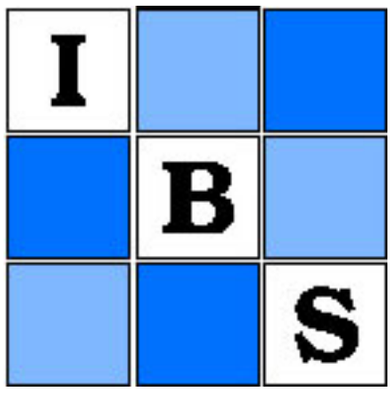

BIOMETRICS *32*, 1–49
March, 1976

A BIOMETRICS INVITED PAPER

THE ANALYSIS AND SELECTION OF VARIABLES IN LINEAR REGRESSION

R. R. HOCKING

Department of Computer Science and Statistics, Mississippi State, Mississippi, U.S.A. 39762

LIST OF CONTENTS

SUMMARY

The problems of subset selection and variable analysis in linear regression are reviewed. The discussion covers the underlying theory, computational techniques and selection criteria. Alternatives to least squares, including ridge and principal component regression, are considered. These biased estimation procedures are related and contrasted with least squares. Examples are included to illustrate the essential points.

1. INTRODUCTION

The primary purpose of this paper is to provide a review of the concepts and methods associated with variable selection in linear regression models. The title of the paper reflects the thought that variable selection is just a part of the more general problem of analyzing the structure of the data. Thus, the scope of the paper has been broadened to include other topics, particularly the problems of multicollinearity and biased estimation.

The problem of determining the "best" subset of variables has long been of interest to applied statisticians and, primarily because of the current availability of high-speed computations, this problem has received considerable attention in the recent statistical literature. Several papers have dealt with various aspects of the problem but it appears that the typical regression user has not benefited appreciably. One reason for the lack of resolution of the problem is the fact that it has not been well defined. Indeed, it is apparent that there is not a single problem, but rather several problems for which different answers might be appropriate. The intent of this review is not to give specific answers but merely to summarize the current state of the art. Hopefully, this will provide general guidelines for applied statisticians.

The problem of selecting a subset of independent or predictor variables is usually described in an idealized setting. That is, it is assumed that (a) the analyst has data on a large number of potential variables which include all relevant variables and appropriate functions of them plus, possibly, some other extraneous variables and variable functions and (b) the analyst has available "good" data on which to base the eventual conclusions. In practice, the lack of satisfaction of these assumptions may make a detailed subset selection analysis a meaningless exercise.

The problem of assuring that the "variable pool" contains all important variables and variable functions is not an easy one. The analysis of residuals (see e.g. Anscombe [1961], Draper and Smith [1966], and Daniel and Wood [1971]) may reveal different functional

forms which might be considered and may even suggest variables which were not initially included. These revelations, especially the latter, seldom occur without considerable skill and effort on the part of the analyst.

The assumption of "good data" includes the usual linear model assumptions such as homogeneity of variance, etc. Again residual plots may suggest transformations and also may reveal bad data points or "outliers." A serious problem which is included under this heading is that of multicollinearity among the independent variables. The consequences of near degeneracy of the matrix of independent variables have been described by a number of authors. For example, see the text by Johnston [1972] or the recent paper by Mason *et al.* [1975]. As observed by these authors, multicollinearity can arise because of sampling in a subspace of the true sample space either by chance or by necessity or simply by including extraneous predictors which are closely related to the actual predictors. Whatever the cause, this degeneracy may result in estimates of the regression coefficients with high variance and which, as a consequence, may be far from the true values. (See e.g. Hoerl and Kennard [1970a].) In addition, the resulting prediction equation may be quite unreliable, especially if it is used outside of the immediate neighborhood of the original data.

Marquardt [1974b] suggested that the two problems, multicollinearity and erratic data, should be tackled simultaneously. The instability of least squares in the presence of near degeneracies suggests that residual plots may not reveal bad data or may give erroneous indications. The need for procedures which are "robust" against such departures is apparent. Marquardt [1974b] suggested that ridge analysis (see Section 5.2) may be an appropriate tool. Beaton and Tukey [1974] discussed robustness in the context of polynomial regression. Holland [1973] suggested a combination of ridge and robust methods. Andrews [1974] proposed robust methods for multivariate regression and provided an illustration using an example from Daniel and Wood [1971]. It is of interest to note that both references reach, essentially, the same conclusions, Andrews [1974] using the robust regression procedures and Daniel and Wood [1971] using a combination of subset analysis and inspection of residual plots. This suggests that an analyst, skilled in one or more of the techniques to be described in this paper, may well be using a robust procedure. The role of the developers of regression methodology is to provide the less skilled user with techniques which are robust while easy to use and understand.

The problem of variable selection will be initially described under the assumption that the two requirements (a) and (b), described above, are met. No attempt will be made to discuss (a) nor will we discuss the use of residual plots to detect erratic data or departures from normality. It should be emphasized that a residual analysis for the candidate subset equations is recommended. A number of recent papers dealing with biased estimation in the presence of multicollinearity will be discussed in Section 5 and related to the subset analysis.

To provide a basis for the discussion, Section 2 contains a review of the consequences of incorrect model specification which provides a theoretical motivation for variable deletion.

The problem of determining an appropriate equation based on a subset of the original set of variables contains three basic ingredients, namely (1) the computational technique used to provide the information for the analysis, (2) the criterion used to analyze the variables and select a subset, if that is appropriate, and (3) the estimation of the coefficients in the final equation. Typically, a procedure might embody all three ideas without clearly identifying them. For example, one might use a standard computer package based on the stepwise regression concept as described by Efroymson [1960]. The basis for this procedure is just the Jordan reduction method for solving linear equations (see Hemmerle [1967])

with a specific criterion for determining the order in which variables are introduced or deleted. However, for a specified stopping rule, stepwise regression also implies the selection of a particular subset of variables. Further, the estimates of the coefficients for the final equation are obtained by applying least squares to the retained variables.

A number of computational techniques are reviewed in Section 3. Criteria for analyzing variables and selection of subsets are described in Section 4. In view of the recent results on biased estimation, it seems reasonable to consider alternatives to subset least squares. A discussion is given in Section 5. Finally, Section 6 contains a suggestion for incorporating the concept of biased estimation into the subset selection process.

In the process of preparing this paper, there was a temptation to conduct a simulation study to arrive at a specific recommendation. This temptation was easily suppressed after outlining what might be a reasonable set of ranges for the many parameters involved. In addition, there was considerable doubt that the results would be, in any general sense, conclusive. In lieu of this a number of examples are presented to illustrate the various points.

A comment is in order on the list of references. In addition to references cited in the text, this list contains some references which are primarily of historical interest and others which might be viewed as collateral. Also, there are numerous occasions where a particular development could have been credited to several authors but for brevity only one or two are cited or perhaps none if the result is simple or well known.

2. NOTATION AND BASIC CONCEPTS

2.1. Notation and Assumptions.

It is assumed that there are $n \geq t + 1$ observations on a t-vector of input variables, $x' = (x_1 \cdots x_t)$, and a scalar response, y, such that the j^{th} response, $j = 1 \cdots n$, is determined by

$$y_j = \beta_0 + \sum_{i=1}^{t} \beta_i x_{ij} + e_j \ . \tag{2.1}$$

The residuals, e_j , are assumed identically and independently distributed, usually normal, with mean zero and unknown variance, σ^2. (The inputs x_{ij} are frequently taken to be specified design variables, but in many cases it is more appropriate to consider them as random variables and assume a joint distribution on y and x, say, multivariate normal.) Note that implicit in these assumptions is the assumption that the variables $x_1 \cdots x_t$ include all relevant variables although extraneous variables may be included.

The model (2.1) is frequently expressed in matrix notation as

$$Y = X\beta + e. \tag{2.2}$$

Here Y is the n-vector of observed responses, X is the design matrix of dimension $n \times (t + 1)$ as defined by (2.1), assumed to have rank $t + 1$, and β is the $(t + 1)$-vector of unknown regression coefficients. In certain situations, it will be convenient to assume that all variables have been expressed as deviations from their observed sample means and in still other cases the variables will also be assumed to have unit sum of squares. In such cases, we shall use (2.2) to denote the model but emphasize the $(n \times t)$ matrix is respectively the "adjusted" design matrix or the "standardized" design matrix. The definition of β will be assumed consistent with that of X.

In the variable selection problem, let r denote the number of terms which are deleted from model (2.1), that is, the number of coefficients which are set to zero. The number of terms which are retained in the final equation will be denoted by $p = t + 1 - r$. Note that the intercept term, β_0 , is included and is hence eligible for deletion although typically it is forced into the equation. In this paper, it will be assumed that β_0 is forced into the equation; hence, the number of variables in the subset equation is $p - 1$. More generally, the analyst may wish to force several terms into the final equation or conditionally force terms into the equation, e.g. the linear term x might be forced in if the quadratic term x^2 is selected for inclusion.

If it is not clear from the context, the convention used is that statistics associated with a p-term model will be subscripted by p while those associated with the full $(t + 1)$-term model will not be subscripted. For example, RSS will denote the residual sum of squares for the least squares fit of the full model and RSS_p will denote the corresponding quantity for a p-term subset model.

In the course of the discussion, it will be convenient to refer to the p-term model with minimum RSS_p among all possible p-term models as the "best" model of size p. It should be emphasized that "best" is defined only in this sense and that the model may, indeed, not be best as a function of its intended use. In addition, it should be emphasized that this definition of best is only applied to the current sample and does not imply that the same relation holds for the population.

2.2. Consequences of Incorrect Model Specification.

There are a variety of practical and economical reasons for reducing the number of independent variables in the final equation. In addition, variable deletion may be desirable in terms of the statistical properties of the parameter estimates and the estimate of the final equation. This section provides a brief review of the consequences of incorrectly specifying the model either in terms of retaining extraneous variables or deleting relevant variables. The properties described here are dependent on the assumption that the subset of variables under consideration has been selected without reference to the data. Since this is contrary to normal practice, the results should be used with caution.

Let the model (2.1) be written in matrix form as

$$Y = X_p\beta_p + X_r\beta_r + e \tag{2.3}$$

where the X matrix has been partitioned into X_p of dimension $n \times p$ and X_r of dimension $n \times r$. The β vector is partitioned conformably. Let $\hat{\beta}$, with components $\hat{\beta}_p$ and $\hat{\beta}_r$, denote the least squares estimate of β and let $\tilde{\beta}_p$ denote the subset least squares estimate of β_p if the variables in X_r are deleted from the model. That is,

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2.4}$$

and

$$\tilde{\beta}_p = (X_p'X_p)^{-1}X_p'Y. \tag{2.5}$$

Further, let $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ represent the residual mean squares for the two situations. Specifically,

$$\hat{\sigma}^2 = Y'(I - X(X'X)^{-1}X')Y/(n - t - 1) \tag{2.6}$$

and

$$\tilde{\sigma}^2 = Y'(I - X_p'X_p)^{-1}X_p')Y/(n - p). \tag{2.7}$$

If model (2.3) is correct, the properties of $\hat{\beta}$ and $\hat{\sigma}^2$ as estimates of β and σ^2 are well known from general linear model theory. In particular, $\hat{\beta}$ and $\hat{\sigma}^2$ are minimum variance unbiased estimators with $\hat{\beta} \sim N(\beta, (X'X)^{-1}\sigma^2)$ and $(n - t - 1)\hat{\sigma}^2 \sim \sigma^2\chi^2(n - t - 1)$.

The properties of $\tilde{\beta}_p$ and $\tilde{\sigma}^2$ have been described by several authors with recent results given by Walls and Weeks [1969], Rao [1971], Narula and Ramberg [1972], Rosenberg and Levy [1972], and Hocking [1974]. If we let

$$A = (X_p'X_p)^{-1}X_p'X_r \tag{2.8}$$

then $\tilde{\beta}_p$ is normally distributed with

$$E(\tilde{\beta}_p) = \beta_p + A\beta_r \tag{2.9}$$

and

$$\text{VAR } (\tilde{\beta}_p) = (X_p'X_p)^{-1}\sigma^2. \tag{2.10}$$

The mean squared error is given by

$$\begin{aligned} \text{MSE } (\tilde{\beta}_p) &= E(\tilde{\beta}_p - \beta_p)(\tilde{\beta}_p - \beta_p)' \\ &= (X_p'X_p)^{-1}\sigma^2 + A\beta_r\beta_r'A'. \end{aligned} \tag{2.11}$$

Also, $(n - p)\tilde{\sigma}^2/\sigma^2$ is distributed as non-central chi-squared with

$$E(\tilde{\sigma}^2) = \sigma^2 + \beta_r'X_r'(I - X_p(X_p'X_p)^{-1}X_p')X_r\beta_r/(n - p). \tag{2.12}$$

The following properties are then easily established:

1. $\tilde{\beta}_p$ is generally biased, interesting exceptional cases being (a) $\beta_r = 0$ and (b) $X_p'X_r = 0$.
2. The matrix VAR $(\hat{\beta}_p)$ − VAR $(\tilde{\beta}_p)$ is positive semi-definite. That is, the estimates of the components of β_p given by $\hat{\beta}_p$ are generally more variable than those given by $\tilde{\beta}_p$.
3. If the matrix VAR $(\hat{\beta}_r) - \beta_r\beta_r'$ is positive semi-definite, then the matrix VAR $(\hat{\beta}_p)$ − MSE $(\tilde{\beta}_p)$ is positive semi-definite.
4. $\tilde{\sigma}^2$ is generally biased upward.

The regression equation is frequently used to predict the response to a particular input, say $x' = (x_p'x_r')$. If we use the full model then the predicted value of the response is $\hat{y} = x'\hat{\beta}$ which has mean $x'\beta$ and prediction variance

$$\text{VARP } (\hat{y}) = \sigma^2(1 + x'(X'X)^{-1}x). \tag{2.13}$$

On the other hand, if the subset model with x_r deleted is used, the predicted response is $\tilde{y}_p = x_p'\tilde{\beta}_p$ with mean

$$E(\tilde{y}_p) = x_p'\beta_p + x_p'A\beta_r \tag{2.14}$$

and

$$\text{VARP } (\tilde{y}_p) = \sigma^2(1 + x_p'(X_p'X_p)^{-1}x_p). \tag{2.15}$$

The prediction mean squared error is given by

$$\begin{aligned} \text{MSEP } (\tilde{y}) &= E(y - \tilde{y}_p)^2 \\ &= \sigma^2(1 + x_p'(X_p'X_p)^{-1}x_p) + (x_p'A\beta_r - x_r'\beta_r)^2. \end{aligned} \tag{2.16}$$

The following properties are then easily established:

5. $\tilde{y}$ is biased unless $X_p'X_r\beta_r = 0$.

6. VARP $(\hat{Y}) \geq$ VARP $(\tilde{y}_p)$.

7. If the matrix VAR $(\hat{\beta}_r) - \beta_r\beta_r'$ is positive semi-definite, then VARP $(\hat{y}) \geq$ MSEP $(\tilde{y}_p)$.

The motivation for variable elimination is provided by properties 2 and 6. That is, even if $\beta_r \neq 0$, β_p may be estimated or future responses may be predicted with smaller variance using the subset model. The penalty is in the bias. In the sense of mean squared error, properties 3 and 7 describe a condition under which the gain in precision is not offset by the bias.

On the other hand, if the variables in X_r are extraneous, that is, $\beta_r = 0$, then properties 2 and 6 indicate a loss of precision in estimation and prediction if these variables are included.

3. COMPUTATIONAL TECHNIQUES

As mentioned in the introduction, this paper will consider the general problem of trying to determine the relations between the input variables, x_i , and their roles, either alone or in conjunction with others, in describing the response, y. One objective of this analysis may well be the selection of a subset of the input variables to be used in a final equation.

To provide the information for such an analysis, one is quite naturally led to consider fitting models with various combinations of the input variables. If the number of inputs, t, is small, one might consider all 2^t combinations assuming β_0 is forced in, but for large t that is economically out of the question.

This section contains a discussion of a number of computational procedures which will provide information on some or all of the subset combinations. Attention is focused primarily on least squares fitting, but mention is made of ridge regression. It should be emphasized that this phase of the problem is viewed as primarily computational. A discussion of how to interpret the output is given in Section 4.

3.1. All Possible Regressions.

If t is not too large, fitting all possible models might be considered, that is, the t models in which only one of the inputs is included, the $\binom{t}{2}$ models in which each pair of inputs is included and so on up to the single model containing all t inputs. Prior to the advent of high-speed computers, such a solution was out of the question for problems involving more than a few variables. The availability of rapid computation has inspired efforts in this direction and there now exist a number of very efficient algorithms for evaluating all possible regressions. One of the earliest, due to Garside [1965], was capable of efficiently handling problems with ten to twelve independent variables. More recently algorithms have been proposed by Schatzoff *et al.* [1968], Furnival [1971], and Morgan and Tatar [1972] which no doubt extend this range.

The basic idea in all of these papers is to perform the computations on the 2^t subsets in such a way that consecutive subsets differ in only one variable. Thus the Jordan reduction or Beaton "sweep operator," (see Beaton [1964]) may be efficiently used to perform the computations. Garside [1965] described an ordering of the subsets such that all subsets will be fit in 2^t sweeps. The papers by Schatzoff *et al.* [1968] and Morgan and Tatar [1972] offer slight modifications, the latter emphasizing that the amount of computation can be substantially decreased by not evaluating the regression coefficients for each subset but rather just the residual sum of squares. Furnival [1971] makes this same point in presenting two algorithms, one based on the sweep operator and the other on Gaussian elimination using only the forward solution, hence, avoiding computation of $(X'X)^{-1}$ and $\hat{\beta}$.

An interesting procedure for evaluating all possible subsets, which is distinct from these, was proposed by Newton and Spurrell [1967a]. Noting that regression sum of squares for any subset is the sum of "basic elements," they developed a scheme for evaluating these basic elements without evaluating all subsets. For example, with four variables only five subsets need to be evaluated to determine the basic elements and, hence, the regression sum of squares for any subset. In addition, they introduced the concept of "element analysis" as a means of identifying the roles of the variables, both alone and in relation to other variables. This technique is illustrated further in Newton and Spurrell [1967b].

A comparison of these algorithms is not a simple matter, but if done it must be based on considerations such as storage requirements, number of computations, computer time, accuracy and amount of information given. No attempt has been made in this regard beyond the comparisons made in the references. If the intent is just to screen the subsets based on residual sum of squares, however, the second Furnival algorithm seems quite efficient.

3.2. Stepwise Methods.

Because of the computational task of evaluating all possible regressions, various methods have been proposed for evaluating only a small number of subsets by either adding or deleting variables one at a time according to a specific criterion. (The sweep operator is efficiently used to perform the computations.) These procedures, which are generally referred to as stepwise methods, consist of variations on two basic ideas called Forward Selection (FS) and Backward Elimination (BE). (See e.g. Efroymson [1966] or Draper and Smith [1966].) A brief discussion of these two ideas follows:

Forward Selection. This technique starts with no variables in the equation and adds one variable at a time until either all variables are in or until a stopping criterion is satisfied. The variable considered for inclusion at any step is the one yielding the largest single degree of freedom (d.f.) F-ratio among those eligible for inclusion. That is, variable i is added to the p-term equation if

$$F_i = \max_i \left(\frac{\mathrm{RSS}_p - \mathrm{RSS}_{p+i}}{\hat{\sigma}_{p+i}^2} \right) > F_{\text{in}} . \tag{3.1}$$

Here the subscript $(p + i)$ refers to quantities computed when variable i is adjoined to the current p-term equation. The specification of the quantity F_{in} results in a rule for terminating the computations. Section 4.6 contains a brief summary of some of the common stopping rules and the results of a simulation study on Forward Selection.

Backward Elimination. Starting with the equation in which all variables are included, variables are eliminated one at a time. At any step, the variable with smallest F-ratio, as computed from the current regression, is eliminated if this F-ratio does not exceed a specified value. That is, variable i is deleted from the p-term equation if

$$F_i = \min_i \left(\frac{\mathrm{RSS}_{p-i} - \mathrm{RSS}_p}{\hat{\sigma}_p^2} \right) < F_{\text{out}} . \tag{3.2}$$

Here RSS_{p-i} denotes the residual sum of squares obtained when variable i is deleted from the current p-term equation. Again, several stopping rules similar to those for Forward Selection have been suggested for determining F_{out} .

These two basic ideas suggest a number of combinations, the most popular being that described by Efroymson [1960] and denoted as ES. This method is basically FS but at each step the possibility of deleting a variable as in BE is considered. (It should be noted that

the term stepwise regression is frequently used to refer specifically to the Efroymson procedure as opposed to the more general meaning used here.)

The stepwise procedures have been criticized on many counts, the most common being that neither FS, BE or ES will assure, with the obvious exceptions, that the "best" subset of a given size will be revealed. Some users have recommended that both FS and BE be performed in the hope of seeing some agreement. Oosterhoff [1963] observes that they need not agree for any value of p except $p = t + 1$. Mantel [1970] critizes FS by illustrating a situation in which an excellent model would be overlooked because of the restriction of adding only one variable at a time.

Another criticism of FS and BE often cited is that they imply an order of importance to the variables. This can be misleading since, for example, it is not uncommon to find that the first variable included in FS is quite unnecessary in the presence of other variables. Similarly, it is easily demonstrated that the first variable deleted in BE can be the first variable included in FS. In defense of the original proponents of stepwise methods, it should be noted that much of the criticism has been directed at properties which were never claimed by the originators. It is unfortunate that many users have attached significance to the order of entry or deletion and assumed optimality of the resulting subset.

The lack of satisfaction of any reasonable optimality criterion by the subsets revealed by stepwise methods, although a valid criticism, may not be as serious a deficiency as the fact that typical computer routines usually reveal only one subset of a given size. As noted by Mantel [1970] and Beale [1970a], if all t input variables are brought in by FS, a total of $t(t + 1)/2$ equations are actually fitted. Similarly, in BE, assuming that it is continued until only one variable remains, only t equations are actually fitted but the residual sum of squares has been computed for $t(t + 1)/2$ subsets. Although $t(t + 1)/2$ may be a small fraction of 2^t, the intuitive basis for these procedures suggests that in moderately well behaved problems the subsets revealed may agree with the best subsets obtained by evaluating all possible regressions. There are, of course, notable exceptions such as that reported by Gugel [1972] in which he observed an improvement of over 37 percent in the value of the squared multiple correlation coefficient when comparing the best subset with that revealed by ES.

As observed by Gorman and Toman [1966], it is unlikely that there is a single best subset but rather several equally good ones. This, coupled with our desire to provide the user with information so that he may obtain insight into the structure of his data, suggests that an evaluation of a fairly large number of subsets might be desirable. An ideal situation would be one in which it could be guaranteed that the best subset of each size and a number of nearly best subsets are observed, without the expense of evaluating all possible subsets. The extent to which this is possible is described in the next section.

3.3. Optimal Subsets.

There is an elementary but fundamental principle in constrained minimization problems which says that if additional constraints are adjoined to a problem, the optimum value of the objective function will be as large or larger than that obtained in the original problem. To see how this idea applies to the subset selection problem, note that the problem may be described as that of minimizing the residual sum of squares for the full model, subject to the restriction that certain of the coefficients are zero. That is,

$$\begin{aligned} &\text{minimize } Q(\beta) \\ &\text{subject to } \beta_i = 0 \qquad i \in R. \end{aligned} \tag{3.3}$$

Here $Q(\beta) = (Y - X\beta)'(Y - X\beta)$ is the residual sum of squares and R is a particular set of indices. The optimality principle states that if R_1 and R_2 are two index sets and Q_1 and Q_2 the corresponding residual sums of squares, then, if R_1 is a subset of R_2 it follows that $Q_1 \leq Q_2$.

Several authors, including Hocking and Leslie [1967], Beale *et al.* [1967], Kirton [1967], Beale [1970b], LaMotte and Hocking [1970], and Furnival and Wilson [1974], have used this principle to develop algorithms which will ensure that the best subset of each size for $1 \leq p \leq t + 1$ will be identified while evaluating only a small fraction of the 2^t subsets.

To illustrate the basic concept, the method described by Hocking and Leslie [1967] will be summarized. Suppose the equation for the full model has been fit and assume that the variables are labelled according to the magnitude of their t-statistics. That is, variable 1 has the smallest t-statistic, variable 2 the next smallest, etc. Now suppose the objective is to identify the best set of four variables to be deleted. Let $Q(5)$ denote the residual sum of squares if variable 5 is deleted and let $Q(1, 2, 3, 4)$ denote the residual sum of squares if variables 1, 2, 3 and 4 are deleted. Then if $Q(1, 2, 3, 4) \leq Q(5)$, the residual sum of squares for deleting any other set of four variables will be at least as great as $Q(1, 2, 3, 4)$. This follows since such a residual sum of squares is guaranteed to be at least as great as $Q(5)$, hence, no other subsets need to be evaluated. If $Q(1, 2, 3, 4) > Q(5)$, additional subsets need to be evaluated.

Extensions of this simple idea were developed by LaMotte and Hocking [1970] and incorporated into a computer program called SELECT (LaMotte [1972]). Early versions of this program were generally inefficient if $t > 30$ but recent efforts have greatly improved the program. One user reported the analysis of a 70-variable problem with a moderate amount of computation. The program described by Furnival and Wilson [1974] is similar to SELECT but performs the computations in a more efficient manner.

To provide an indication of the effectiveness of the optimality principle as used in SELECT, note that for the 15-variable data reported by McDonald and Schwing [1973], the determination of a best subset of each size required the evaluation of only 1,465 subsets as opposed to a possible $2^{15} = 32{,}768$. For the 26-variable data reported by LaMotte and Hocking [1970], SELECT required the evaluation of 3,546 out of a total of 67,108,864 possible subsets.

As in the computation of all possible regressions, the amount of computation required is a function of how much information is desired. For example, it seems reasonable to do a preliminary run determining only the residual sum of squares and then for a rather small number of subsets compute more detailed information, such as values of regression coefficients, etc. The Furnival and Wilson [1974] algorithm utilizes this approach to substantially reduce the total amount of computation.

With the optimal regression programs there is another consideration. In addition to the best subset of size p, a number of "nearly best" subsets are identified. It is natural to ask if the next best subset is included in the output. Although there is generally no guarantee of this, it is frequently the case, being more likely with the less efficient algorithm described by Hocking and Leslie [1967] than with the SELECT algorithm. The point is that by specifying values for certain program parameters, more information can be obtained but at an increased cost. The Furnival and Wilson [1974] program contains an option for guaranteeing the m-best subsets rather that the single best subset. For $m = 10$, the amount of computation is approximately doubled if this option is invoked.

The number of subsets which must be evaluated to determine the optimum subsets for these algorithms is highly dependent on the data. This should be contrasted with the

Newton and Spurrell [1967a] algorithm for which the number of subsets required to determine the basic elements depends only on the number of variables. The possibility of combining both of these concepts is of interest but has not been considered.

3.4. Sub-Optimal Methods.

As a compromise between the limited output of stepwise procedures and the guaranteed results of the optimal procedures, Gorman and Toman [1966] proposed a procedure based on a fractional factorial scheme in an effort to identify the better models with a moderate amount of computation. With the same objective, Barr and Goodnight [1971] in the Statistical Analysis System (SAS) regression program proposed a scheme based on maximum-R^2-improvement. This is essentially an extension of the stepwise concept but the search is more extensive. For example, to determine the best p-term equation, starting with a given $(p-1)$-term equation, the currently excluded variable causing the greatest increase in R^2 is adjoined to that subset. Given this subset, a comparison is made to see if replacing a variable by one currently excluded will increase R^2. If so, the best switch is made. This process is continued until it is found that no switch will increase R^2. The resulting p-term equation is thus labelled "best," but it should be emphasized that this subset can be inferior to the one determined by SELECT.

Although no effort has been made to compare these methods with the optimal methods, it is not surprising to note that several users have reported situations in which best subsets were missed.

3.5. Ridge Regression.

Hoerl and Kennard [1970a] suggested the biased "ridge" estimator for problems involving non-orthogonal predictors. In particular, they considered the estimator

$$\beta(k) = (X'X + kI)^{-1}X'Y \tag{3.4}$$

where X is in standardized form. The constant k is to be determined by inspection of the "ridge trace," that is, plots of $\beta(k)$ versus k. A more detailed discussion of the ridge estimator appears in Section 5. In the context of the present section, note that although ridge regression is not designed for the purpose of variable selection, there is an inherent deletion of variables, namely those whose coefficients from (3.4) go to zero rapidly with increasing k. Hoerl and Kennard [1970b] suggested that such variables "cannot hold their predicting power" and should be eliminated. With respect to computational considerations, ridge regression is quite efficient since reasonably good ridge plots can be obtained using only a few values of k. The difficult question, which is discussed in Section 5, is that of determining the value of k, and of course, the question of how small must $\beta_i(k)$ be to justify deleting x_i . Marquardt [1974] suggested that variable deletion is not a zero-one situation but rather, that all variables might be retained with decreased influence if $\beta_i(k)$ is small. This, of course, ignores the economic and practical motives for deleting variables.

3.6. Examples.

In this section two examples are presented to illustrate the possible differences between SELECT, FS, and BE as computational procedures. These examples are considered again in later sections to illustrate various other problems. For convenience, all data are analyzed in standard form; hence, for example, the values for the residual mean squares (RMS_p) and the regression coefficients need to be scaled up for comparison with the references from which the data were taken.

3.6.1 Example 1: Gas Mileage Data. In an attempt to predict gasoline mileage for 1973–1974 automobiles, road tests were performed by "Motor Trend" magazine in which gasoline mileage and ten physical characteristics of various types of automobiles were recorded. The data were taken from the March, April, June and July issues of 1974. This example was suggested by Dr. R. J. Freund, Institute of Statistics, Texas A&M University.

A description of the variables is given in Table 1. The correlation matrix and eigenvalues of $X'X$ are shown in Table 2. To illustrate some of the points made in this section, the data were run on the 1970 version of SELECT, a FS program, a BE program and, as a check, all possible regressions were evaluated.

SELECT revealed, at least, the best four subsets for all values of p except $p = 8$ where only one subset was evaluated and $p = 6$ where the second and fourth best were not among those evaluated. BE gave the best subset for all cases except $p = 3$ where the third best was obtained. FS did somewhat worse, disagreeing with SELECT for all except $p = 2$ and $p = 3$. A summary of the results for SELECT and FS is given in Table 3. The column labelled VARIABLES indicates the variables added or deleted (−) as p is increased. The last column shows the rank of the subset obtained by FS relative to the best subset. For example, with $p = 5$, FS yielded the tenth best subset.

Inspection of the SELECT output suggests that variables 3, 9 and 10 play a fundamental role in predicting gasoline mileage. Indeed, the results of Section 4.9.1 indicate that this subset may be best for prediction. It is of interest to note that variable 9 is the best single variable, but variables 3 and 10 rank seventh and tenth, respectively, when used alone. It is not until three variables are allowed that their combined effect is observed. FS selected variable 2 as its second choice and as a result was led astray and failed to recognize the role of variables 3, 9 and 10. It is also noted that variable 2, which is the second best single variable, was the first variable deleted by BE. As a result, BE was in closer agreement with the optimal choice.

3.6.2. Example 2: Air Pollution Data. The data for this example ($t = 15$) are taken from McDonald and Schwing [1973] and the reader is referred to that paper for details. Their paper contained a discussion of the SELECT results and compared subset regression with ridge analysis. These topics will be discussed in Sections 4.9.2 and 5.5.2. In this section attention is directed to the computational techniques.

Whereas the gas mileage example indicated that BE was superior to FS with respect to identifying best subsets, the situation is reversed for this example. FS agrees with SELECT for all cases except $p = 5$, 10 and 11, where the FS subset is among the first five. On the other hand, BE is not optimal for $p = 3$ through 9 as indicated in the partial summary in Table 4. In this case the rank indicated is relative to the subsets observed by

TABLE 1

DESCRIPTION OF VARIABLES FOR THE GAS MILEAGE DATA
(32 OBSERVATIONS)

Number	Description
X1	Engine Shape (Straight (1) or V(0))
X2	Number of Cylinders
X3	Transmission Type (Manual (1) or Auto (0))
X4	Number of Transmission Speeds
X5	Engine Size (Cubic Inches)
X6	Horsepower
X7	Number of Carburetor Barrels
X8	Final Drive Ratio
X9	Weight (Pounds)
X10	Quarter Mile Time (Seconds)
Y	Gasoline Mileage (MPG)

TABLE 2

CORRELATION MATRIX AND EIGENVALUES OF $X'X$ FOR GAS MILEAGE DATA

	Correlations									
X1	1.000									
X2	-0.811	1.000								
X3	0.168	-0.523	1.000							
X4	0.206	-0.493	0.794	1.000						
X5	-0.710	0.902	-0.591	-0.556	1.000					
X6	-0.723	0.832	-0.243	-0.126	0.791	1.000				
X7	-0.570	0.527	0.058	0.274	0.395	0.750	1.000			
X8	0.440	-0.700	0.713	0.700	-0.710	-0.449	-0.091	1.000		
X9	-0.555	0.782	-0.692	-0.583	0.888	0.659	0.428	-0.712	1.000	
X10	0.745	-0.591	-0.230	-0.213	-0.434	-0.708	-0.656	0.091	-0.175	1.000
Y	0.664	-0.852	0.600	0.480	-0.848	-0.776	-0.551	0.681	-0.868	0.419
	Eigenvalues									
	5.760	2.650	0.597	0.270	0.222	0.210	0.133	0.081	0.054	0.024

SELECT. There may be many other intermediate subsets which would be revealed by an all possible algorithm. BE is led astray at $p = 11$ by eliminating variable 14. While this choice is optimal at that time, the permanent removal of variable 14 is not desirable as p is decreased. It appears that variables 12 and 13 are a substitute for variable 14 at $p = 11$, but for smaller values of p variables 12 and 13 are less effective. The role of these two variables is examined in Section 5.5.2.

TABLE 3

A COMPARISON OF SELECT AND FORWARD SELECTION FOR THE GAS MILEAGE DATA ($RMS_p \times 10^3$)

	SELECT		FORWARD SELECTION		
p	VARIABLES	RMS_p	VARIABLES	RMS_p	RANK
2	9	8.24	9	8.24	1
3	2	5.85	2	5.85	1
4	3,10,-2	5.37	6	5.62	2
5	6	5.27	3	5.59	10
6	5	5.24	10	5.46	7
7	8	5.33	5	5.36	3
8	4	5.50	8	5.51	2
9	7	5.71	4	5.72	3
10	1	5.96	7	5.96	2
11	2	6.24	1	6.24	1

TABLE 4

A COMPARISON OF SELECT AND BACKWARD ELIMINATION FOR THE AIR POLLUTION DATA ($RMS_p \times 10^3$)

SELECT			BACKWARD ELIMINATION		
p	VARIABLES	RMS_p	VARIABLES	RMS_p	RANK
2	9	10.10	9	10.10	1
3	6	7.67	12	9.82	10
4	2	6.44	13	7.19	8
5	1,14,-6	5.51	6	6.22	9
6	6	5.24	2	5.60	> 10
7	3	5.00	5	5.40	> 10
8	5	4.92	3	5.10	7
9	4	4.93	1	4.99	3
10	12,13,-14	4.86	4	4.86	1

4. SELECTION CRITERIA

4.1. Users of Regression.

The availability of good computer algorithms for computing subset regressions now raises the question of how the information should be evaluated. As emphasized by Lindley [1968], the criterion used to decide on the appropriate subset or subsets should certainly be related to the intended use. Mallows [1973b] provided the following list of potential uses of the regression equation:

a. *Pure Description*
b. *Prediction and Estimation*
c. *Extrapolation*
d. *Estimation of Parameters*
e. *Control*
f. *Model Building.*

Although these terms should be self-explanatory, a brief explanation is given in the following paragraphs.

If the objective is to obtain a good description of the response variable and the criterion for fitting the data is least squares, then a search for equations with small residual sums of squares is indicated. In this sense, the best solution is to retain all variables but in some cases little will be sacrificed if some variables are deleted. Most users would prefer to look at the squared multiple correlation coefficient, R^2, (defined below) as an equivalent measure which is between zero and one and, hence, appears to be easier to interpret. In this regard, we mention the paper by Crocker [1972] in which it is suggested that the statistical significance of R^2 may not give a true picture of the adequacy of the model. The recommendation of that paper is that, in some cases, it may be more appropriate to consider the percent reduction in standard deviation of the response variable achieved by the model. Another

limitation of R^2, noted by Barrett [1974], is that for fixed residual sum of squares, R^2 increases with the steepness of the regression surface.

The distinction between prediction of a future response and the estimation of the mean response for a given input is recognized in most texts on regression. The important issue here is that the variance of the estimate of the mean response is given by

$$\text{VAR}\ (x'\hat{\beta}) = \sigma^2 x'(X'X)^{-1}x \tag{4.1}$$

in contrast with the expression (2.13) for the prediction variance. It is clear that in the case of prediction, the contribution to the prediction variance due to the variability in estimating the coefficients, namely equation (4.1), may be small relative to the inherent variability of the system being studied.

The danger of extrapolating beyond the range of the data used to develop the estimates is apparent since the current model may no longer apply. However, even if the model is appropriate, a predictor which is adequate within this range may be very poor outside of this region because of poor parameter estimates resulting from near degeneracy of the X-matrix (See Mason *et al.* [1975].)

If parameter estimation is the objective, then one should consider the bias resulting from deleting variables as well as the estimated variance. Again if X is nearly degenerate, several authors recommend biased estimates which, in addition to giving better parameter estimates, may lead to a predictive equation which is more effective in extrapolation. (See Section 5.)

The concept of control, as defined by Draper and Smith [1966], is concerned with controlling the level of output by varying the level of the inputs. In this case, accurate estimates of the regression coefficients are desirable.

In many studies, the objective of the study is to develop a model for the response as a function of the observed inputs and various functions of these inputs. In this situation, it would appear that computational methods for evaluating subset regressions could be profitably used in an interactive mode to reveal relations between sets of variables.

4.2. Criteria Functions.

With the objectives of Section 4.1 in mind, a number of criteria have been proposed for deciding on an appropriate subset. These criteria are stated in terms of the behavior of certain functions as a function of the variables included in the subset. Many of these criteria functions are simple functions of the residual sum of squares for the p-term equation denoted by RSS_p . Some of the more common ones are

1. The residual mean square,

$$\text{RMS}_p = \frac{\text{RSS}_p}{n - p}.$$

2. The squared multiple correlation coefficient,

$$R_p^2 = 1 - \frac{\text{RSS}_p}{\text{TSS}}.$$

3. The adjusted R^2,

$$\bar{R}_p^2 = 1 - (n - 1)(1 - R_p^2)/(n - p).$$

4. The average prediction variance,

$$J_p = (n + p)\ \mathrm{RMS}_p/n.$$

(Ref. Mallows [1967], Rothman [1968], and Hocking [1972].)

5. The total squared error,

$$C_p = \mathrm{RSS}_r/\hat{\sigma}^2 + 2p - n.$$

(Ref. Gorman and Toman [1966], Mallows [1973a].)

6. The average prediction mean squared error,

$$S_p = \mathrm{RMS}_p/(n - p - 1).$$

(Ref. Tukey [1967], Sclove [1971].)

7. The standardized residual sum of squares,

$$\mathrm{RSS}_p^* = e_p' D_p^{-1} e_p, \quad \text{where} \quad e_p = Y - \hat{Y}_p \quad \text{and}$$
$$D_p = \mathrm{DIAG}\ (I - X_p(X_p'X_p)^{-1}X_p').$$

(Ref. Schmidt [1973a].)

8. The prediction sum of squares,

$$\mathrm{PRESS} = e_p' D_p^{-2} e_p\ .$$

(Ref. Allen [1971b], Schmidt [1973a] and Stone [1974].)

The question of how these functions should be used and which criterion is appropriate in view of the intended use remains to be answered. An attempt is made in the following paragraphs to provide some general guidelines.

4.3. The Evaluation of Subset Regressions.

Prediction and parameter estimation are two of the more frequent goals of regression analysis. Recall from Section 2 that if the matrix VAR $(\hat{\beta}_r) - \beta_r\beta_r'$ is positive semi-definite, then it is possible to estimate parameters and predict responses with smaller mean squared error using the subset equation. In particular, writing VAR $(\hat{\beta}_r) = B_{rr}\sigma^2$ where B_{rr} is the appropriate submatrix of $X'X^{-1}$, the required condition is satisfied if

$$\beta_r' B_{rr}^{-1} \beta_r/\sigma^2 \leq 1. \tag{4.2}$$

Of course, the parameters β_r and σ^2 are unknown, but if they are estimated from the current data using the full model, the condition (4.2) can be stated in terms of the F-statistic associated with testing the hypothesis $\beta_r = 0$. In particular,

$$F = \frac{\hat{\beta}_r' B_{rr}^{-1} \hat{\beta}_r}{r\hat{\sigma}^2} \leq 1/r. \tag{4.3}$$

Thus, assuming that the t-variate model equation (2.3) is correct, then, based on using the current data for fitting the equation, it seems reasonable to delete the variables in X_r if condition (4.3) is satisfied. The claim here is that with respect to mean squared error the subset equation will yield better estimates of the parameters, β_p and also yield a better prediction equation. Further, since this result is true for any input vector x, extrapolation beyond the range of the current data is permissible. The user should proceed with caution when extrapolating beyond the range of the current data. The results of Section 2 were based on the assumptions that (i) the t-variate model was valid for all x and (ii) the p-term

subset was selected without reference to the data. These conditions are rarely met in practice.

A commonly used criterion for deleting variables (see e.g. Efroymson [1966]) is that the t-statistics associated with the parameter estimates for the full model be less than one in absolute value. This criterion has a basis in the present development since a necessary condition for positive semi-definiteness of $B_{rr}\hat{\sigma}^2 - \hat{\beta}_r\hat{\beta}_r'$ is that the t-statistics associated with the r parameters in β_r are less than one in magnitude. It is clear that condition (4.3) is more restrictive.

Pursuing the distinction between predicting in the neighborhood of the current data and extrapolating outside of this region, and still assuming that model (2.3) is correct, it may be argued that condition (4.3) is appropriate for extrapolation but too restrictive for prediction since it applies for any input vector, x. The requirement that VARP $(\hat{y}_i)$ − MSEP $(\tilde{y}_{p_i})$, when averaged over a specified set of inputs, be non-negative seems like a reasonable compromise. In particular, using the current data, X, yields

$$\frac{1}{n}\sum_{i=1}^{n}(\text{VARP}\ (\hat{y}_i) - \text{MSEP}\ (\tilde{y}_{p_i})) = \frac{r}{n}\sigma^2(1 - \beta_r'B_{rr}^{-1}\beta_r/r\sigma^2). \tag{4.4}$$

Replacing the parameters in (4.4) by their estimates when fitting the full equation yields the following condition:

$$F = \frac{\hat{\beta}_r'B_{rr}^{-1}\hat{\beta}_r}{r\hat{\sigma}^2} \leq 1. \tag{4.5}$$

Based on this discussion, recognizing the ideal conditions under which the results were developed, one might consider using (4.3) if extrapolation is the objective and the less restrictive condition (4.5), allowing the deletion of more variables, if prediction is the objective. If the primary concern is accurate estimates of the regression coefficients, β_p , then satisfaction of (4.3) is demanded.

The discussion thus far has focused on conditions which offer improvement in prediction and estimation by using a subset model. As a measure of the degree of improvement relative to the full model, define the relative gain for prediction as

$$\text{RGP} = \frac{\text{VARP}\ (\hat{y}) - \text{MSEP}\ (\tilde{y}_p)}{\text{VARP}\ (\hat{y})}. \tag{4.6}$$

The relative gain gives an indication of the decrease in the width of the prediction interval for the subset model. This concept also allows the assessment of subsets which might give an increase in the prediction interval width but are desirable for other reasons. Thus, (4.6) might be negative but the loss in precision might be offset by other considerations.

To illustrate, consider the evaluation of (4.6) when considering the average performance over the current data. In view of (4.4),

$$\text{RGP} = r(1 - \beta_r'B_{rr}^{-1}\beta_r/r\sigma^2)/(n + t + 1). \tag{4.7}$$

The role of sample size in this expression is deceptive. It appears that for large n, the relative loss in precision might be small even though important variables are eliminated. Recall, however, that for a given model $\beta_r'B_{rr}^{-1}\beta_r$ will also increase with increasing sample size.

If the objective is to estimate mean response, the relative gain may be defined similarly. In terms of averaging over the current data, the expression for relative gain is the same as (4.7) with the exception that the denominator is replaced by $t + 1$. The difference reflects

the fact that in the prediction of a single response, the inherent variability in the system may dominate the variability due to estimating the regression coefficients.

With this discussion as background, we now turn to a discussion of how the previously mentioned selection criteria might be used. It has been argued that, since the first six criteria are all simple functions of the residual sums of squares, it makes no difference which one is used. This is true, of course, but it is important to establish how the criteria should be interpreted. Of course, all of these recommendations must be considered heuristic since the exact properties of these procedures have not been developed.

4.4. Interpretation of C_p-Plots.

The C_p-statistic has been selected to illustrate the concepts discussed in Section 4.3 because plots of C_p versus p appear to lend themselves to easy interpretation. Other statistics will be related to C_p so that the analagous interpretations can be made.

By way of review, recall that the C_p-statistic and the interpretation of C_p-plots were initially described by Mallows [1964] and [1966] and subsequently discussed by Gorman and Toman [1966], Daniel and Wood [1971], and Mallows [1973a]. C_p is an estimate of the standardized total mean squared error of estimation for the current data, X. Denoting this by Γ_p yields

$$\Gamma_p = \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{MSE}\,(\tilde{y}_i) = \frac{E(\text{RSS}_p)}{\sigma^2} + 2p - n. \tag{4.8}$$

(Note that the total mean squared error of prediction, obtained by using MSEP $(\tilde{y}_i)$ is $\Gamma_p + n$.) Replacing $E(\text{RSS}_p)$ and σ^2 by appropriate estimates yields

$$C_p = \frac{\text{RSS}_p}{\hat{\sigma}^2} + 2p - n. \tag{4.9}$$

A plot of C_p as a function of p for all subsets or at least the contending subsets is recommended by Mallows [1973a] as a means of providing information about the structure of the problem. With respect to subset selection, it is suggested that subsets with small C_p , and C_p close to p be considered, the latter condition indicating small bias.

In order to properly interpret C_p in terms of the intended use, some guidelines are necessary. Mallows [1973a] provides some assistance in the interpretation of C_p-plots and shows how they may be calibrated. To relate to the current development, observe that if $\hat{\sigma}^2$ in (4.9) is the residual mean square from the full model then the relation between C_p and the F-ratio in (4.3) and (4.5) is

$$C_p - p = r(F - 1). \tag{4.10}$$

The conditions (4.3) and (4.5) can then be easily translated into conditions on C_p . Noting that C_p is bounded below by $p - r$, the line of constant RRS_p , subsets satisfying (4.3) which might be appropriate for extrapolation and parameter estimation will satisfy the condition

$$2p - t - 1 \leq C_p \leq 2p - t. \tag{4.11}$$

On the other hand, the less restrictive condition (4.5) reduces to

$$C_p \leq p. \tag{4.12}$$

Condition (4.12) is consistent with the recommendations noted above. Other conditions on the F-ratio can easily be translated into conditions on C_p . For example, the condition

$F \leq 2$ which does not promise any improvement in the sense of the development of Section 4.3 but is a favorite of some users translates into $C_p \leq t + 1$.

In terms of relative gain, an estimate may be obtained by replacing the parameters in (4.7) by their full model estimates. In terms of C_p the result is

$$\widehat{\text{RGP}} = \frac{t + 1 - C_p}{n + t + 1}. \tag{4.13}$$

Note that subsets with $C_p > t + 1$ may give a small loss in precision of prediction relative to the full model. The use of the full model as a standard in the expression for relative gain is convenient, but the gain (or loss) for any subset should be compared with that achievable by using the subset with minimum C_p . The indication that the relative loss goes to zero with increasing sample size is misleading. It can be shown that for a given model, C_p increases with n.

If the objective is to estimate mean response, the expression for relative gain is given by deleting the n in the denominator of (4.13).

4.5. Other Criteria Functions.

Many statisticians voice a preference for the residual mean square, RMS_p , as a criterion function. Plots of RMS_p versus p are inspected and the choice of p is based on (i) the minimum RMS_p , (ii) the value of p such that $\text{RMS}_p = \text{RMS}$ for the full equation or (iii) the value of p such that the locus of smallest RMS_p turns sharply upward. While it is clear that there is a direct correspondence between the C_p and RMS_p-plots, it is of interest to investigate the source of these criteria and contrast them with those previously discussed for C_p . First, the choice of minimum RMS_p is apparently based on the fact that the true model minimizes $E(\text{RMS}_p)$. (See Theil [1961], Schmidt [1973a] or property 5 in Section 2.) In view of this, it would seem appropriate to use the criterion, minimum RMS_p , if the objective is extrapolation or estimation of parameters. Noting that

$$C_p = (n - p)\left(\frac{\text{RMS}_p}{\hat{\sigma}^2} - 1\right) + p, \tag{4.14}$$

the condition (4.11) may be written in terms of RMS_p . Thus, the subset with minimum RMS_p may be recommended for extrapolation or parameter estimation if, in addition, it satisfies

$$\frac{(n - t - 1)}{(n - p)}\text{RMS} \leq \text{RMS}_p \leq \frac{(n - t)}{(n - p)}\text{RMS}. \tag{4.15}$$

Recall that RMS is the residual mean square obtained when fitting the full model. The left inequality is, of course, guaranteed but is included to emphasize that this requirement is very restrictive.

The suggestions (ii) and (iii) are both based on the requirement that the ratio RMS_p/RMS is approximately equal to one. Reference to (4.14) shows that this is analagous to the requirement $C_p \doteq p$; however, the factor $(n - p)$ in (4.14) can magnify the difference between this ratio and one yielding large values of C_p . Thus, these suggestions would appear to be appropriate if the model is designed for prediction, keeping in mind the discussion related to (4.13).

The function, R^2, is probably the most commonly used criterion function. Typically, the plot of R_p^2 versus p may yield a locus of maximum R_p^2 which remains quite flat as p is decreased and then turns sharply downward. The value of p at which this "knee" in the

R_p^2 plot occurs is frequently used to indicate the number of terms in the model. It is of interest to ask if this criterion is appropriate for any of the intended uses and to relate it to the inspection of C_p-plots. It has been observed that R^2 is just a measure of the residual sum of squares to the total sum of squares and, hence, would appear to be a reasonable measure of data description.

The relation of R_p^2 to C_p is given by

$$C_p = (n - t - 1)(1 - R_p^2)/(1 - R^2) + 2p - n. \tag{4.16}$$

It is clear from this relation that, while the R_p^2 plot may be quite flat for a given range on p, the coefficient $(n - t - 1)$ can magnify small differences causing C_p to increase dramatically as p is decreased. As a result, the R^2-criterion may suggest the deletion of more variables than the minimum C_p-criterion. Simulation studies by Feiverson [1973] and Radhakrishnan [1974] indicate that essential variables may be deleted using the R^2-criterion. Also, lacking a precise definition of the knee, the qualitative inspection of R^2-plots is dependent on the scale. It would appear that the C_p-plots are more amenable to a graphical analysis.

As an alternative to R^2, some users recommend the adjusted squared multiple correlation coefficient, $\bar{R}^2$, and suggest using the value of p for which $\bar{R}_p^2$ is maximum. This procedure is exactly equivalent to looking for the minimum RMS_p. Indeed, there appears to be no advantage in using $\bar{R}_p^2$ over RMS_p in view of the simple relation

$$\bar{R}_p^2 = 1 - \frac{n-1}{\text{Total}}\text{RMS}_p . \tag{4.17}$$

The two remaining functions which are simply related to RMS_p, namely J_p and S_p, are less frequently used. J_p arises by computing the total prediction variance over the current data for a given subset and then estimating σ^2 by RMS_p. On theoretical grounds, the objection to this statistic is that it ignores the bias in prediction.

The function S_p has an appeal similar to C_p in the sense that it arises by considering the average mean squared error of prediction. In this case, the average is taken over all x, assuming y and x are multivariate normal. Specifically, the expected value of (2.16) is

$$E(\text{MSEP}\ (\tilde{y})) = \frac{n+1}{n}\frac{n-2}{n-p-1}(\sigma^2 + \beta_r'\Sigma_{rr.p}\beta_r) \tag{4.18}$$

where σ^2 is the residual variance of y conditioned on x and $\sum_{rr.p}$ is the residual covariance matrix for the deleted variables conditioned on the retained variables. Following Mallows development of C_p note from (2.12) that

$$E[\text{RMS}_p] = \sigma^2 + \beta_r'\Sigma_{rr.p}\beta_r . \tag{4.19}$$

Substituting into (4.18) yields

$$E[\text{MSEP}\ (\tilde{y})] = \frac{(n+1)(n-1)}{n}\cdot\frac{E(\text{RMS}_p)}{n-p-1}. \tag{4.20}$$

The statistic S_p then arises by deleting the factor involving only n and replacing $E(\text{RMS})$ by RMS_p. (Tukey [1967] gives a different argument for the same statistic.)

If the assumption of multivariate normality is acceptable, then this development suggests looking at subsets with values of S_p close to minimum S_p if the objective is to use the resulting equation for prediction. Further, since the average is taken over all x, this equation may be appropriate for modest extrapolation.

The expression for relative gain when (4.20) is used in (4.6) is

$$\text{RGP} = 1 - \frac{n - t - 2}{n - p - 1} \frac{E(\text{RMS}_p)}{\sigma^2}. \tag{4.21}$$

The obvious estimate of relative gain suggests consideration of subsets for which S_p is not appreciably greater than $s^2/(n - t - 2)$. Again, the gain (or loss) for any subset should be contrasted with that attained using the minimum S_p .

Lindley [1968] developed a Bayesian approach to the problems of subset selection for prediction and control, including in his analysis the cost of observing the input variables. With regard to prediction from what Lindley terms a random experiment, the criterion is to choose a p-term equation so as to minimize the quantity,

$$\hat{\beta}_r' B_{rr}^{-1} \hat{\beta}_r / n + U_p . \tag{4.22}$$

Here r and p are as previously defined and U_p is the cost of observing the p-terms selected. Note that this criterion is generally consistent with those discussed previously which require that the first term of (4.22) be small. If U_p is independent of p then this criterion will suggest that all terms be included. This is contrary to the previous discussions which allow for decreased mean squared error of prediction for a subset equation. Lindley specifically mentions the case of polynomial regression in which case a polynomial of either degree $n - 1$ or zero would be used.

If control of the output at level y_0 is the objective, the appropriate criterion is given by adding to (4.22) the terms

$$\sigma^2[r/n + y_0^2/(\sigma^2 + \hat{\beta}_p'(X_p'X_p)\hat{\beta}_p)]. \tag{4.23}$$

The difference in the two criteria is primarily caused by the fact that the standard errors of the regression coefficients were found to be irrelevant in the prediction problem.

4.6. Stopping Rules for Stepwise Methods.

The subsets encountered by any of the stepwise procedures can be evaluated according to any of the criteria used with the all-possible or optimal procedures. The sequential nature of the computations is such that if F_{in} is sufficiently large for FS, the computations will be terminated before all variables are included. Conversely, if F_{out} is sufficiently small for BE, not all variables will be eliminated. As a consequence, schemes for choosing F_{in} and F_{out} are referred to as "stopping rules." For example, one might consider

$$F_{\text{in}} = F(\alpha, 1, n - p - 1) \tag{4.24}$$

and

$$F_{\text{out}} = F(\alpha, 1, n - p). \tag{4.25}$$

Pope and Webster [1972] describe conditions under which the level of significance is meaningful for FS, noting that these conditions are rarely satisfied in practice. Kennedy and Bancroft [1971] developed expressions for bias and mean squared error of prediction for FS and BE but under restrictive conditions.

Some users prefer to choose F_{in} and F_{out} so that the computations will run the full course and reveal one subset of each size for FS and BE. The subsets are then evaluated by criteria of the type described in Section 4.2. The determination of the final subset size is generally referred to as a stopping rule.

Bendel and Afifi [1974], in a simulation study, compared eight different stopping rules

for FS. Their study includes C_p and S_p (their U_m if the coefficient $(n^2 - n - 2)/n$ is deleted), the univariate $-F$ procedure using F_{in} as in (4.24), the lack-of-fit-F

$$L_f = (\text{RSS}_p - \text{RSS})/(\hat{\sigma}^2(t + 1 - p)), \tag{4.26}$$

the adjusted R^2 and some variations. The measure of comparison is the value of $\bar{p}_e$, the mean normalized prediction error. Their results indicate a preference for the univariate $-F$ or a test of $S_p = S_{p+1}$ at level α for small degrees of freedom (d.f.) with $0.1 \leq \alpha \leq 0.4$. For higher d.f. (40 or more) C_p and S_p rank high but the univariate $-F$ with $\alpha = 0.2$ or 0.25 does quite well. They suggest that the best overall test is the univariate $-F$ with $\alpha = 0.15$. These results on the significance level are consistent with those reported by Kennedy and Bancroft [1971] who recommended $\alpha = 0.25$ for FS and $\alpha = 0.10$ for BE in the univariate $-F$ procedures.

It should be emphasized that the Bendel and Afifi [1974] results are based on an evaluation of the subsets revealed by the FS algorithm. Substantial improvement in the value of $\bar{p}_e$ might be noted if all possible subsets, or at least the optimal subset, were evaluated.

The use of stopping rules based on sequentially adding terms in polynomial regression was discussed by Beaton and Tukey [1974]. They warn against the practice of stopping when the term of next highest degree gives no improvement over the current equation.

4.7. Validation and Assessment.

Having decided upon a particular subset and estimated the coefficients, it is natural to attempt to assess the performance of the equation in terms of its intended use. If, for example, prediction is the objective then one might consider gathering additional data, hopefully under comparable conditions, and comparing observed and predicted values. Letting y_{0i} and $\tilde{y}_{pi}$, $i = 1 \cdots m$ denote the new observed values and their predicted values using the p-term equation, a reasonable measure of performance is the prediction mean squared error

$$\text{PMS}_p = \sum_{i=1}^{m} (y_{0i} - \tilde{y}_{pi})^2/m. \tag{4.27}$$

This quantity might then be compared with RMS_p noting that the mean of RMS_p is given by (2.12) while the mean of PMS_p is given by

$$E[\text{PMS}_p] = \sigma^2 + \{\text{tr}\ (X_{0_p}'X_{0_p})(X_p'X_p)^{-1}\sigma^2 + ((X_{0_p}A - X_{0_r})\beta_r)^2\}/m. \tag{4.28}$$

Here the input matrix leading to the new responses is $X_0 = (X_{0_p}\ ,\ X_{0_r})$ and A is defined by (2.8).

Frequently it is not convenient to collect additional data for assessment; hence, some users suggest splitting the available data into two groups: one for analysis (e.g. choice of subset equation and estimation of coefficients) and the other for assessment. The paper by Anderson *et al.* [1970] illustrates the performance of a particular set of data with various partitions, using the full model as well as the models developed by BE and FS (as modified by Efroymson [1960]. As suggested by comparing (2.12) and (4.28), they observe that PMS_p is larger than RMS_p . The more interesting point they make is that PMS_p is substantially larger when using the full model than for the subset models (with FS being slightly superior to BE). This is in agreement with the results obtained in Section 2.

The basic idea behind partitioning the data is that the data used for analysis should not be used for assessment. The question of how to partition the data and the decision as to whether this luxury can be afforded has led several authors to consider integrating the two

concepts of analysis and assessment. If all but the i^{th} observation is used to obtain a p-term predictor of y_i , say $\hat{y}_p(i)$, then an assessment function proposed by Allen [1971b], Schmidt [1973a] and Stone [1974] is the sum of squares of differences between the observed and predicted value. Allen has labelled this function PRESS_p . That is,

$$\text{PRESS}_p = \sum_{i=1}^{n} (y_i - \hat{y}_p(i))^2. \tag{4.29}$$

Letting e_p denote the vector of residuals for the p-term equation and D_p denote the diagonal matrix whose diagonal elements are those of $I - X_p(X_p'X_p)^{-1}X_p'$, it can be shown that

$$\text{PRESS}_p = e_p'D_p^{-2}e_p . \tag{4.30}$$

As a criterion for determining a subset regression, the suggestion is to evaluate PRESS_p for all possible subsets and make a selection based on small values of PRESS_p . Allen [1971b] suggested a computational scheme for this purpose and also suggested a stepwise procedure for the case when t is large. An algorithm analagous to SELECT which would guarantee minimum PRESS_p has not been developed.

A criterion function closely related to PRESS_p, proposed by Schmidt [1973a], is the standardized residual sum of squares,

$$\text{RSS}_p^* = e_p'D_p^{-1}e_p . \tag{4.31}$$

He observed that the true model minimizes $E(\text{RSS}_p^*)$.

Inspection of (4.30) and (4.31) shows that they are both weighted sums of squares of the residuals and, hence, direct comparison of these two functions with those previously discussed is difficult. It would appear, however, that for large samples both (4.30) and (4.31) would be close to RSS_p . This follows since in this case one would expect that $\hat{y}_p(i)$ would be approximately equal to the estimator $\hat{y}_{p_i}$ obtained by using all n observations.

In summary, PRESS_p has an intuitive appeal if the objective is prediction. A detailed description of how to use plots of PRESS_p such as that given for C_p would be desirable. The increased computational demands of PRESS_p may outweigh any advantages these functions have over those depending only on RSS_p .

4.8. Ridge Regression as a Selection Criterion.

As noted earlier, ridge regression can be viewed as a variable selection procedure if variables for which $\tilde{\beta}_i(k)$ is small are deleted. The problem of choosing k is fundamental and deferred to Section 5.

Allen [1972], [1974] considered using a generalized ridge estimator for the purpose of selecting variables. Specifically, he considered the estimator

$$\beta(K) = (X'X + K)^{-1}X'Y \tag{4.32}$$

where K is a diagonal matrix whose components are to be determined. (Here X is in adjusted form as opposed to the standardized form in (3.4).) The suggestion is to express a criterion function as a function of K and then determine K to minimize (if that is appropriate) this function. For example, he considers the function

$$\text{PRESS}\ (K) = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 \tag{4.33}$$

where $\hat{y}_{(i)}$ is the estimator of $E[y_i]$ using (4.32) and excluding the i^{th} observation. In Allen

[1972], a computational procedure utilizing nonlinear regression techniques was suggested for minimizing PRESS (K), or other criterion functions, with respect to K.

The appealing concept here is that Allen considers a generalization of the ridge estimator suggested by Hoerl and Kennard [1970a] and in addition chooses the diagonal elements based on a specific criterion. With respect to subset selection, Allen suggested that a large diagonal element in K (e.g. greater than 10^{15}) indicates that the corresponding variable may be deleted. This procedure removes the subjectivity from determining the amount of ridge "shrinkage" and may well yield a good subset estimator. It has the undesirable feature of focusing on a single subset which is contrary to the ideas previously discussed.

4.9. Examples.

4.9.1. Variable Analysis for Gas Mileage Data. The analysis of the role of the automobile characteristics in describing gas mileage is aided by the graphical display in Figures 1 and 2. Referring to the C_p-plot in Fig. 1, observe that for $4 \leq p \leq 9$, the ten best subsets all satisfy $C_p \leq p$. Further, for $6 \leq p \leq 11$ the best subset and a number of other subsets satisfy $C_p \leq 2p - t$. For $p = 3$, only three subsets satisfy $C_p \leq p$. In view of the discussion in Section 4.4, there are a number of candidate equations for prediction ($C_p \leq p$) and estimation and extrapolation ($C_p \leq 2p - t$). Inspection of the variables contained in these subsets reveals that variables 3, 9 and 10, which constitute the best subset for $p = 4$ with

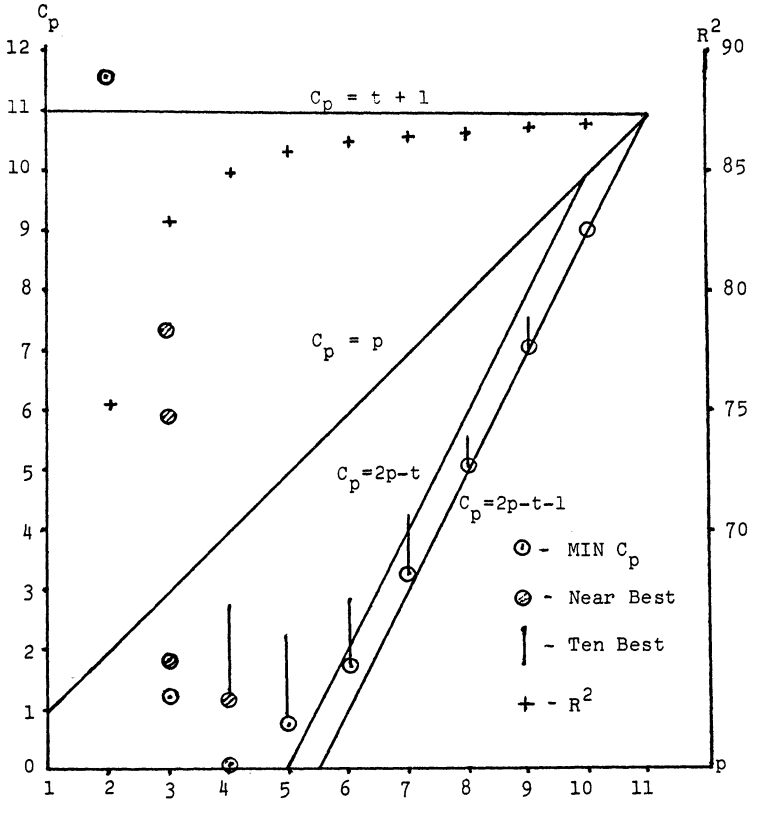

FIGURE 1

C_p AND R^2 ($\times 10^2$) PLOTS FOR GAS MILEAGE DATA

$C_p = 0.103$, are contained in all of the best subsets and virtually all of the other subsets identified for $p > 4$. Variables 9 and 10 give the third best subset for $p = 3$ but the pairs (3, 10) and (3, 9) are not among the top ten. It is apparent that the combined effect of these three variables is essential, a fact not obvious from their individual or pairwise performances.

By contrast, variable 2, which is the second best single variable and along with variable 9 constitutes the best pair of variables, rarely occurs in any of the subsets identified. The subset (2, 6, 9) which is identified by FS is the second best for $p = 4$ with $C_p = 1.15$. This subset offers a relative gain in precision of prediction of 22.9 percent as computed from (4.13). The subset (3, 9, 10) yields a 25.3 percent relative gain while the pair (2, 9) and the best four (3, 6, 9, 10) yield relative gains of 22.7 percent and 23.7 percent, respectively.

The selection of a single subset for almost any use should apparently contain variables 3, 9 and 10. The minimum C_p criterion recommends this particular set for prediction. The minimum S_p criterion (Fig. 2) is in agreement but the minimum is not as well defined, suggesting $p = 5$ or even 6. For estimation of parameters and extrapolation, the criterion $C_p \leq 2p - t$ (Fig. 1) suggests $p \geq 6$ while the minimum RMS_p criterion (Fig. 2) yields $p = 6$. The R^2-plot is less definite suggesting $p \leq 5$ with a dramatic drop at $p = 3$.

The analysis of the ridge trace (see Section 5.5.1) does not suggest the important role of variables 3, 9 and 10 but does show a significant decrease in magnitude of the coefficients for variables 6, 9 and 10 in the range $0 \leq k \leq 0.20$. Variable 5 also indicates "instability"

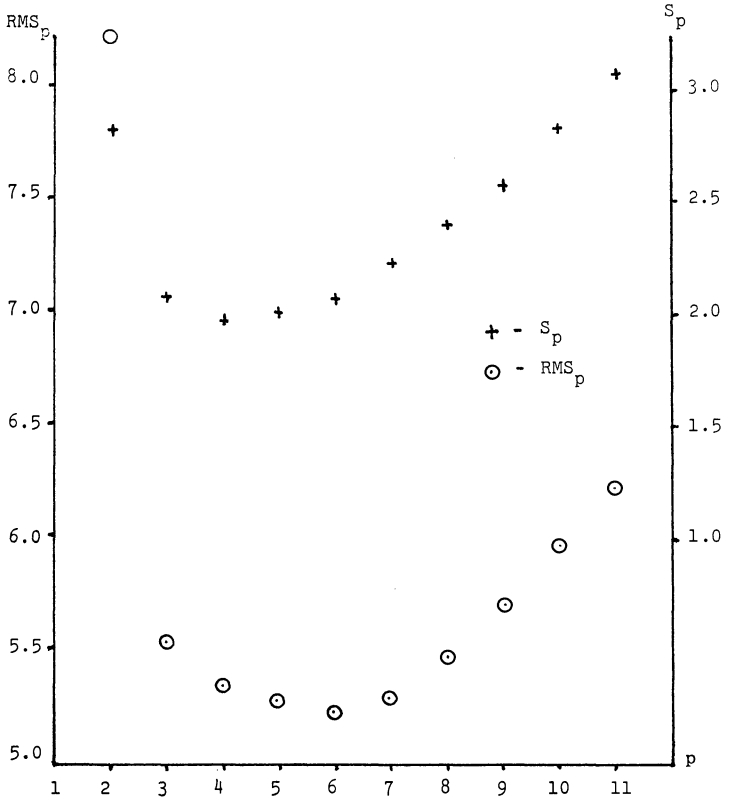

FIGURE 2

RMS_p ($\times 10^3$) AND S_p ($\times 10^4$) PLOTS FOR GAS MILEAGE DATA

by a change in the sign of its coefficient while variables 2 and 3 are relatively stable. The fact that the ridge trace judges as unreliable two of the three "essential" variables identified in the subset analysis is disturbing and warrants further investigation. This and a discussion of the choice of k are given in Section 5.5.1.

For comparison and later reference, Table 5 lists the coefficient estimates for the best subsets for $p = 4$, 5 and 6.

4.9.2. Variable Analysis for Air Pollution Data. Inspection of the C_p-plot, Fig. 3, shows that for $6 \leq p \leq 16$ the best subset and a number of other candidates satisfy the criterion $C_p \leq p$ and for $11 \leq p \leq 16$ the more demanding criterion $C_p \leq 2p - t$ is satisfied by the best subsets. Plots of RMS_p and S_p are not shown but they take on their minima at $p = 8$ and $p = 7$, respectively, the latter agreeing with minimum C_p. Again the R^2-plot, shown in Fig. 3, is less distinctive, dropping more rapidly for $p < 5$.

McDonald and Schwing [1973] suggested the subset (1, 2, 3, 6, 9, 14) which yields minimum C_p. In view of the apparent purpose of the study, i.e. model building, the inclusion of one or two more variables might be appropriate, namely, variables 5 and 4 in that order.

McDonald and Schwing also considered the ridge regression criteria of Hoerl and Kennard [1970b] for eliminating variables. The subset (1, 2, 8, 9, 14) was suggested. Variables 3, 5, 12 and 13 were eliminated because of instability and the remainder for having small coefficients. McDonald and Schwing suggested a least squares fit for these six variables which is in contrast to the preference of Hoerl and Kennard for not reestimating but, instead, using the estimates for the full model. The subset suggested by ridge is observed to be the third best subset for $p = 7$ with $C_p = 5.5$.

Table 6 lists the coefficient estimates for the best subset for $p = 7$, 8, 9 as well as the ridge subset. These can be contrasted with the biased estimators discussed in Section 5.5.2.

5. BIASED ESTIMATION

In the preceding sections, consideration has been given to the problems of identifying relevant variables and examining their interrelations. With the exception of ridge regression, the statistics used were based on fitting the equations by least squares. The implication is that once we decide on the variables to be included, the least squares fit should be used.

The current literature contains a number of alternatives to least squares which, although they produce biased estimates, may be preferable. Depending upon the intended use of the regression equation, the objection to bias may not be strong; indeed, subset estimators are

TABLE 5

SUBSET LEAST SQUARES ESTIMATES FOR GAS MILEAGE DATA ($\times 10$)

	VARIABLE				
p	3	5	6	9	10
4	2.43	--	--	-6.36	3.63
5	2.42	--	-2.01	-5.26	2.40
6	2.87	0.231	-2.41	-6.63	2.99

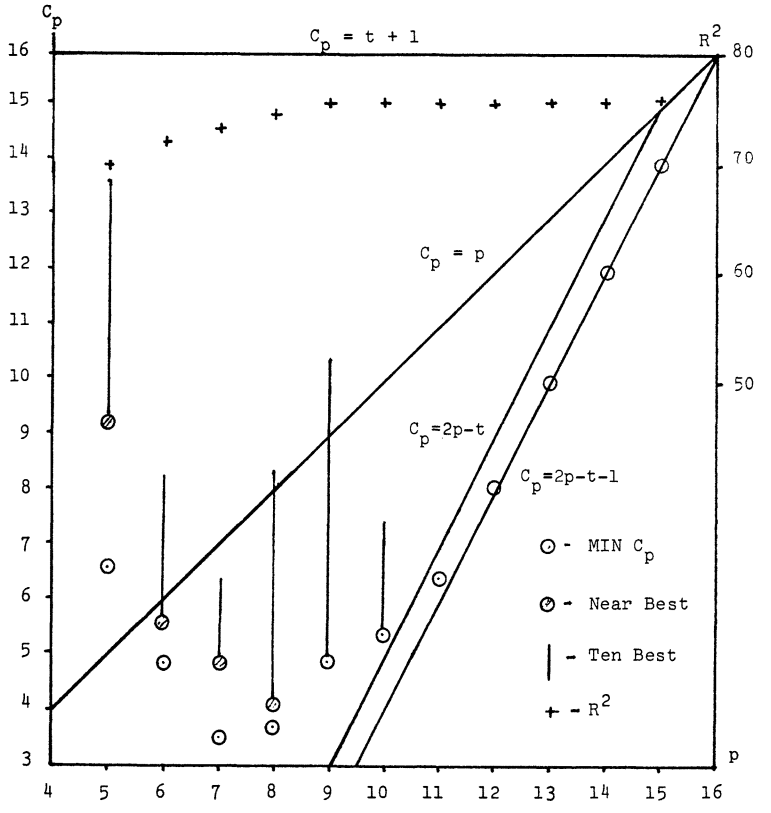

FIGURE 3

C_p AND R^2 ($\times 10^2$) PLOTS FOR AIR POLLUTION DATA

generally biased. The important issue would appear to be whether or not the resulting estimators or predictors do a better job.

In this section three biased estimation procedures, namely, Stein Shrinkage, ridge regression and some variations of principal component regression are reviewed and related, at least geometrically. The estimation problem is discussed without reference to subset

TABLE 6

SUBSET LEAST SQUARES ESTIMATES FOR AIR POLLUTION DATA ($\times 10$)

	VARIABLE								
p	1	2	3	4	5	6	8	9	14
7	2.88	-2.42	-1.80	--	--	-1.85	--	6.57	2.65
8	2.65	-3.09	-1.76	--	-1.35	-2.31	--	7.48	2.30
9	3.31	-3.51	-2.17	-1.55	-2.21	-2.70	--	6.92	2.31
Ridge p = 7	2.39	-2.67	--	--	--	-1.57	0.97	5.94	2.49

selection. The problem of integrating biased estimation with subset selection is considered in Section 6.

To simplify the presentation in this section, assume that all variables, dependent as well as independent, have been standardized. Thus the components of $X'X$ and $X'Y$ are sample correlations. The symbol $\hat{\beta}$ will always refer to the least squares estimator while other estimates will be distinguished by appropriate subscripts or superscripts.

5.1. Stein Shrinkage.

If it is assumed that the dependent and independent variables are multivariate normal with known means, Stein [1960] showed that the maximum likelihood estimator of β is inadmissible for $t \geq 3$, $n \geq t + 2$.

Specifically, Stein considers the loss function

$$L = \frac{(\tilde{\beta} - \beta)'\Gamma(\tilde{\beta} - \beta)}{\sigma^2} \tag{5.1}$$

where Γ is the covariance matrix of the t-vector of input variables, σ^2 is the residual variance in the conditional distribution of y on x and $\tilde{\beta}$ is any estimator of β. If we consider $\tilde{\beta} = \hat{\beta}$, the least squares or maximum likelihood estimator, then the risk (expected loss) is the constant $t/(n - t - 1)$. Stein [1960] showed that the estimator

$$\tilde{\beta} = \left(1 - \frac{b(1 - R^2)}{a(1 - R^2) + R^2}\right)\hat{\beta}, \tag{5.2}$$

for appropriate choices of a and b, has uniformly smaller risk than $\hat{\beta}$. (Here R^2 is the sample multiple correlation coefficient as defined in Section 4.) In particular, he noted that the risk for $a = 0$ is less than maximum likelihood if $b = (t - 2)/(n - t - 2)$ and suggested the estimator $\tilde{\beta}_S = c\hat{\beta}$, where

$$c = \max\left[\left(1 - \frac{t - 2}{n - t + 2}\frac{1 - R^2}{R^2}\right), 0\right]. \tag{5.3}$$

Geometrically, Stein's recommendation is to shrink the least squares estimator toward the origin. Stated differently, $\tilde{\beta}_s$ is the solution to the problem

$$\begin{aligned} &\underset{\beta}{\text{minimize}}\ (\beta - \hat{\beta})'(\beta - \hat{\beta}) \\ &\text{subject to: } \beta'\beta \leq d^2, \end{aligned} \tag{5.4}$$

the radius d being determined from the constant c.

Sclove [1968] suggests that it might be appropriate to apply the shrinkage factor to a subset of the variables. Since the shrinkage is applied to the coefficients in the space of orthogonal regressions, the discussion of this idea is deferred to Section 5.3.

5.2. Ridge Regression.

The concept of ridge regression, as described by Hoerl and Kennard [1970a], [1970b], was motivated by the fact that models for which the correlation matrix has a non-uniform eigenvalue structure can lead to least squares estimates which are far removed from the true parameter point. To see this, let L_1 be the Euclidean distance from the least squares point, $\hat{\beta}$, to the true parameter point β. Then if λ_i , $i = 1 \cdots t$ are the eigenvalues of $X'X$,

$$E[L_1^2] = \sigma^2 \sum_{i=1}^{t} 1/\lambda_i \tag{5.5}$$

and

$$\text{VAR}\,[L_1^2] = 2\sigma^4 \sum_{i=1}^{t} 1/\lambda_i^2. \tag{5.6}$$

If one or more of the λ_i are very small it is clear that $\hat{\beta}$, although unbiased, may be far removed from β.

The ridge estimator is similar to the Stein estimator in the sense that it shrinks the least squares estimator toward the origin, but in this case the shrinkage is done with respect to the contours of $X'X$. Specifically, the ridge estimator proposed by Hoerl and Kennard [1970a],

$$\tilde{\beta}_R = (I + k(X'X)^{-1})^{-1}\hat{\beta} \tag{5.7}$$

is the solution to the problem

$$\begin{aligned} &\underset{\beta}{\text{minimize}}\ (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \\ &\text{subject to: } \beta'\beta \leq d^2 \end{aligned} \tag{5.8}$$

where the radius d depends on k. The relation between $\tilde{\beta}_S$ and $\tilde{\beta}_R$ as a function of the radius of the constraining sphere is shown for $t = 2$ in Fig. 4. (Note that Stein does not recommend shrinkage for $t = 2$.)

The ridge regression concept has generated considerable interest in the literature. Marquardt [1970] noted the relation between ridge estimators and a generalized inverse estimator and Marquardt [1974b] noted the relation to robust regression. Mayer and Wilke [1973] considered a general class of estimators based on linear transforms of least squares estimators which included ridge and shrunken estimators as special cases. Marquardt and Snee [1973] and McDonald and Schwing [1973] provided applications to real data sets. Newhouse and Oman [1971] and McDonald and Galarneau [1975] presented the results of simulation studies, the former being quite critical of ridge estimators. Coniffe and Stone [1973] examined the concept of ridge regression and were generally critical.

Much of the discussion of ridge regression centers around the choice of the constant k in (5.7). Hoerl and Kennard [1970a] established the existence of a constant k which yields an estimator with smaller average distance from β than the least squares estimator. They recommended inspection of the "ridge trace" as a means of estimating k. It is of interest to ask if the optimality properties they cite still apply when k is estimated from the data.

Other authors have recommended alternative schemes for estimating k. Marquardt [1970] and [1974a] and Marquardt and Snee [1973] suggested using the value of k for which the maximum variance inflation factor (VIF) is "between one and ten and closer to one." The VIF associated with each coefficient represents the amount by which the variance of that coefficient is inflated by the correlations between the variables. Specifically, the VIF's are the diagonal elements of

$$\text{VAR}\,(\tilde{\beta}_R)/\sigma^2 = (X'X + kI)^{-1}X'X(X'X + kI)^{-1}. \tag{5.9}$$

Mallows [1973a] extended the concept of C_p-plots to C_k-plots which may be used to determine k. Specifically, he suggested plotting C_k versus V_k where

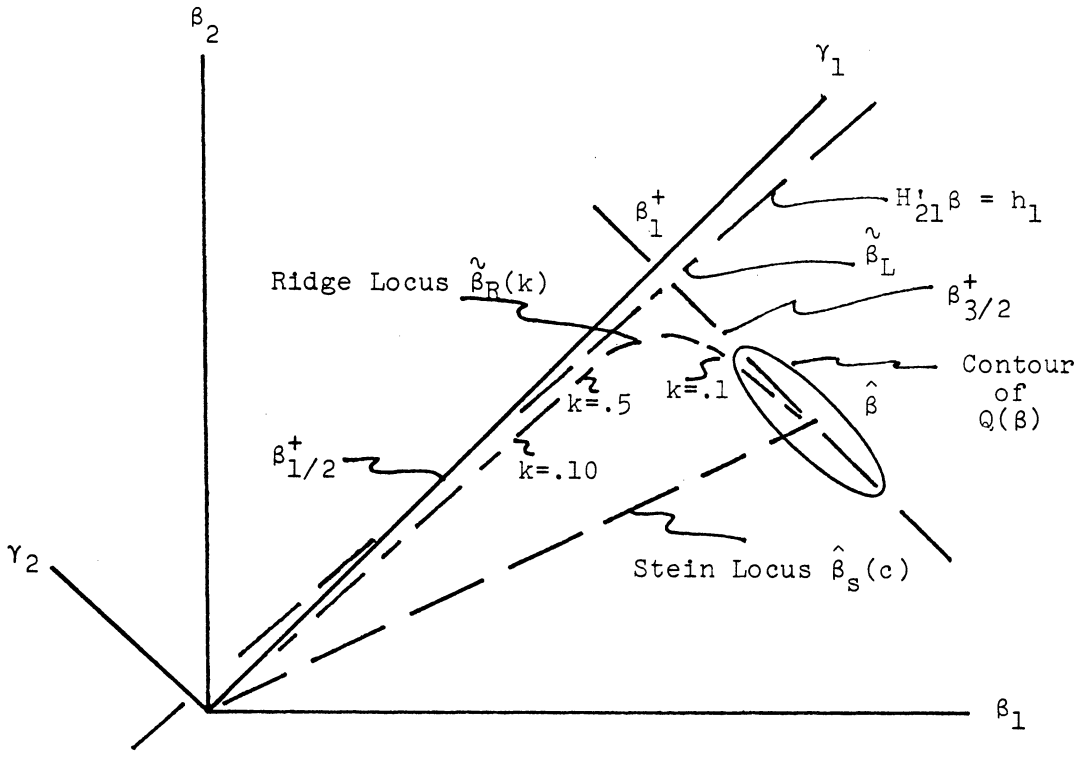

FIGURE 4
BIASED ESTIMATORS FOR $t = 2$, $p = .9$

$$C_k = \frac{\text{RSS}_k}{\hat{\sigma}^2} - n + 2 + 2 \text{ tr } (XL), \tag{5.10}$$

$$V_k = 1 + \text{tr } (X'XLL') \tag{5.11}$$

and

$$L = (X'X + kI)^{-1}X'. \tag{5.12}$$

Here RSS_k is the residual sum of squares as a function of k. The suggestion is to choose k to minimize C_k. It is of interest to note that for the data discussed by Gorman and Toman [1966], the ridge trace inspection indicates that k be in the interval $0.20 \leq k \leq 0.30$ whereas Mallows' procedure yields $k \doteq 0.02$. The VIF criterion yields $0.02 \leq k \leq 0.10$.

Farebrother [1975] suggested $k = \hat{\sigma}^2/\hat{\beta}'\hat{\beta}$ which, for the Gorman-Toman data, yields $k = 0.003$. With respect to this formula, it is of interest to note that in the case $X'X = I$, the choice of k which will minimize $E(L_1^2(k))$ is $k = t\sigma^2/\beta'\beta$. Hoerl *et al.* [1975] recommended $k = t\hat{\sigma}^2/\hat{\beta}'\hat{\beta}$ as a general rule where the parameters are estimated from the full equation least squares fit. Their simulation studies suggest that the resulting ridge estimator yields coefficient estimates with generally smaller mean squared error than that obtained from least squares. In a later paper, Hoerl and Kennard [1975] suggest an iterative procedure where $k_{i+1} = t\hat{\sigma}^2/\beta_i'\beta_i$ with $\beta_i = \tilde{\beta}_R(k_i)$.

Newhouse and Oman [1971] conducted a simulation study of ridge estimators. Their

study was restricted to the case of two predictors for two different values of r, the correlation between predictors and a number of schemes for choosing k. Their conclusions indicated, at least for the case $t = 2$, that ridge estimators may well be worse than ordinary least squares and, in general, fail to establish any superiority. They further stated that there is nothing to suggest that the results in higher dimensions ($t > 2$) would be substantially different.

McDonald and Galarneau [1975] performed a similar study for the case $t = 3$ using two new methods of estimating k. The basic idea is to choose k so that

$$\tilde{\beta}_R'\tilde{\beta}_R = \hat{\beta}'\hat{\beta} - \hat{\sigma}^2 \sum_{i=1}^{t} 1/\lambda_i \ . \tag{5.13}$$

If the right side of (5.13) is negative, they suggest two modifications. Although neither method was better than least squares in all cases, they concluded, based on an optimal rule for choosing k, that there is sufficient potential improvement to warrant further investigation of ridge estimators. They also considered the case $t = 2$ and found that their results were comparable to those of Newhouse and Oman [1971]. They suggest that there may be some advantage to ridge estimators in higher dimensions which is not available for $t = 2$. This is consistent with the results of Stein [1960]. They also report that the values of k chosen by their optimal rule, as well as the practical rules, were generally less than those obtained from the ridge trace. For example, their analysis of the Gorman and Toman [1966] data suggested $k = 0.007$.

A natural extension of the ridge estimator is to consider a general diagonal matrix K rather than the scalar matrix kI. Newhouse and Oman [1971] included such an extension in their simulation study. Hoerl and Kennard [1970a] also considered this but in the space of orthogonalized predictors. In this case they obtained explicit expressions for the elements of K and recommended an iterative procedure for estimating them. Hemmerle [1974] obtained a closed form solution for the estimates, eliminating the need for the iterative procedure. These extensions add to the flexibility of the ridge estimator but the visual features of the single parameter ridge trace, favored by Hoerl and Kennard [1970a], are lost.

As a final comment on ridge regression, it can be shown (see e.g. Marquardt [1970]) that the ridge estimator is equivalent to a least squares estimator in which the data has been augmented by a fictitious set of points such that the response is zero and a diagonal matrix is added to $X'X$. The possibility of actually collecting additional data which would improve the stability of $X'X$ is thus suggested. The paper by Gaylor and Merrill [1968] describes methods for achieving this.

5.3. Principal Component Regression.

Mason *et al.* [1975] cite a number of sources of multicollinearity in regression variables. The occurrence of small eigenvalues of $X'X$ is a warning of the presence of one or more of these problems. It is clear that if there are s zero eigenvalues, the number of input variables can be reduced by s. If the small eigenvalues are near zero, the situation is not so clear. They may represent actual linear dependencies and departures from zero may be due to measurement and computational inaccuracies. On the other hand they may be nonzero but represent "near" dependencies. The procedure in this case is not clear. Ridge regression ignores the nature of the dependency and "distorts" the data to yield a set of biased estimators. Other authors, for example, Kendall [1957], Massy [1965], Jeffers [1967], Lott [1973] Hawkins [1973] and Greenberg [1975], recommend transforming to the space of orthogonal predictors determined by the eigenvectors and deleting the variables corresponding to the

small eigenvalues. It is of interest to contrast this procedure with those described earlier in this section and to mention some extensions.

Let Λ denote the diagonal matrix of eigenvalues, λ_i, of $X'X$ and T denote the orthogonal matrix of eigenvectors. That is,

$$T'X'XT = \Lambda \tag{5.14}$$

and

$$T'T = I. \tag{5.15}$$

Letting $Z = XT$ and $\beta = T\gamma$, the linear model (2.2) becomes

$$Y = Z\gamma + e \tag{5.16}$$

with least squares estimate determined by the solution of

$$\Lambda\hat{\gamma} = Z'Y. \tag{5.17}$$

Or, in terms of the original parameters,

$$\hat{\beta} = T\hat{\gamma}. \tag{5.18}$$

Now if s of the eigenvalues are zero, it follows that the corresponding columns of Z are zero and these variables drop out of the orthogonal model (5.16). The equation (5.17) is then of dimension $g = t - s$. That is if we write in partitioned form,

$$T = (T_g , T_s),$$

$$\gamma' = (\gamma_g', \gamma_s')$$

and

$$\Lambda = \text{Diag } (\Lambda_g , \Lambda_s), \tag{5.19}$$

then

$$\Lambda_g\hat{\gamma}_g = T_g'X'Y. \tag{5.20}$$

In terms of the original parameters,

$$\beta_g^{+} = T_g\hat{\gamma}_g . \tag{5.21}$$

The technique which has become known as "principal component" regression is based on this analysis. That is, if X is actually of rank t, but s of the eigenvalues are judged to be "sufficiently small," they are set to zero and β is estimated by (5.21).

It is of interest to compare this procedure with the shrunken estimators and the ridge estimators as characterized by (5.4) and (5.8), respectively. In particular, we observe that (5.21) is the solution to the problem

$$\underset{\beta}{\text{minimize}}\ (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})$$

$$\text{subject to: } T_s'\beta = 0. \tag{5.22}$$

An alternative expression for (5.21) is

$$\beta_g^{+} = (I - T_sT_s')\hat{\beta} \tag{5.23}$$

which is reminiscent of (5.7) but distinctly different.

To illustrate for the case $t = 2$ and $g = 1$, note that

$$X'X = \begin{bmatrix} 1 & p \\ p & 1 \end{bmatrix}, \tag{5.24}$$

the eigenvalues are $\lambda_1 = 1 + p$ and $\lambda_2 = 1 - p$ and the eigenvectors are given by the matrix

$$T = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \tag{5.25}$$

Thus, for example, if p is close to one and, hence, λ_2 is close to zero, the linear constraint in (5.22) is $\beta_1 = \beta_2$. The estimator $\beta_g{}^+$ is indicated in Fig. 4.

Marquardt [1970] suggested that the assumption of an integral number of zero roots of $X'X$ may be too restrictive. He noted that if $X'X$ is of rank g, then (5.21) is the (Moore-Penrose) generalized inverse solution of the normal equations. That is, the generalized inverse of $X'X$ is given by

$$(X'X)^+ = T_g \Lambda_g{}^{-1} T_g'. \tag{5.26}$$

In the case where X is actually of rank t but has small eigenvalues, he suggested the concept of fractional rank of X. Thus if we assume X is of rank f, where $g < f < g + 1$, then the recommendation is to use

$$(X'X)^+ = T_g \Lambda_g{}^{-1} T_g' + \frac{f - g}{\lambda_{g+1}} t_{g+1} t_{g+1}'. \tag{5.27}$$

The "generalized inverse" solution recommended by Marquardt is then given by

$$\beta_f{}^+ = (X'X)^+ X'Y \tag{5.28}$$

where $(X'X)^+$ is given by (5.27). An alternative expression which reveals the relation of this solution to the principal component solution is

$$\beta_f{}^+ = (1 - f + g)\beta_g{}^+ + (f - g)\beta_{g+1}{}^+. \tag{5.29}$$

Here $\beta_g{}^+$ and $\beta_{g+1}{}^+$ denote, respectively, the principal component solution (5.21) in which we set either s or $s - 1$ of the eigenvalues to zero.

The advantage of fractional rank is apparent. The decision as to whether to assume λ_{g+1} equal to zero is avoided by choosing a solution along the line joining the two solutions obtained by either setting λ_{g+1} to zero or leaving it alone. The fractional rank solutions for $f = 1/2$ and $f = 3/2$ are shown in Fig. 4. Hocking *et al.* [1975] discuss the determination of f and g.

To relate to (5.22) note that the generalized inverse solution solves the problem

$$\begin{aligned} &\text{minimize } (\beta - \hat{\beta})X'X(\beta - \hat{\beta}) \\ &\text{subject to: } T_s'\beta = \delta. \end{aligned} \tag{5.30}$$

Here δ is a vector of zeros with the exception of the component corresponding to the row t_{g+1}' in T_s'. This component of δ is

$$\frac{f - g}{\lambda_{g+1}} t_{g+1}' X'Y. \tag{5.31}$$

Sclove [1968] suggested the application of Stein shrinkage to a subset of the coefficients.

For example, one might choose to shrink the coefficients corresponding to the s variables with smallest eigenvalues. In the present notation, this estimator is given by

$$\tilde{\beta}_{ss} = T\begin{bmatrix} I_g & 0 \\ 0 & cI_s \end{bmatrix}T'\hat{\beta} \tag{5.32}$$

where

$$c = \max\left[\left(1 - c^* \frac{1 - R^2}{w}\right), 0\right]. \tag{5.33}$$

Here,

$$w = \hat{\beta}'T_s\Lambda_s T_s'\hat{\beta} \quad \text{and} \quad 0 \leq c^* \leq \frac{2(s-2)}{n-t+2}.$$

For example, the value of $c_0^* = (s-2)/(n-t+2)$ is analagous to the recommendation of Stein [1960]. Actually, Sclove [1968] suggested a preliminary test which might suggest setting $c = 0$. It is of interest to note that this corresponds to the "principal component" estimator β_g^+ given by (5.21). In fact, $\hat{\beta}_{ss}$ is just another modification of the principal component estimator.

The estimator $\tilde{\beta}_{ss}$ can be obtained by solving the problem

$$\begin{aligned} &\text{minimize } (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \\ &\text{subject to: } \beta'T_s\Lambda_s T_s'\beta \leq d^2. \end{aligned} \tag{5.34}$$

The relation between β_g^+, β_f^+, and $\tilde{\beta}_{ss}$ is illustrated for the case $t = 3, g = 1, s = 2$ in Fig. 5. It should be noted that Sclove [1968] recommended this estimator for $s \geq 3$ in analogy with Stein [1960].

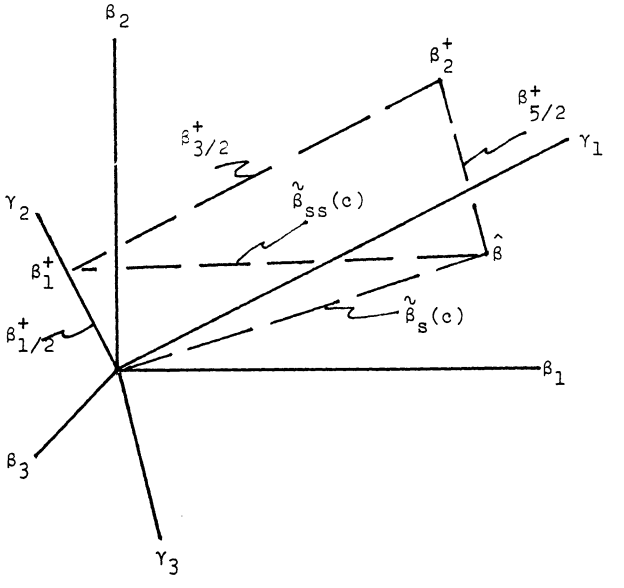

FIGURE 5

GENERALIZED INVERSE AND SHRUNKEN ESTIMATORS FOR $t = 3$, $s = 2$

In yet another variation on principal component regression, Webster *et al.* [1974] proposed what they called latent root regression. To relate their suggestion to the present development, let $A = [Y \mid X]$ be the $n \times (t + 1)$ matrix of standardized variables, let θ denote the diagonal matrix of eigenvalues, θ_i , of $A'A$ and H denote the corresponding orthogonal matrix of eigenvectors, h_i . That is,

$$H'A'AH = \theta \tag{5.35}$$

and

$$H'H = I. \tag{5.36}$$

Further, let the first row of H be denoted by h' and the remaining components by H_t so that in partitioned form

$$H = \begin{bmatrix} h' \\ H_t \end{bmatrix}. \tag{5.37}$$

It is then easily shown that the least squares estimate of β is given by

$$\hat{\beta} = -H_t v \tag{5.38}$$

where v is the solution to the problem

$$\begin{aligned} &\text{minimize } v'\theta v \\ &\text{subject to: } h'v = 1. \end{aligned} \tag{5.39}$$

To see this, let $\alpha = Hv$ and observe that the solution is $\hat{\alpha}' = (1, -\hat{\beta})$.

Webster *et al.* [1974] observed that if $\theta_i = 0$, there is an exact relation between the dependent and the independent variables. If, in addition, the corresponding component of h' is zero, there is a linear dependence among the independent variables. Based on this observation, they suggested that if these two quantities are small, the problem is unstable and, hence, recommended that these quantities be assumed to be zero in the above development. That is, the corresponding components of v are assumed equal to zero, or equivalently, v is the solution to the problem

$$\begin{aligned} &\text{minimize } v'\theta v \\ &\text{subject to: } h'v = 1 \\ &\qquad\qquad (0 \;\; I_s)v = 0. \end{aligned} \tag{5.40}$$

In this case the last s components of v are set to zero.

To provide a geometric description of the resulting estimate of β, let $\alpha = Hv$ in (5.40) and observe that an equivalent problem is

$$\begin{aligned} &\text{minimize } (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \\ &\text{subject to: } H_{ts}'\beta = h_s . \end{aligned} \tag{5.41}$$

Here the matrix H_{ts} is the $t \times s$ matrix consisting of the last s columns of H_t and the vector h_s contains the corresponding components of h.

A comparison of latent root regression, that is, the problem (5.41) with principal component regression as described by the problem (5.22) can now be made. In (5.22), the constraint space is the intersection of a set of hyperplanes passing through the origin. In (5.41) the hyperplanes may have different coefficients and need not pass through the

origin. The extent to which these problems differ may be explained by examining these two sets of equations. Recalling that Webster *et al.* [1974] required that the components of h_s be small, it remains to compare the elements of T_s in (5.22) with those of H_{ts} in (5.42). It can be shown that the condition $\lambda_i = 0$ in (5.14) is equivalent to the condition that θ_{i+1} and the $(i + 3)$st component of h are both zero. In this case the constraints in (5.22) and (5.42) are identical. Thus the problems will differ as a function of the criterion for assuming that these components are small.

The distinctions among the principal component solution, $\beta_g{}^+$, the generalized inverse solution, $\beta_f{}^+$, and the latent root solution, $\tilde{\beta}_L$ are indicated in Fig. 4 for the case $t = 2$, $s = 1$.

5.4. Relation of Ridge to Principal Component Estimators.

Based on the geometric relations obtained by expressing the biased estimators as solutions to constrained minimization problems, it is clear that these estimators are generally distinct. However, depending on the choice of parameters for the method and the parameters of the model, the estimators might be in close agreement.

Allen [1974] provided a characterization of ridge estimation which can be used to relate ridge and principal component estimators. Suppose the original data is augmented by dummy observations yielding the model

$$\begin{bmatrix} Y \\ 0 \end{bmatrix} = \begin{bmatrix} X \\ M \end{bmatrix} \beta + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \tag{5.42}$$

where Var $(e_2) = \sigma^2 W^{-1}$, $W = \text{Diag}\ (w_i)$ and M is, for the moment, arbitrary. The least squares estimator of β for (5.42) is given by

$$\begin{aligned} \tilde{\beta} &= (X'X + M'WM)^{-1}X'Y \\ &= (I - (X'X)^{-1}M'[W^{-1} + M(X'X)^{-1}M']^{-1}M)\hat{\beta}. \end{aligned} \tag{5.43}$$

Letting $M = T'$, as defined by (5.14), the first expression for $\tilde{\beta}$ in (5.43) is just the generalized ridge estimator described by Hoerl and Kennard [1970a]. The elements of W, which are the ridge parameters, reflect the "degree of belief" in the dummy observations. Letting $M = T_s{}'$, as in (5.19), and assuming the elements of W are large, the second form of (5.43) gives the principal component estimator as in (5.23). The generalized inverse estimator (5.29) is obtained by allowing the component of W corresponding to the column t_{g+1} of T_s to be finite and given by $w_{g+1} = \lambda_{g+1}(1 - f + g)/(f - g)$. The Sclove shrunken estimator (5.32) follows by letting all $w_i = \lambda_i(1 - c)/c$ where c is defined by (5.33).

Hocking *et al.* [1975] described a class of biased estimators which allows for additional algebraic and geometric comparisons of these biased estimators. The results of Hemmerle [1974] on iterative estimation of the biasing parameters in ridge regression are extended to include iteration on other biased estimators.

Although simulation and theoretical results are not overwhelming at this time, there does appear to be some potential merit to biased estimators, especially for $t \geq 3$. The generalized inverse estimator and Sclove's shrunken estimator appear to have the advantage of flexibility over their predecessors, principal component regression and Stein shrinkage. The results of Hocking *et al.* [1975] show that simple ridge regression includes all of these as limiting cases and is more flexible with regard to reflecting the correlation structure of the input variables. The analysis of the principal components might well be included

with the ridge analysis to identify the sources of degeneracy. The generalized ridge estimator includes all others as either special or limiting cases and merits further study.

5.5. Examples.

5.5.1. Biased Estimates for the Gas Mileage Data. Fig. 6 shows the ridge trace plots for the more interesting variables and suggests that stability is achieved for $0.15 \leq k \leq 0.25$ with rather substantial changes in the magnitude of the regression coefficients for these variables. The maximum variance inflation factor for $k = 0$ is VIF $= 21.6$ which decreases rapidly to VIF $= 1.0$ for $k = 0.20$. Based on the recommendation of Marquardt [1970], the range on k would appear to be $0.05 \leq k \leq 0.10$. The computation of (5.13) as recommended by McDonald and Galarneau [1975] suggests $k > 1.0$ while Farebrother [1975] would recommend $k = \hat{\sigma}^2/\hat{\beta}'\hat{\beta} = 0.01$. and Hoerl *et al.* [1975] obtain $k = t\hat{\sigma}^2/\hat{\beta}'\hat{\beta} = 0.1$. The marked disagreement between these methods of choosing k may not be typical, but it indicates that the choice of k is not well defined.

The eigenvalue analysis of $X'X$ reveals three small roots, namely $\lambda_{10} = 0.024$, $\lambda_9 = 0.054$ and $\lambda_8 = 0.081$ whose reciprocals account for over 76 percent of the total of $\sum \lambda_i^{-1} = 94.9$. The vector associated with the smallest root, λ_{10}, suggests a rather strong linear relation between variables 5, 7 and 9, which, in view of their descriptions, seems reasonable. The vectors associated with the other two roots are less definite but suggest a relation between

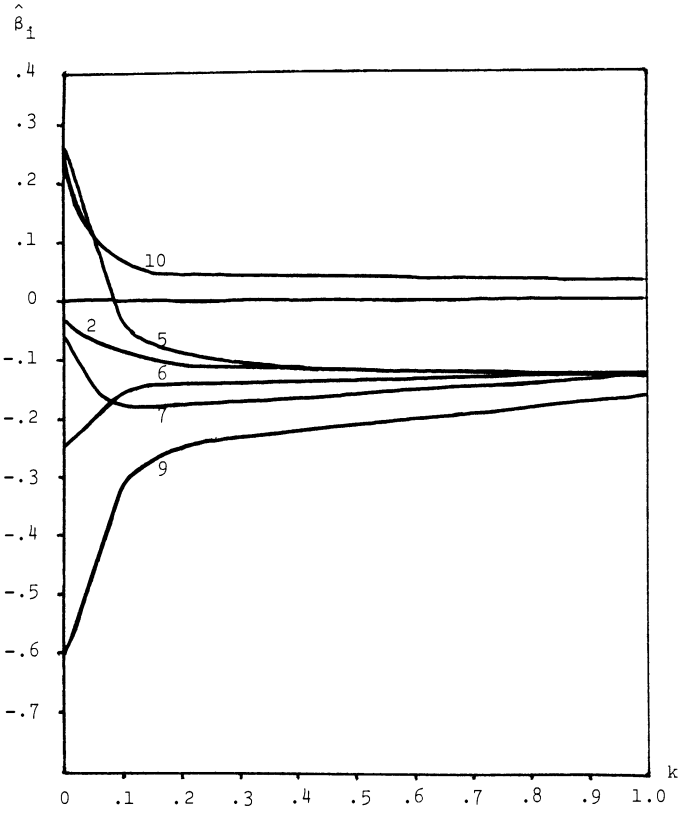

FIGURE 6

RIDGE TRACE FOR GAS MILEAGE DATA

variables 2, 6 and 10. It is apparently these relations which account for the behavior of the ridge plots for these variables. There appears to be justification for a reduction of rank of at least one and possibly two.

The eigenvalue analysis of $A'A$ as recommended by Webster *et al.* [1974] does not yield any roots satisfying their criteria, but the vector associated with the smallest root suggests the same relation between variables 5, 7 and 9 noted above.

Table 7 shows the least squares estimates, the ridge estimates for $k = 0.08$ (VIF = 2.5) and the principal component estimates for a reduction in rank of one and two. Generalized inverse estimates are not shown but can be computed by taking appropriate linear combinations as in (5.29). Similarly, the Stein and Sclove shrinkage estimators are not shown but are easily computed using $c = 0.94$ for equation (5.3) and $c_0 = 0.586$ for equation (5.33), shrinking on the three smallest eigenvalues.

Comparing the estimates in Table 7 and contrasting them with the subset estimators reveals a number of points. The ridge estimator for $k = 0.08$ and principal component estimator for $s = 2$ are reasonably close, the agreement being even better for $k = 0.20$. There is evidence that the near degeneracy has resulted in magnifying the estimates of certain coefficients, in particular those on variables 5, 6, 9 and 10, while the effect of variable 7 has been underestimated.

Recalling the linear relations mentioned earlier in this section and the analysis of Section 4.9.1, it appears that subset analysis has been reasonably effective in picking up the degeneracies. The magnitude of the coefficient on variable 9 reflects the fact that this variable has assumed the role of variable 7 as well.

Space does not permit a continuation of this detailed analysis, but it is evident that a combination of the subset and biased procedures is required to provide a good understanding of this data.

5.5.2. Biased Estimators for the Air Pollution Data. Fig. 7 shows the ridge trace plots for some of the interesting variables for this data. The complete plots and a detailed analysis are given in McDonald and Schwing [1973]. Those authors suggested that stability is achieved for $k = 0.20$. The variance inflation factor decreases rapidly in the range $0 \leq k \leq 0.02$ and is approximately one for $k = 0.18$, suggesting the range $0.02 \leq k \leq 0.08$. The evaluation of (5.13) suggests $k = 0.01$ while $\hat{\sigma}^2/\hat{\beta}'\hat{\beta} = 0.002$ and $t\hat{\sigma}^2/\hat{\beta}'\hat{\beta} = 0.03$. In this

TABLE 7

LEAST SQUARES, RIDGE AND PRINCIPAL COMPONENT ESTIMATES FOR GAS MILEAGE DATA ($\times 10$)

ESTIMATOR	VARIABLE									
	1	2	3	4	5	6	7	8	9	10
LEAST SQUARES	0.27	-0.33	2.09	0.80	2.74	-2.44	-0.54	0.70	-6.03	2.43
RIDGE k = 0.08	0.39	-0.73	1.76	0.78	-0.341	-1.49	-1.77	0.86	-3.31	0.97
PRINCIPAL COMPONENT s = 1	0.23	0.64	2.84	1.23	-1.24	-1.01	-2.63	1.02	-2.93	1.53
PRINCIPAL COMPONENT s = 2	-0.15	-1.19	1.76	0.81	-1.29	0.39	-2.62	0.83	-2.77	1.01

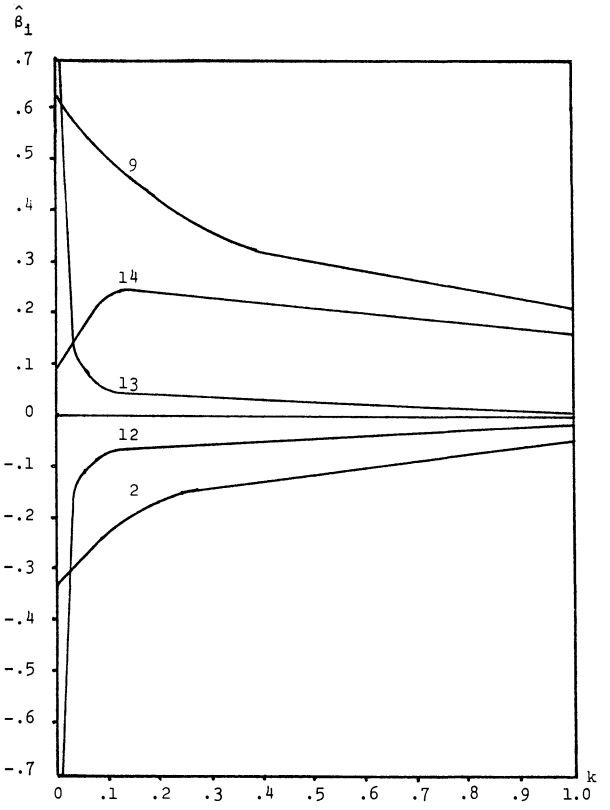

FIGURE 7.

RIDGE TRACE FOR AIR POLLUTION DATA

case, these methods for choosing k are in closer agreement than the previous example but the choice is not obvious.

The eigenvalue analysis of $X'X$ reveals two small roots, $\lambda_{15} = 0.0049$ and $\lambda_{14} = 0.046$. The sum of the reciprocals of these two roots accounts for 86 percent of the total of $\sum \lambda_i^{-1}$. The vector associated with the smallest root shows a strong linear relation between variables 12 and 13 which accounts for the behavior of the ridge trace. This relation was already evident from the correlation of 0.984 between these two variables. The vector associated with λ_{14} suggests a linear relation between variables 2, 4, 7, 9, 10. The ridge plots of variables 2 and 9 in Fig. 7 are probably a result of this relation.

Table 8 contains the least squares estimates, the ridge estimates for $k = 0.06$ (VIF = 2.6) and two principal component solutions for some of the variables. Those not listed had small coefficients. Contrasting the results of the principal component analysis with the subset analysis for $p = 9$ in Table 6 shows again that the optimal subset analysis was effective in detecting the degeneracies and in this case yields estimates which are in good agreement with those for $s = 2$.

The behavior of the coefficient of variable 14 relative to those for variables 12 and 13 is of interest. Ridge regression attempts to suppress the offsetting effects of variables 12 and 13 in favor of variable 14. The degeneracy observed suggests the deletion of either

TABLE 8

LEAST SQUARE, RIDGE AND PRINCIPAL COMPONENT ESTIMATES FOR ESSENTIAL VARIABLES IN AIR POLLUTION DATA (×10)

ESTIMATOR	VARIABLE										
	1	2	3	4	5	6	8	9	12	13	14
LEAST SQUARES	3.06	-3.19	-2.36	-2.12	-2.31	-2.33	0.84	6.40	-9.71	9.75	0.91
RIDGE k = 0.06	2.86	-2.41	-1.57	-0.96	-0.97	-1.35	1.09	5.53	-0.98	0.87	2.39
PRINCIPAL COMPONENT s = 1	3.20	-3.10	-2.30	-1.52	-2.03	-1.67	0.86	6.61	-0.29	0.01	2.39
PRINCIPAL COMPONENT s = 2	2.99	-2.77	-2.25	-1.16	-1.88	-1.59	0.86	6.92	-0.38	-0.36	2.39

variable 12 or 13 initially with subsequent deletion of the remaining one as indicated by the principal component or subset analysis.

The Stein and Sclove shrinkage coefficients in this case are 0.901 and 0.931 respectively.

5.5.3. Example 3: Biased Analysis of Artificial Data. The analysis of the preceding examples provides only relative comparisons since the "true" answers are not known. In this section, the artificial data described by Webster *et al.* [1974] is subjected to the same analysis. Their six variable data contained a near degeneracy in that the sum of the first four variables is essentially constant. Table 9 contains the true values of the β_i (in standard

TABLE 9

COMPARISON OF ESTIMATES FOR EXAMPLE 3 (×10)

ESTIMATOR	VARIABLE					
	1	2	3	4	5	6
TRUE	1.85	0.92	0.19	-1.84	2.77	9.23
LEAST SQUARES	-5.56	-7.81	-9.35	-10.81	3.78	8.72
RIDGE k = 0.02	1.60	-1.03	-0.56	-1.59	3.82	8.53
LATENT ROOT, s=1	2.35	-0.37	0.22	-0.68	3.89	8.75
PRINCIPAL COMPONENT, s=1	2.05	-0.65	-0.14	-1.06	3.88	8.75
p = 4	2.76	--	--	--	3.61	9.09
p = 5	2.83	--	0.82	--	3.95	8.73
p = 6	--	-2.60	-2.66	-3.72	3.88	8.75

form), the least squares, ridge, latent root and principal component estimates. In addition the best subset estimates are shown for $p = 4$, 5 and 6 yielding C_p values of 1.71 (min), 3.37 and 5.16, respectively.

As is evident by comparing the ridge estimator for $k = 0.02$ (VIF = 1.7) with the least squares estimator, the ridge trace on variables 1–4 changes dramatically for slight increases in k, yielding estimates which are substantially better than least squares.

The latent root procedure suggested by Webster *et al.* [1974] yields estimates which are roughly comparable to those obtained via principal component regression as was anticipated from the discussion of Section 5.3. Based on this isolated example, there is little to choose between the three biased procedures. Again, it should be emphasized that use of the ridge estimator without identifying the source of the degeneracy is not recommended.

The subset analysis picks up the degeneracy but especially for $p = 6$ does, at first glance, appear to do a poor job of estimating the coefficients. However, this conclusion is misleading since to arrive at comparable "true" coefficients, the relation $\sum x_i$ = constant should be imposed on the true model.

The fact that the C_p-plot indicates the deletion of two or three variables should not be considered as an indication of further reduction in rank but that, for prediction purposes, the smaller models are adequate as a consequence of the choice of σ^2.

6. ANALYSIS OF SUBSETS WITH BIASED ESTIMATORS

6.1. RIDGE-SELECT.

In Sections 3 and 4, the emphasis has been on the examination of subset regressions to enhance the understanding of the structure of the data and, in some cases, to suggest the use of a subset of the original variables. In Section 5, the motivation for biased estimators has been reviewed. It seems reasonable to conclude that the researcher might wish to include both concepts, that is, biased estimation and subset examination, in his analysis.

One suggestion might be to perform the subset analysis as described and then use a biased procedure for determining the estimators for either the full or subset equation as is appropriate. If the original t-variable model justifies a biased estimation procedure, for example, because of multicollinearity, then it is reasonable to conclude that the subset analysis should also be based on biased estimators. Unfortunately, the subset analyses found in the literature are generally based on least squares estimates. It seems reasonable to conclude that in the presence of near-degeneracy these subset analyses might be misleading, suggesting data structures which are not really present or suggesting subset equations which are inferior to other candidate subsets.

A possible solution to the problem would be to perform the biased analysis on all possible subsets to obtain plots of the type described in Section 4 for analysis. Thus, for example, if residual sum of squares is the criterion, the residual sum of squares obtained from the biased procedure would be plotted against subset size and the subset examinations performed analagous to those described in Section 4.

Apart from conceptual questions, an obvious drawback to this recommendation is the amount of computation required. If ridge regression is used, it would require, in addition to the solution of the normal equations, a determination of the value of k for each subset. If one of the variants of principal component regression is used, it would require an eigenvalue analysis for each subset. It is clear that such a recommendation is not feasible for even a modest number of independent variables. If such a procedure is to be considered,

there is clearly a need for an algorithm such as SELECT which will reveal interesting subsets without requiring the evaluation of all subsets. Unfortunately, there does not appear to be a simple solution to this problem.

There is, however, a compromise procedure which is conceptually quite simple and may be useful. This procedure is particularly simple when applied to ridge regression, as the current SELECT program can be used to perform the computations. The restriction which is forced on us is that the same value of k must be used for all subsets. Of course, the analysis could be repeated for several values of k and the combined results analyzed.

To explain the procedure it is necessary to recall the fundamental optimality principle that led to SELECT and to note that any of the biased estimation procedures can be described as a constrained minimization problem. Specifically, the problem of determining the ridge estimator for a subset of variables may be described as

$$\begin{aligned} &\text{minimize } Q(\beta) \\ &\text{subject to: } \beta'\beta = d^2 \\ &\qquad \beta_i = 0 \quad i \in R. \end{aligned} \tag{6.1}$$

Here d is the radius of the hypersphere which is determined by the specified value of k and R is the set of indices corresponding to variables which have been deleted. The addition of the constraint $\beta'\beta = d^2$ does not affect the basic monotonicity property which led to SELECT and, hence, it is possible to develop an algorithm which will determine the subset of size p whose ridge estimator for a given value of k is "best" without evaluating all subsets. These comments are also valid for the other types of biased estimators. Thus, for principal component regression, the constraint $\beta'\beta = d^2$ in (6.1) is replaced by the constraint $T_*'\beta = 0$. Note that this is equivalent to assuming the same reduction in rank of X, namely s, for all subsets and that the same linear constraints apply for each subset. This seems to be harder to accept than the assumption of fixed k in ridge regression.

Returning to subset analysis with ridge regression, it is of interest to note that the current SELECT program can be used to perform the computations. The procedure is to use, as input to SELECT, the model

$$\begin{bmatrix} Y \\ 0 \end{bmatrix} = \begin{bmatrix} X \\ \sqrt{k}\, I \end{bmatrix} \beta + e \tag{6.2}$$

where, as in Section 5, the data have been standardized. As observed by Marquardt [1970] and Banerjee and Carr [1971], the least squares analysis of this model will yield the ridge estimator. For fixed k, the least squares subset analysis for this model will yield the ridge subset analysis. It should be noted that the residual sum of squares obtained by fitting (6.2) or any subset model will be greater by an amount $k\tilde{\beta}_R'\tilde{\beta}_R$ over the correct residual sum of squares obtained by solving (6.1). Fortunately, this quantity is a constant for all subsets for a given value of k since $k\tilde{\beta}_R'\tilde{\beta}_R = kd^2$. Thus, the relative magnitudes of the subset residual sums of squares are maintained. Plots of residual sums of squares may thus be obtained and analyzed as in Section 4. Repeating the analysis for various values of k may provide information as to the invariance of the data structure to the value of k.

At this time it is not obvious that a similar trick will yield a simple computational procedure to performing a subset analysis with principal component regression.

The merits of this RIDGE-SELECT procedure have not been investigated. An essential point to be considered is whether biased estimation is justified in the subset if it is justified for the full equation. As observed by Hoerl and Kennard [1970a], the elimination of vari-

ables may not remove near degeneracies from $X'X$ and, hence, biased estimation may still be appropriate. It may be noted that a Ridge-Stepwise procedure follows immediately by applying the usual stepwise algorithm to the data as suggested by equation (6.2).

6.2. RIDGE-SELECT for Gas Mileage Data.

Table 10 contains a summary of the best subsets identified by the RIDGE-SELECT procedure for $k = 0.05$ and $k = 0.10$. Recalling that the eigenvalue analysis suggested linear relations between variables 5, 7 and 9 and variables 2, 6 and 10, it is again apparent that these near degeneracies were detected by the subset analysis. The treatment is somewhat different in that RIDGE-SELECT appears to prefer variables 7 and 9 from the first group while SELECT chose 5 and 9. This is in agreement with the principal component results in Table 7. From the second group RIDGE-SELECT shows a preference for variables 2 and 6 while SELECT generally ignores variable 2. The significance of these differences is not clear but appears to be supported by the principal component results. The combined efforts of ridge regression and subset analysis in removing degeneracies plus the information provided by the subset analysis appears to have some merit.

7. SUMMARY

A number of problems arise because of the non-orthogonality of regression data. If X has orthogonal columns, the effects of individual variables are clear and the problems of estimation and subset selection are elementary. Unfortunately, with undesigned experiments the columns of X are rarely orthogonal and there may exist near dependencies which, in turn, cause high correlations between variables or sets of variables. In such cases, it is generally difficult to assess the effects of individual factors as their relations with other factors are often complex.

In an attempt to provide the user with information on which he can base his analysis,

TABLE 10

RIDGE-SELECT RESULTS FOR GAS MILEAGE DATA

p	VARIABLES (k=0.05)	VARIABLES (k=0.10)
2	9	9
3	2	2
4	6	6
5	3, 10, -2	3
6	2, 7, -10	7
7	8, 10, -2	8
8	4	10
9	2	4
10	1	1
11	5	5

it has been recommended that a number of subsets be evaluated and a variety of schemes for interpreting the results have been reviewed. With respect to computational procedures, the advantages of stepwise methods seem to be outweighed by the additional, and often superior, information provided by the all-possible or optimal algorithms.

The problem of interpreting the subset information is complex. The question of whether any of the proposed criteria are effective in identifying "true" variables is difficult to answer.

The motivations for variable selection and biased estimation are similar and it seems natural to consider a simultaneous application of these procedures. Thus, for example, it may well be appropriate to apply additional biasing, according to one of the techniques in Section 5, to the subsets suggested as in Section 4. Alternatively, the bias may be introduced first so as to reflect the choice of subsets as in Section 6.

The multicollinearity problem seems to have been given too little attention in the statistics literature. Quite apart from any preference for biased or unbiased estimators, an eigenvalue examination should be an integral part of regression analysis. The actions taken as a result of this examination may vary. The user may conclude that ordinary least squares is appropriate, he may choose to use a biased estimator or he may decide to take additional observations according to an orthogonal design in the factor space. The important point is that the user should be aware of the presence of near-singularities. A considerable amount of work is required to provide adequate guidelines on the following:

(a) *What constitutes a serious multicollinearity problem?*

(b) *What type of biased estimator is preferable?*

(c) *How much bias should be allowed?*

The objective of this paper has been primarily to review the literature on variable analysis in regression and, to some extent, offer suggestions for the user. No definitive solutions were expected or realized. A rigorous mathematical analysis of most of the techniques is difficult but, in any event, the value of a particular method can only be established by its performance in practice. Since this is difficult to assess with real data, the use of simulated data is suggested. Potential investigators are cautioned to examine the magnitude of the required simulation before proceeding.

8. ACKNOWLEDGMENTS

I would like to express my appreciation to Dr. F. M. Speed for his valuable comments and to Mr. M. J. Lynn for his help with the computations. Special thanks to Mrs. Cheryl Dees and Mrs. Carolyn Jones for their help in the preparation of the manuscript.

9. REFERENCES

Abt, K. [1967]. Significant independent variables in linear models. *Metrika 12*, 2–15.

Aitkin, M. A. [1974]. Simultaneous inference and the choice of variable subsets in multiple regression. *Technometrics 16*, 221–8.

Akaike, H. [1969]. Fitting autoregressive models for control. *Ann. Statist. Math. 21*, 243–7.

Akaike, H. [1971]. Autoregressive model fitting for control. *Ann. Statist. Math. 23*, 163–80.

Allen, D. M. [1971a]. Mean square error of prediction as a criterion for selecting variables. *Technometrics 13*, 469–75.

Allen, D. M. [1971b]. The prediction sum of squares as a criterion for selecting prediction variables, Tech. Report No. 23, Dept. of Statist., Univ. of Kentucky.

Allen, D. M. [1972]. Biased prediction using multiple linear regression. Tech. Report No. 36, Dept. of Statist., Univ. of Kentucky.

Allen, D. M. [1974]. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics 16*, 125–7.

Anderson, R. L., Allen, D. M. and Cady, F. B. [1970]. Selection of predictor variables in linear multiple

regression. *Statistical Papers in Honor of George W. Snedecor*, Bancroft, T. A. (ed.), Iowa State Univ. Press, 3–17.

Anderson, R. L. and Battiste, E. L. [1970]. The use of prior information in linear regression analysis. *Presented at 7th International Biometrics Conference*, Hanover, Germany.

Andrews, D. [1974]. A robust method for multiple linear regression. *Technometrics 16*, 523–31.

Anscombe, F. J. [1961]. Examination of residuals. Univ. of California, *Proc. 4th Berkeley Symposium*, 1–36.

Anscombe, F. J. and Tukey, J. W. [1963]. The examination and analysis of residuals. *Technometrics 5*, 141–60.

Anscombe, F. J. [1967]. Topics in the investigation of linear relations fitted by the method of least squares. *J. R. Statist. Soc. 29*, 1–52.

Anscombe, F. J. [1973]. Graphs in statistical analysis. *The Amer. Statistician 27*, 17–21.

Arvesen, J. N. and McCabe, G. P. [1975]. Variable selection in regression analysis. *J. Amer. Statist. Assoc. 70*, 166–70.

Banerjee, K. S. and Carr, R. N. [1971]. A comment on ridge regression, biased estimation for non-orthogonal problems. *Technometrics 13*, 895–8.

Bargmann, R. E. [1962]. Representative selection of variables. *Presented at the 122nd annual meeting of the Amer. Statist. Assoc.*, Minneapolis, Minnesota.

Barr, A. J. and Goodnight, J. H. [1971]. Statistical analysis system. Raleigh Student Supply Store, North Carolina State Univ.

Barrett, J. P. [1974]. The coefficient of determination—some limitations. *The Amer. Statistician 28*, 19–20.

Beale, E. M. L. [1970a]. A note on procedures for variable selection in multiple regression. *Technometrics 12*, 909–14.

Beale, E. M. L. [1970b]. Selecting an optimum subset. *Integer and Nonlinear Programming.* Abadie, J. (ed.), North Holland Publishing Co., Amsterdam.

Beale, E. M. L.; Kendall, M. G., and Mann, D. W. [1967]. The discarding of variables in multivariate analysis. *Biometrika 54*, 357–66.

Beaton, A. E. [1964]. The use of special matrix operators in statistical calculus. Res. Bull. RB-64-51, Educational Testing Service, Princeton, New Jersey.

Beaton, A. E. and Tukey, J. W. [1974]. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics 16*, 147–85.

Bendel, R. B. and Afifi, A. A. [1974]. Comparison of stopping rules in forward stepwise regression. *Presented at Joint Statistical Meetings*, St. Louis, Missouri.

Birch, M. W. [1964]. A note on the maximum likelihood estimation of a linear structural relationship. *J. Amer. Statist. Assoc. 59*, 1175–8.

Box, G. E. P. [1966]. Use and abuse of regression. *Technometrics 8*, 625–9.

Boyce, D. E., Farhi, A., and Weischedel, R. [1974]. Optimal subset selection. Lecture notes in *Economics and Mathematical Systems*, Beckman, H. E. and Kunzi, H. P. (eds.) Springer-Verlag, New York.

Breaux, H. J. [1968]. A modification of Efroymson's technique for stepwise regression analysis. *Communication of the ACM 11*, 556–7.

Breaux, H. J., Campbell, L. W. and Torrey, J. C. [1966]. Stepwise multiple regression statistical theory and computer program description. Report No. 1330, Ballistic Research Laboratories.

Browne, M. W. [1970]. A critical evaluation of some reduced-rank regression procedures. Res. Bull. 70-21, Educational Testing Service, Princeton, New Jersey.

Browne, M. W. [1969]. Precision of prediction. Res. Bull. 69-69, Educational Testing Service, Princeton, New Jersey.

Cady, F. B. and Allen, D. M. [1972]. Combining experiments to predict future yield data. *Agronomy J. 64*, 21–4.

Chambers, J. M. [1971]. Regression updating. Bell Telephone Laboratories, Inc., Murray Hill, New Jersey.

Cohen, A. C., Jr. [1957]. Restriction and selection in multinormal distributions. *Ann. Math. Statist. 28*, 731–41.

Cohen, A. C., Jr. [1955]. Restriction and selection in samples from bivariate normal distributions. *J. Amer. Statist. Assoc. 50*, 884–93.

Coniffe, D. and Stone, J. [1973]. A critical view of ridge regression. *The Statistician 22*, 181–7.

Crocker, D. C. [1972]. Some interpretations of the multiple correlation coefficient. *The Amer. Statistician 26*, 31–3.

Daling, J. R. and Tamura, H. [1970]. Use of orthogonal factors for selection of variables in a regression equation—an illustration. *Applied Statist. 19*, 260–8.

Daniel, C. and Wood, F. S. [1971]. *Fitting Equations to Data.* Wiley, New York.

Debertin, D. L. and Freund, R. J. [1973]. Significance tests and selection of variables in multiple regression. Preliminary report, Institute of Statist., Texas A&M Univ., College Station, Texas.

Dickinson, A. W. [1968]. Pitfalls in the use of regression analysis. *Presented at Share XXX, Monsanto Company*, St. Louis, Missouri.

Diehr, G. and Hoflin, D. R. [1974]. Approximating the distribution of the sample R^2 in best subset regressions. *Technometrics 16*, 317–20.

Draper, N. R. and Smith, H. [1966]. *Applied Regression Analysis*. Wiley, New York.

Edwards, J. B. [1969]. The relation between the F-test and $\bar{R}^2$. *The Amer. Statistician 23*, 28.

Efroymson, M. A. [1960]. Multiple regression analysis. *Mathematical Methods for Digital Computers*. Ralston A. and Wilf, H. S., (eds.), Wiley, New York.

Efroymson, M. A. [1966]. Stepwise regression—a backward and forward look. *Presented at the Eastern Regional Meetings of the Inst. of Math. Statist.*, Florham Park, New Jersey.

Farebrother, R. W. [1975]. The minimum mean square error linear estimator and ridge regression. *Technometrics 17*, 127–8.

Farebrother, R. W. [1973]. The minimum mean square error linear estimater and ridge regression. Dept. of Econometrics, Univ. of Manchester, Manchester, U. K.

Feiveson, A. H. [1973]. Selecting variables in regression and classification by thresholding. Ph.D. Dissertation, Texas A&M Univ., College Station, Texas.

Forsythe, A. B., Engelman, L., and Jennrich, R. [1973]. Stopping rule for variable selection in multiple regression. *J. Amer. Statist. Assoc. 68*, 75–7.

Freund, R. J., Vail, R. W., and Clunies-Ross, C. W. [1961]. Residual analysis. *J. Amer. Statist. Assoc. 56*, 98–104.

Furnival, G. M. [1971]. All possible regressions with less computation. *Technometrics 13*, 403–8.

Furnival, G. M. and Wilson, R. W., Jr. [1974]. Regression by leaps and bounds. *Technometrics 16*, 499–512.

Garside, M. J. [1965]. The best subset in multiple regression analysis. *Applied Statist. 14*, 196–200.

Garside, M. J. [1971]. Some computational procedures for the best subset problem. *Applied Statist. 20*, 8–15.

Gaylor, D. W. and Merrill, J. A. [1968]. Augmenting existing data in multiple regression. *Technometrics 10*, 73–81.

Gaylor, D. W. and Sweeney, H. C. [1965]. Design for optimal prediction in simple linear regression. *J. Amer. Statist. Assoc. 60*, 205–16.

Goldberger, A. S. [1968]. *Topics in Regression Analysis*. Macmillan Co., Toronto, Canada.

Goldberger, A. S. [1961]. Stepwise least squares: residual analysis and specifications. *J. Amer. Statist. Assoc. 56*, 998–1000.

Goldberger, A. S. and Jochems, D. B. [1961]. Note on stepwise least squares. *J. Amer. Statist. Assoc. 56*, 105–10.

Gorman, J. W. and Toman, R. J. [1966]. Selection of variables for fitting equations to data. *Technometrics 8*, 27–51.

Greenberg, E. [1975]. Minimum variance properties of principal component regression. *J. Amer. Statist. Assoc. 70*, 194–7.

Gross, A. J. [1966]. An application of the multiple correlation coefficient in the reduction of variates. The Rand Corp., Santa Monica, California.

Gugel, H. W. [1972]. SELECT—A computer program for isolating "best" regressions. General Motor Corp. Res. Pub., GMR, 1216.

Gunst, R. F. and Mason, R. L. [1973]. Some additional indices for selecting variables in regression. *Biometrics 30*, Abstract #2179, 382.

Hamaker, H. C. [1962]. On multiple regression analysis. *Statistica Neerlandica 16*, 31–56.

Hawkins, D. M. [1973]. On the investigation of alternative regressions by principal component analysis. *Applied Statist. 22*, 275–86.

Helms, R. W. [1974]. The average estimated variance criterion for the selection-of-variables problem in general linear models. *Technometrics 16*, 261–74.

Hemmerle, W. J. [1967]. *Statistical Computations on a Digital Computer*. Blaisdell, Waltham, Massachusetts.

Hemmerle, W. J. [1974]. An explicit solution for generalized ridge regression. *Technometrics 17*, 309–14.

Hocking, R. R. and Leslie, R. N. [1967]. Selection of the best subset in regression analysis. *Technometrics 9*, 531–40.

Hocking, R. R. [1972]. Criteria for selection of a subset regression: which one should be used. *Technometrics 14*, 967–70.

Hocking, R. R. and LaMotte, L. R. [1973]. Using the SELECT program for choosing subset regressions.

Proc. of Univ. of Kentucky conference on regression with a large number of predictor variables. Thompson, W. O. and Cady, F. B. (eds.) Dept. of Statist., Univ. of Kentucky, Lexington, Kentucky.

Hocking, R. R. [1974]. Misspecification in regression. *The Amer. Statistician 28*, 39–40.

Hocking, R. R. [1975]. Selection of the best subset of regression variables. To appear in *Statistical Methods for Digital Computers*, Enslein, Ralston and Wilf, (eds.)

Hocking, R. R., Speed, F. M. and Lynn, M. J. [1975]. A class of biased estimators in linear regression. *Presented at ASA, IMS. ENAR Regional Meetings*, St. Paul, Minnesota.

Hoerl, A. E. [1962]. Application of ridge analysis to regression problems. *Chem. Eng. Progress 58*, 54–9.

Hoerl, A. E. [1964]. Ridge analysis. *Chem. Eng. Progress Symposium 60*, 67–77.

Hoerl, A. E. [1970]. Ridge regression. *Presented at the Joint Meeting*, Chapel Hill, North Carolina.

Hoerl, A. E. and Kennard, R. W. [1970a]. Ridge regression: biased estimation for non-orthogonal problems. *Technometrics 12*, 55–67.

Hoerl, A. E. and Kennard, R. W. [1970b]. Ridge regression: applications to non-orthogonal problems. *Technometrics 12*, 69–82.

Hoerl, A. E. and Kennard, R. W. [1975], Ridge regression: iterative estimation of the biasing parameter. (Preliminary report.) *IMS Bull.* (Abstract) *4*, 135.

Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. [1975]. Ridge regression: some simulations. *Comm. in Statist. 4*, 105–23.

Holland, P. W. [1973]. Weighted ridge regression: combining ridge and robust regression methods. National Bureau of Economic Research, Working Paper No. 11. Cambridge, Massachusetts.

Holms, A. G. [1974]. Chain pooling to minimize prediction error in subset regression. NASA T. M. A-71645. *Presented at Joint Statistical Meetings*, St. Louis, Missouri.

Horst, P. [1934]. Item analysis by the method of successive residuals. *J. Experimental Education 2*, 254–63.

Hotelling, H. [1940]. The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Ann. Math. Statist. 11*, 271–83.

Jeffers, J. N. R. [1967]. Two case studies in the application of principal component analysis. *Applied Statist. 3*, 225–36.

Johnston, J. [1972]. *Econometric Methods*, (2nd ed.), McGraw-Hill, New York.

Jones, H. L. [1946]. Linear regression functions with neglected variables. *J. Amer. Statist. Assoc. 41*, 356–69.

Jones, R. H. [1973]. Selection of a subset regression. Dept. of Information and Computer Science, Univ. of Hawaii.

Kendall, M. G. [1957]. *A Course in Multivariate Analysis*. Hafner Publishing Co., New York.

Kennard, R. W. [1971]. A note on the C_p statistic. *Technometrics 13*, 899–900.

Kennedy, W. J. and Bancroft, T. A. [1971]. Model-building for prediction in regression based on repeated significance tests. *Ann. Math. Statist. 42*, 1273–84.

Kirton, H. C. [1967]. "Best" models in multiple regression analysis. N.S.W. Dept. of Agriculture, Sydney, Australia.

Laird, R. J. and Cady, F. B. [1969]. Combined analysis of yield data from fertilizer experiments. *Agronomy J. 61*, 829–34.

LaMotte, L. R. [1972]. The SELECT routines: a program for identifying best subset regression. *Applied Statist. 21*.

LaMotte, L. R. and Hocking, R. R. [1970]. Computational efficiency in the selection of regression variables. *Technometrics 12*, 83–93.

Larsen, W. A. and McCleary, S. J. [1972]. The use of partial residual plots in regression analysis. *Technometrics 14*, 781–90.

Larson, H. J. and Bancroft, T. A. [1963]. Sequential model building for prediction in regression analysis. *Ann. Math. Statist. 34*, 462–79.

Lawley, D. N. and Maxwell, A. E. [1973]. Regression and factor analysis. *Biometrika 60*, 331–8.

Lindley, D. V. [1968]. The choice of variables in multiple regression. *J. R. Statist. Soc. 30*, 31–53.

Linhart, H. [1960]. A criterion for selecting variables in a regression analysis. *Psychometrika 25*, 45–8.

Lott, W. F. [1973]. The optimal set of principal component restrictions on a least squares regression. *Comm. Statist. 2*, 449–64.

Lotto, G. [1962]. On the generation of all possible stepwise combinations. *Math. Comput. 16*, 241–3.

Lowerre, J. M. [1974]. On the mean square error of parameter estimates for some biased estimators. *Technometrics 16*, 461–4.

Lucas, H. L. and Linnerud, A. C. [1967]. Observations on the selection of predictors. Biostatistics Seminar, Univ. of North Carolina, Chapel Hill, North Carolina.

McCabe, G. P., Jr. and Ross, M. A. [1973]. A stepwise algorithm for selecting regression variables using cost criteria. Mimeograph Series No. 349, Dept. of Statist., Purdue Univ.

McCabe, G. P., Jr. and Arvesen, J. N. [1974]. A subset selection procedure for regression variables. *J. Statist. Comput. Simul. 3*, 137–46.

McCabe, G. P., Jr. [1973]. Computations for variable selection in discriminant analysis. *Technometrics 17*, 103–9.

McCornack, R. L. [1968]. A comparison of three predictor selection techniques in multiple regression. SP-3029, System Dev. Corp., Santa Monica, California.

McDonald, G. C. and Galarneau, D. I. [1975]. A Monte Carlo evaluation of some ridge-type estimators. *J. Amer. Statist. Assoc. 70*, 407–16.

McDonald, G. C. and Schwing, R. C. [1973]. Instabilities of regression estimates relating air pollution to mortality. *Technometrics 15*, 463–81.

Mallows, C. L. [1964]. Choosing variables in a linear regression: a graphical aid. *Presented at the Central Regional Meeting of the Inst. of Math. Statist.*, Manhattan, Kansas.

Mallows, C. L. [1966]. Choosing a subset regression. *Presented at Joint Statistical Meetings*, Los Angeles, California.

Mallows, C. L. [1967]. Choosing a subset regression. Bell Telephone Laboratories, unpublished report.

Mallows, C. L. [1973a]. Some comments on C_p. *Technometrics 15*, 661–75.

Mallows, C. L. [1973b]. Data analysis in a regression context. *Proc. of Univ. of Kentucky conference on regression with a large number of predictor variables*. Thompson, W. O. and Cady, F. B. (eds.), Dept. of Statist., Univ. of Kentucky, Lexington, Kentucky.

Mantel, N. [1970]. Why stepdown procedures in variable selection. *Technometrics 12*, 591–612.

Marquardt, D. W. [1970]. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics 12*, 591–612.

Marquardt, D. W. and Snee, R. D. [1973]. Ridge regression. *Proc. of Univ. of Kentucky conference on regression with a large number of predictor variables*. Thompson, W. O. and Cady, F. B. (eds.), Dept. of Statist., Univ. of Kentucky, Lexington, Kentucky.

Marquardt, D. W. [1974a]. Biased estimators in regression. *Presented at SREB Summer Research Conference*, Winter Park, Florida.

Marquardt, D. W. [1974b]. Discussion (of Beaton and Tukey [1974]) *Technometrics 15*, 189–92.

Mason, R. L., Webster, J. T., and Gunst, R. F. [1975]. Sources of multicollinearity in regression analysis. *Comm. in Statist. 4*, 277–92.

Massy, W. F. [1965]. Principal components regression in exploratory statistical research. *J. Amer. Statist. Assoc. 60*, 234–46.

Mayer, L. S. and Wilke, T. A. [1973]. On biased estimation in linear models. *Technometrics 15*, 497–508.

Morgan, J. A. and Tatar, J. F. [1972]. Calculation of the residual sum of squares for all possible regressions. *Technometrics 14*, 317–25.

Narula, S. and Ramberg, J. S.]1972]. Letter to the editor. *The Amer. Statistician 26*, 42.

Newton, R. G. and Spurrell, D. J. [1967a]. A development of multiple regression for the analysis of routine data. *Applied Statist. 16*, 51–65.

Newton, R. G. and Spurrell, D. J. [1967b]. Examples of the use of elements for clarifying regression analysis. *Applied Statist. 16*, 165–71.

Newhouse, J. P. and Oman, S. D. [1971]. An evaluation of ridge estimators. Rand Report R-716-PR. Santa Monica, California.

Oosterhoff, J. [1963]. On the selection of independent variables in a regression equation. Preliminary Report 5319. Amsterdam. *Stichting Mathematisk Centrum.*

Pope, P. T. and Webster, J. T. [1972]. The use of an F-statistic in stepwise regression procedures, *Technometrics 14*, 327–40.

Pyne, D. A. [1970]. Relationships between sets of variables found by several model building procedures. Unpublished M.S. Thesis, Iowa State Univ.

Radhakrishnan, S. [1974]. Selection of variables in multiple regression. Ph.D. Dissertation, Univ. of Houston, Houston, Texas.

Rao, P. [1971]. Some notes on misspecification in multiple regression. *The Amer. Statistician 25*, 37–9.

Rosenberg, S. H. and Levy, P. S. [1972]. A characterization on misspecification in the general linear regression model. *Biometrics 28*, 1129–32.

Rothman, David. [1968]. Letter to the editor. *Technometrics 10*, 432.

Schatzoff, M., Fienberg, S., and Tsao, R. [1968]. Efficient calculation of all possible regressions. *Technometrics 10*, 769–79.

Schmidt, P. [1973a]. Methods of choosing among alternative linear regression models. Univ. of North Carolina, Chapel Hill, North Carolina.
Schmidt, P. [1973b]. Calculating the power of the minimum standard error choice criterion. *Intern. Econ. Review 14*.
Sclove, S. L. [1967]. Improved estimation of regression parameters. Tech. Report No. 125, Dept. of Statist., Stanford Univ., Palo Alto, California.
Sclove, S. L. [1968]. Improved estimators for coefficients in linear regression. *J. Amer. Statist. Assoc. 63*, 597–606.
Sclove, S. L. [1971]. On criteria for choosing a regression equation for prediction. Tech. Report No. 28, Dept. of Statist., Carnegie-Mellon Univ., Pittsburgh, Pennsylvania.
Scott, J. T., Jr. [1966]. Factor analysis and regression. *Econometrica 34*, 552–62.
Silvey, S. O. [1969]. Multicollinearity and imprecise information. *J. R. Statist. Soc. 31*, 539–52.
Snee, R. D. [1973]. Some aspects of nonorthogonal data analysis, part I. Developing prediction equations. *J. of Quality Technology 5*, 67–79.
Stein, C. M. [1960]. Multiple regression. *Contributions to Probability and Statistics*. Essays in Honor of Harold Hotelling, Olkin, I. (ed.), Stanford Univ. Press, 424–43.
Stone, M. [1974]. Cross-validatory choice and assessment of statistical predictions. *J. R. Statist. Soc. 30*.
Stowe, R. A. and Mayer, R. P. [1969]. Pitfalls of stepwise regression analysis. *Ind. and Eng. Chem. 61*, 11–16.
Summerfield, A. and Lubin, A. [1951]. A square root method of selecting a set of variables in multiple regression: I. The method. *Psychometrika 16*, 271–84.
Swindel, B. F. [1974]. Good ridge estimators based on prior information. *Presented at the Joint Statistical Meetings*, St. Louis, Missouri.
Theil, H. [1961]. *Economic Forecasts and Policy*. North Holland Publishing Co., Amsterdam.
Tomoyuke, T. and Kudo, Akio. [1974]. An algorithm related to all possible regression and discriminant analysis. *J. Japan Statist. Soc. 4*, 47–56.
Toro-Vizcarrondo, C. and Wallace, T. D. [1968]. A test of the mean square error criterion for restrictions in linear regression. *J. Amer. Statist. Assoc. 63*, 558–72.
Tukey, J. W. [1967]. Discussion (of Anscombe [1967]). *J. R. Statist. Soc. 29*, 47–8.
Wallace, T. D. [1964]. Efficiencies for stepwise regressions. *J. Amer. Statist. Assoc. 59*, 1179–82.
Walls, R. E. and Weeks, D. L. [1969]. A note on the variance of a predicted response in regression. *The Amer. Statistician 23*, 24–6.
Ward, J. F., Jr. [1974]. Restricted least squares and ridge estimators. *Presented at Joint Statistical Meetings*, St. Louis, Missouri.
Waugh, F. V. [1963]. The analysis of regression in subsets of variable. *J. Amer. Statist. Assoc. 31*, 729–30.
Webb, S. R. [1966]. Efficiency of a computer routine for matrix inversion with application to exhaustive regression. Rocketdyne, RM 1205-351, Conoga Park, California.
Webster, J. T. [1965]. On the use of a biased estimator in linear regression. *J. Indian Statist. Assoc. 3*, 82–90.
Webster, J. T., Gunst, R. F., and Mason, R. L. [1974]. Latent root regression analysis. *Technometrics 16*, 513–22.
Wherry, R. J. [1931]. A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Ann. Math. Statist. 2*, 440–51.
Williams, J. D. and Lindem, A. C. [1971]. Setwise regression analysis—a stepwise procedure for sets of variables. *Educational and Psychological Measurement 31*, 747–8.
Wong, Y. K. [1937]. On the elimination of variables in multiple correlation. *J. Amer. Statist. Assoc. 32*, 357–60.
Wood, F. S. [1972]. The linear and nonlinear least-squares curve fitting programs. *Presented at SHARE San Francisco Meetings*, San Francisco, California.
Wood, F. S. [1973]. The use of individual effects and residuals in fitting equations to data. *Technometrics 15*, 677–95.

Received January 1975, *Revised September* 1975

Key Words: Linear regression, Subset selection, Biased estimation.