.016	.036	.054
	18	30	.022	.052	.091	.035	.099	.135	.022	.052	.095
		60	.021	.054	.085	.031	.064	.091	.021	.056	.087
		90	.018	.049	.082	.024	.056	.091	.018	.049	.085
	24	30	.021	.075	.133	.079	.221	.222	.021	.075	.138
		60	.024	.063	.114	.031	.093	.132	.024	.064	.114
		90	.024	.061	.102	.027	.070	.118	.024	.062	.105
0.4	12	30	.020	.033	.059	.021	.053	.075	.020	.033	.060
		60	.016	.032	.052	.020	.034	.056	.016	.032	.052
		90	.020	.037	.061	.026	.039	.064	.020	.037	.061
	18	30	.022	.055	.098	.038	.096	.125	.022	.056	.100
		60	.023	.053	.089	.026	.067	.103	.023	.053	.089
		90	.021	.051	.084	.024	.061	.090	.021	.052	.084
	24	30	.019	.069	.132	.055	.206	.202	.019	.069	.138
		60	.023	.054	.107	.027	.078	.124	.023	.053	.106
		90	.023	.060	.102	.024	.074	.115	.023	.060	.103
0.8	12	30	.017	.036	.062	.024	.050	.074	.017	.036	.064
		60	.021	.036	.060	.021	.040	.064	.021	.036	.060
		90	.017	.031	.052	.018	.032	.057	.017	.031	.052
	18	30	.018	.058	.092	.045	.106	.137	.018	.058	.093
		60	.025	.049	.085	.026	.059	.106	.025	.049	.089
		90	.020	.049	.085	.022	.051	.091	.020	.050	.085
	24	30	.020	.078	.138	.095	.235	.232	.020	.081	.138
		60	.020	.065	.109	.028	.092	.136	.020	.065	.111
		90	.024	.062	.100	.028	.075	.113	.024	.062	.101

[a]Note: See Table 6 note.

The Standard Error of the Mean Value of C_A by Method and I/D Level [a]

			Method								
			STEPWISE			BACKWARD			FORWARD		
			I/D Level (α)								
$\rho_{x_i x_j}$	P	N	α_p	.05	.15	α_p	.05	.15	α_p	.05	.15
0.0	12	30	.029	.050	.074	.041	.064	.085	.029	.050	.074
		60	.042	.057	.074	.046	.062	.076	.042	.057	.075
		90	.049	.068	.078	.054	.070	.079	.049	.068	.078
	18	30	.028	.051	.069	.038	.073	.080	.028	.052	.071
		60	.040	.060	.071	.047	.065	.072	.040	.060	.072
		90	.046	.070	.078	.053	.073	.079	.046	.070	.078
	24	30	.022	.046	.074	.031	.087	.092	.022	.046	.076
		60	.032	.062	.071	.037	.065	.076	.032	.062	.071
		90	.040	.065	.077	.045	.070	.079	.040	.065	.077
0.4	12	30	.025	.037	.054	.028	.050	.066	.025	.037	.055
		60	.030	.036	.050	.032	.042	.059	.030	.036	.050
		90	.032	.035	.049	.033	.041	.058	.032	.035	.049
	18	30	.020	.039	.061	.026	.066	.083	.020	.040	.062
		60	.028	.038	.054	.028	.045	.063	.028	.039	.056
		90	.032	.039	.051	.032	.044	.058	.032	.039	.053
	24	30	.022	.038	.067	.038	.086	.088	.022	.039	.070
		60	.026	.039	.057	.027	.047	.070	.026	.040	.059
		90	.030	.036	.052	.031	.043	.059	.030	.036	.053
0.8	12	30	.020	.033	.054	.035	.050	.068	.020	.033	.056
		60	.027	.037	.053	.034	.047	.059	.027	.037	.053
		90	.029	.040	.053	.036	.048	.058	.029	.040	.054
	18	30	.020	.035	.055	.035	.067	.081	.020	.035	.056
		60	.024	.035	.055	.033	.051	.070	.024	.035	.055
		90	.027	.038	.050	.032	.046	.061	.027	.038	.051
	24	30	.020	.038	.065	.045	.088	.096	.020	.038	.068
		60	.023	.039	.059	.023	.052	.068	.023	.039	.061
		90	.028	.040	.055	.033	.048	.058	.028	.040	.056

[a]Note: See Table 6 note.

Standard Errors of the Mean[a]

METHOD = STEPWISE
Inclusion/Deletion Level = α_p

$\rho_{x_i x_j}$	P	N	Freq	P_N	R^2	R_k^2	R_p^2
0.0	12	30	68	5.078	0.010	0.009	0.0044
0.0	12	60	111	3.206	0.007	0.006	0.0045
0.0	12	90	145	2.040	0.005	0.004	0.0036
0.0	18	30	65	5.705	0.012	0.012	0.0002
0.0	18	60	88	4.250	0.009	0.008	0.0043
0.0	18	90	108	3.247	0.006	0.006	0.0029
0.0	24	30	55	6.669	0.012	0.012	0.0000
0.0	24	60	89	4.659	0.007	0.006	0.0000
0.0	24	90	116	3.573	0.006	0.005	0.0016
0.4	12	30	64	5.741	0.012	0.012	0.0091
0.4	12	60	89	3.827	0.005	0.005	0.0038
0.4	12	90	121	2.785	0.003	0.003	0.0020
0.4	18	30	47	7.036	0.013	0.013	0.0000
0.4	18	60	87	4.721	0.006	0.006	0.0006
0.4	18	90	108	3.150	0.005	0.005	0.0019
0.4	24	30	53	6.481	0.011	0.011	0.0000
0.4	24	60	74	5.352	0.007	0.007	0.0001
0.4	24	90	102	3.939	0.005	0.004	0.0008
0.8	12	30	38	7.468	0.014	0.014	0.0075
0.8	12	60	66	5.405	0.008	0.008	0.0054
0.8	12	90	92	3.919	0.004	0.004	0.0020
0.8	18	30	49	7.094	0.010	0.010	0.0000
0.8	18	60	65	5.822	0.007	0.006	0.0005
0.8	18	90	79	4.976	0.004	0.004	0.0011
0.8	24	30	45	7.213	0.014	0.013	0.0000
0.8	24	60	54	6.664	0.007	0.007	0.0000
0.8	24	90	88	4.716	0.005	0.005	0.0000

[a]Note: See Table 6 note.

Standard Errors of the Mean[a]

METHOD = STEPWISE
Inclusion/Deletion Level = .05

$\rho_{x_i x_j}$	P	N	Freq	P_N	R^2	R_k^2	R_p^2
0.0	12	30	157	3.128	0.011	0.010	0.0067
0.0	12	60	194	2.196	0.006	0.006	0.0047
0.0	12	90	217	1.583	0.005	0.004	0.0041
0.0	18	30	183	3.094	0.011	0.011	0.0032
0.0	18	60	206	2.491	0.008	0.007	0.0046
0.0	18	90	230	2.194	0.005	0.005	0.0031
0.0	24	30	204	2.665	0.011	0.011	0.0007
0.0	24	60	221	2.405	0.007	0.007	0.0020
0.0	24	90	238	2.127	0.005	0.005	0.0021
0.4	12	30	136	3.525	0.010	0.010	0.0056
0.4	12	60	163	3.054	0.005	0.005	0.0028
0.4	12	90	199	2.497	0.004	0.003	0.0022
0.4	18	30	171	3.184	0.011	0.010	0.0040
0.4	18	60	192	2.798	0.006	0.006	0.0020
0.4	18	90	214	2.570	0.004	0.004	0.0017
0.4	24	30	193	2.600	0.012	0.011	0.0000
0.4	24	60	206	2.667	0.006	0.006	0.0010
0.4	24	90	216	2.559	0.004	0.004	0.0012
0.8	12	30	121	4.095	0.010	0.010	0.0038
0.8	12	60	135	3.678	0.006	0.006	0.0036
0.8	12	90	155	3.021	0.004	0.004	0.0021
0.8	18	30	158	3.201	0.011	0.010	0.0001
0.8	18	60	166	3.366	0.006	0.005	0.0013
0.8	18	90	184	3.046	0.004	0.003	0.0007
0.8	24	30	179	2.809	0.012	0.012	0.0013
0.8	24	60	183	3.012	0.007	0.006	0.0004
0.8	24	90	203	2.817	0.004	0.004	0.0004

[a]Note: See Table 6 note.

Standard Errors of the Mean[a]

METHOD = STEPWISE
Inclusion/Deletion Level = .15

$\rho_{x_i x_j}$	P	N	Freq	P_N	R^2	R_k^2	R_p^2
0.0	12	30	229	2.203	0.010	0.009	0.0075
0.0	12	60	247	1.692	0.007	0.006	0.0054
0.0	12	90	248	1.391	0.005	0.004	0.0045
0.0	18	30	246	1.800	0.011	0.011	0.0064
0.0	18	60	249	1.648	0.008	0.007	0.0055
0.0	18	90	250	1.368	0.005	0.005	0.0044
0.0	24	30	249	1.466	0.013	0.012	0.0046
0.0	24	60	246	1.449	0.007	0.007	0.0042
0.0	24	90	249	1.276	0.005	0.005	0.0041
0.4	12	30	217	2.388	0.010	0.009	0.0053
0.4	12	60	236	2.140	0.005	0.005	0.0030
0.4	12	90	242	2.071	0.004	0.004	0.0027
0.4	18	30	239	1.969	0.012	0.011	0.0053
0.4	18	60	245	1.732	0.007	0.006	0.0035
0.4	18	90	245	1.821	0.005	0.004	0.0027
0.4	24	30	246	1.398	0.012	0.012	0.0043
0.4	24	60	247	1.613	0.007	0.006	0.0022
0.4	24	90	248	1.552	0.005	0.004	0.0020
0.8	12	30	205	2.680	0.010	0.009	0.0049
0.8	12	60	227	2.583	0.006	0.005	0.0031
0.8	12	90	225	2.395	0.004	0.003	0.0021
0.8	18	30	232	1.882	0.011	0.010	0.0037
0.8	18	60	239	2.096	0.006	0.005	0.0023
0.8	18	90	240	1.967	0.004	0.004	0.0014
0.8	24	30	245	1.418	0.013	0.012	0.0036
0.8	24	60	247	1.592	0.007	0.006	0.0020
0.8	24	90	249	1.463	0.005	0.004	0.0016

[a]Note: See Table 6 note.

Standard Errors of the Mean[a]

METHOD = FORWARD
Inclusion/Deletion Level = α_p

$\rho_{x_i x_j}$	P	N	Freq	P_N	R^2	R_k^2	R_p^2
0.0	12	30	68	5.078	0.010	0.009	0.0044
0.0	12	60	111	3.206	0.007	0.006	0.0045
0.0	12	90	145	2.040	0.005	0.004	0.0036
0.0	18	30	65	5.705	0.012	0.012	0.0002
0.0	18	60	88	4.250	0.009	0.008	0.0043
0.0	18	90	108	3.247	0.006	0.006	0.0029
0.0	24	30	55	6.669	0.012	0.012	0.0000
0.0	24	60	89	4.659	0.007	0.006	0.0000
0.0	24	90	116	3.573	0.006	0.005	0.0016
0.4	12	30	64	5.741	0.012	0.012	0.0091
0.4	12	60	89	3.827	0.005	0.005	0.0038
0.4	12	90	121	2.785	0.003	0.003	0.0020
0.4	18	30	47	7.036	0.013	0.013	0.0000
0.4	18	60	87	4.721	0.006	0.006	0.0006
0.4	18	90	108	3.150	0.005	0.005	0.0019
0.4	24	30	53	6.481	0.011	0.011	0.0000
0.4	24	60	74	5.352	0.007	0.007	0.0001
0.4	24	90	102	3.939	0.005	0.004	0.0008
0.8	12	30	38	7.468	0.014	0.014	0.0075
0.8	12	60	66	5.405	0.008	0.008	0.0054
0.8	12	90	92	3.919	0.004	0.004	0.0020
0.8	18	30	49	7.094	0.010	0.010	0.0000
0.8	18	60	65	5.822	0.007	0.006	0.0005
0.8	18	90	79	4.976	0.004	0.004	0.0011
0.8	24	30	45	7.213	0.014	0.013	0.0000
0.8	24	60	54	6.664	0.007	0.007	0.0000
0.8	24	90	88	4.716	0.005	0.005	0.0000

[a]Note: See Table 6 note.

Standard Errors of the Mean[a]

METHOD = FORWARD
Inclusion/Deletion Level = .05

$\rho_{x_i x_j}$	P	N	Freq	P_N	R^2	R_k^2	R_p^2
0.0	12	30	157	3.128	0.011	0.010	0.0067
0.0	12	60	194	2.196	0.006	0.006	0.0047
0.0	12	90	217	1.582	0.005	0.004	0.0041
0.0	18	30	183	3.085	0.011	0.011	0.0032
0.0	18	60	206	2.480	0.008	0.007	0.0046
0.0	18	90	230	2.191	0.005	0.005	0.0031
0.0	24	30	204	2.662	0.011	0.011	0.0007
0.0	24	60	221	2.396	0.008	0.007	0.0020
0.0	24	90	238	2.133	0.005	0.005	0.0021
0.4	12	30	136	3.525	0.010	0.010	0.0056
0.4	12	60	163	3.054	0.005	0.005	0.0028
0.4	12	90	199	2.497	0.004	0.003	0.0022
0.4	18	30	171	3.184	0.011	0.010	0.0040
0.4	18	60	192	2.800	0.006	0.006	0.0020
0.4	18	90	214	2.572	0.004	0.004	0.0018
0.4	24	30	193	2.586	0.012	0.011	0.0000
0.4	24	60	206	2.668	0.006	0.006	0.0010
0.4	24	90	216	2.559	0.004	0.004	0.0012
0.8	12	30	121	4.097	0.010	0.010	0.0038
0.8	12	60	135	3.678	0.006	0.006	0.0036
0.8	12	90	155	3.021	0.004	0.004	0.0021
0.8	18	30	158	3.201	0.011	0.010	0.0001
0.8	18	60	166	3.366	0.006	0.005	0.0013
0.8	18	90	184	3.048	0.004	0.003	0.0007
0.8	24	30	179	2.790	0.013	0.012	0.0013
0.8	24	60	183	3.012	0.007	0.006	0.0004
0.8	24	90	203	2.818	0.004	0.004	0.0004

[a]Note: See Table 6 note.

Standard Errors of the Mean[a]

METHOD = FORWARD

Inclusion/Deletion Level = .15

$\rho_{x_i x_j}$	P	N	Freq	P_N	R^2	R_k^2	R_p^2
0.0	12	30	229	2.193	0.010	0.009	0.0076
0.0	12	60	247	1.689	0.007	0.006	0.0055
0.0	12	90	248	1.386	0.005	0.004	0.0045
0.0	18	30	246	1.804	0.012	0.011	0.0065
0.0	18	60	249	1.635	0.008	0.007	0.0056
0.0	18	90	250	1.368	0.005	0.005	0.0044
0.0	24	30	249	1.469	0.013	0.012	0.0045
0.0	24	60	246	1.425	0.007	0.006	0.0041
0.0	24	90	249	1.272	0.005	0.005	0.0041
0.4	12	30	217	2.376	0.010	0.009	0.0054
0.4	12	60	236	2.137	0.005	0.005	0.0031
0.4	12	90	242	2.071	0.004	0.004	0.0027
0.4	18	30	239	1.964	0.012	0.011	0.0052
0.4	18	60	245	1.720	0.007	0.006	0.0036
0.4	18	90	245	1.811	0.005	0.004	0.0027
0.4	24	30	246	1.389	0.012	0.012	0.0044
0.4	24	60	247	1.604	0.007	0.006	0.0022
0.4	24	90	248	1.550	0.005	0.004	0.0020
0.8	12	30	205	2.673	0.010	0.009	0.0051
0.8	12	60	227	2.574	0.006	0.005	0.0031
0.8	12	90	225	2.394	0.004	0.003	0.0021
0.8	18	30	232	1.872	0.011	0.010	0.0037
0.8	18	60	239	2.079	0.006	0.005	0.0024
0.8	18	90	240	1.965	0.004	0.004	0.0014
0.8	24	30	245	1.405	0.013	0.012	0.0024
0.8	24	60	247	1.590	0.007	0.006	0.0020
0.8	24	90	249	1.463	0.005	0.004	0.0016

[a]Note: See Table 6 note.

Standard Errors of the Mean[a]

METHOD = BACKWARD
Inclusion/Deletion Level = α_p

$\rho_{x_i x_j}$	P	N	Freq	P_N	R^2	R_k^2	R_p^2
0.0	12	30	76	4.738	0.013	0.012	0.0100
0.0	12	60	111	3.201	0.007	0.006	0.0052
0.0	12	90	151	2.030	0.005	0.004	0.0038
0.0	18	30	74	5.361	0.014	0.013	0.0010
0.0	18	60	99	3.932	0.010	0.009	0.0063
0.0	18	90	117	2.855	0.006	0.006	0.0030
0.0	24	30	64	5.730	0.021	0.020	0.0152
0.0	24	60	92	4.593	0.008	0.007	0.0000
0.0	24	90	120	3.449	0.006	0.005	0.0016
0.4	12	30	64	5.717	0.012	0.012	0.0091
0.4	12	60	95	3.705	0.005	0.005	0.0037
0.4	12	90	122	2.852	0.004	0.004	0.0026
0.4	18	30	57	5.411	0.017	0.016	0.0000
0.4	18	60	91	4.458	0.006	0.006	0.0007
0.4	18	90	107	3.379	0.005	0.005	0.0019
0.4	24	30	57	5.828	0.021	0.019	0.0105
0.4	24	60	74	5.320	0.008	0.007	0.0001
0.4	24	90	103	3.901	0.004	0.004	0.0008
0.8	12	30	52	5.752	0.014	0.013	0.0098
0.8	12	60	75	4.886	0.007	0.007	0.0048
0.8	12	90	99	3.680	0.004	0.004	0.0022
0.8	18	30	54	6.159	0.020	0.018	0.0125
0.8	18	60	74	5.263	0.007	0.006	0.0019
0.8	18	90	88	4.604	0.004	0.004	0.0012
0.8	24	30	58	5.799	0.027	0.025	0.0185
0.8	24	60	56	6.301	0.008	0.008	0.0000
0.8	24	90	96	4.387	0.005	0.005	0.0000

[a]Note: See Table 6 note.

Standard Errors of the Mean[a]

METHOD = BACKWARD
Inclusion/Deletion Level = .05

$\rho_{x_i x_j}$	P	N	Freq	P_N	R^2	R_k^2	R_p^2
0.0	12	30	175	2.712	0.011	0.010	0.0083
0.0	12	60	198	2.078	0.007	0.006	0.0051
0.0	12	90	219	1.555	0.005	0.004	0.0041
0.0	18	30	203	2.435	0.013	0.012	0.0069
0.0	18	60	212	2.333	0.008	0.007	0.0049
0.0	18	90	237	2.090	0.005	0.005	0.0033
0.0	24	30	218	1.926	0.016	0.015	0.0100
0.0	24	60	232	2.263	0.008	0.007	0.0026
0.0	24	90	239	1.976	0.006	0.005	0.0025
0.4	12	30	151	3.209	0.011	0.010	0.0080
0.4	12	60	172	2.912	0.005	0.005	0.0031
0.4	12	90	203	2.431	0.004	0.003	0.0023
0.4	18	30	186	2.599	0.013	0.012	0.0066
0.4	18	60	199	2.566	0.007	0.006	0.0026
0.4	18	90	217	2.573	0.005	0.004	0.0019
0.4	24	30	207	2.013	0.017	0.015	0.0094
0.4	24	60	212	2.404	0.007	0.006	0.0011
0.4	24	90	223	2.432	0.005	0.004	0.0012
0.8	12	30	148	3.254	0.011	0.010	0.0067
0.8	12	60	150	3.136	0.006	0.005	0.0036
0.8	12	90	172	2.635	0.004	0.003	0.0021
0.8	18	30	180	2.551	0.014	0.012	0.0078
0.8	18	60	186	2.855	0.006	0.005	0.0017
0.8	18	90	198	2.885	0.004	0.003	0.0007
0.8	24	30	211	1.937	0.017	0.015	0.0127
0.8	24	60	207	2.563	0.007	0.006	0.0012
0.8	24	90	212	2.601	0.005	0.004	0.0007

[a]Note: See Table 6 note.

Standard Errors of the Mean[a]

METHOD = BACKWARD
Inclusion/Deletion Level = .15

$\rho_{x_i x_j}$	P	N	Freq	P_N	R^2	R_k^2	R_p^2
0.0	12	30	233	1.906	0.011	0.010	0.0088
0.0	12	60	248	1.541	0.006	0.006	0.0056
0.0	12	90	249	1.413	0.005	0.004	0.0046
0.0	18	30	248	1.330	0.012	0.011	0.0105
0.0	18	60	250	1.432	0.007	0.007	0.0059
0.0	18	90	250	1.324	0.005	0.005	0.0045
0.0	24	30	248	0.895	0.011	0.012	0.0140
0.0	24	60	248	1.204	0.007	0.007	0.0053
0.0	24	90	248	1.197	0.005	0.005	0.0043
0.4	12	30	232	2.018	0.010	0.009	0.0065
0.4	12	60	240	2.035	0.005	0.005	0.0032
0.4	12	90	242	1.979	0.004	0.004	0.0028
0.4	18	30	247	1.473	0.012	0.011	0.0080
0.4	18	60	247	1.565	0.007	0.006	0.0041
0.4	18	90	244	1.634	0.005	0.004	0.0029
0.4	24	30	249	0.667	0.010	0.011	0.0131
0.4	24	60	249	1.345	0.007	0.006	0.0032
0.4	24	90	249	1.438	0.005	0.004	0.0024
0.8	12	30	227	2.003	0.010	0.009	0.0069
0.8	12	60	236	2.155	0.005	0.005	0.0033
0.8	12	90	235	1.947	0.004	0.003	0.0024
0.8	18	30	246	1.452	0.012	0.011	0.0093
0.8	18	60	242	1.796	0.007	0.005	0.0034
0.8	18	90	244	1.708	0.004	0.004	0.0018
0.8	24	30	249	0.819	0.012	0.013	0.0158
0.8	24	60	249	1.324	0.007	0.006	0.0035
0.8	24	90	249	1.310	0.005	0.004	0.0019

[a]Note: See Table 6 note.