Lockheed Electronics Company, Inc.
A SUBSIDIARY OF LOCKHEED CORPORATION
1830 NASA Road 1, Houston, Texas 77058
Tel. 713-333-5411

JSC-12700

Ref: 642-6932
Contract NAS 9-15800
Job Order 73-715-12

TECHNICAL MEMORANDUM

REGRESSIONS BY LEAPS AND BOUNDS AND BIASED ESTIMATION TECHNIQUES IN YIELD MODELING

By

N. Marquina

Approved By: T. C. Minter
T. C. Minter, Supervisor
Techniques Development Section

February 1979

LEC-12379

CONTENTS

TABLES

FIGURES

SYMBOLS

p	Total number of regressor variables
n	Number of observations
x	Independent or regressor variables
$\underline{X}_j$	The n-dimensional vector of observed values of the j*th* regressor variable
X	The n x p matrix whose columns are the vectors $\underline{X}_j$
$(X'X)^+$	Moore-Penrose pseudoinverse of the matrix X'X
y	The n-dimensional vector of observed values of the dependent variable
Y_i	The i*th* observation of the dependent variable
$\overline{Y}$	Average of the components of y
$\hat{Y}_i$	Estimated value of Y_i
β	True parameters or regression coefficients
b	Estimate of β
ε	The error term
σ	Variance of the components of error
ℓ_j	The j*th* eigenvalue of X'X in order of increasing magnitude
V_j	A normalized eigenvector of X'X associated with ℓ_j
S	The matrix whose columns are the eigenvectors V_j
L	The diagonal matrix of eigenvalues ℓ_j
S_r	Submatrix of S consisting of r columns

REGRESSIONS BY LEAPS AND BOUNDS AND BIASED ESTIMATION TECHNIQUES IN YIELD MODELING

1. INTRODUCTION

The prediction of yield estimates based on meteorological variables is discussed in this technical memorandum. The primary statistical tool for the analysis is linear parameter regression. Multiple linear regression analysis is a procedure for the analysis of the relationships between two sets of variables, independent or regressor variables and dependent or response variables, whose values are believed to be related to the set. Estimation of the coefficients of the regression model is usually performed using least squares.

The least-squares estimator of the regression coefficients has the desirable property of being unbiased and having minimum variance among the class of unbiased linear estimators. However, when near-linear relationships exist among the regressor variables (a situation known as multicollinearity) this minimum variance can be quite large. Thus, estimation procedures other than least squares appear to be desirable when multicollinearity exists among the regressor variables.

The meteorological variables currently used in yield modeling are highly correlated among themselves. A consequence of the multicollinearity present in the meteorological variables is the large variance of the regression coefficients. Many of the applications of regression analysis in yield modeling either explicitly or implicitly place reliance on individual parameter estimates. Inferences about cause-effect relationships between the response and regressor variables based on individual coefficient estimates can be misleading, even erroneous, when multicollinearity is present in the data. In the presence of multicollinearities, the estimated coefficients are highly unstable; the addition of one or more new observations can change the size and even the sign of some of the parameters (ref. 1).

In this memorandum it will be shown that techniques other than ordinary least squares (OLS) exist to deal with the problem of estimation with correlated predictor variables. In particular, latent root regression, principal components regression, ridge, and generalized ridge will be examined. Texas and Oklahoma weather data are used for the illustrations. The programs that implement these techniques were developed by the author while attending the University of Houston; the Industrial Engineering Department of the University of Houston provided the computer time for the sample runs.

In section 2, ordinary least squares are reviewed and the multicollinearity problem is defined. Section 3 deals with the problem of finding the best subset of variables to enter the regression analysis. Section 4 discusses three of the most important biased estimation techniques. In section 5, these ideas are applied to the weather and trend data for the Texas-Oklahoma Panhandle and Oklahoma, which were provided by National Aeronautics and Space Administration, Lyndon B. Johnson Space Center (NASA/JSC) personnel. Conclusions and recommendations are presented in section 6. References are listed in section 7.

2. ORDINARY LEAST SQUARES

2.1 LEAST-SQUARES ESTIMATION

The multiple linear regression model can be written as

$$Y_i = \beta_0^* + \beta_1^* x_{i1}^* + \beta_2^* x_{i2}^* + \cdots + \beta_p^* x_{ip}^* + \varepsilon_i \quad ; \quad i = 1, 2, \cdots, n \qquad (1)$$

where

Y_i = the yield for the i*th* year

$\beta_0^*, \beta_1^*, \cdots, \beta_p^*$ = unknown parameters referred to as the regression coefficients

x_{ij}^* = the value of the j*th* weather variable for the i*th* year

ε_i = random error term for the i*th* year

For the purposes of increased computational accuracy, the relationship (eq. (1)) can be transformed (ref. 2) by letting

$$x_{ij} = \frac{\left(x_{ij}^* - \bar{x}_j^*\right)}{\left[\sum_{i=1}^{n} \left(x^*_{ij} - \bar{X}^*\right)^2\right]^{1/2}}$$

where

$$\bar{x}_j^* = \frac{1}{n} \sum_{i=1}^{n} x_{ij}^*$$

The predictor variables are now standardized so that

$$\sum_{i=1}^{n} x_{ij} = 0$$

and

$$\sum_{i=1}^{n} x_{ij}^2 = 1 \quad ; \quad j = 1, 2, \cdots, p$$

Eq. (1) now becomes

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad ; \quad i = 1, 2, \cdots, n \tag{2}$$

where

$$\beta_0 = \beta_0^* + \sum_{j=1}^{p} \beta_j^* \bar{x}_j^*$$

and

$$\beta_j = \beta_j^* \left[\sum_{i=1}^{n} \left(x_{ij}^* - \bar{x}_j^* \right)^2 \right]^{1/2}$$

In vector notation, eq. (2) is written as

$$y = \beta_0 \underline{1} + X\beta + \varepsilon \tag{3}$$

where

y = an $n \times 1$ vector of yield measurements

$\underline{1}$ = an $n \times 1$ vector of 1's

$X = [\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_p]$ = an $n \times p$ matrix of constants

$\beta' = (\beta_1, \beta_2, \cdots, \beta_p)$ = a $1 \times p$ vector of unknown parameters

ε = an $n \times 1$ vector of random error terms

The following assumptions are used in this memorandum:

a. The elements of $X = [X_1, X_2, \cdots, X_p]$ are nonstochastic.

b. X has rank $p < n$.

c. The elements of y are observable random variables.

d. The elements of ε are unobservable random variables with $E[\varepsilon] = \underline{0}$ and $E[\varepsilon\varepsilon'] = \sigma^2 I_n$.

e. In addition, the assumption $\varepsilon \sim N(\underline{0}, \sigma^2 I_n)$ will be included when hypothesis testing is required.

It should be noted that assumption (b) states that the number of observations exceeds the number of parameters to be estimated and that no exact linear relations exist among the columns of X.

The least-squares estimate b of β is found by minimizing $\varepsilon'\varepsilon$ with respect to β, i.e.,

$$\text{minimize } \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta)$$
$$= y'y - 2\beta X'y + \beta'X'X\beta$$

The problem is reduced to solving

$$X'X\beta = X'y \tag{4}$$

which is obtained by using the fact that $\beta'X'y = y'X\beta$, differentiating $\varepsilon'\varepsilon$ with respect to β, and solving

$$\frac{\partial \varepsilon'\varepsilon}{\partial \beta} = 0$$

or $$-2X'y + 2X'X\beta = \underline{0}$$

These are the so-called normal equations. The solution to eq. (4) is given by

$$b = (X'X)^{-1}X'y \tag{5}$$

The least-squares estimator is unbiased and has minimum variance in the class of unbiased estimators of the regression coefficients. If the normality assumption (e) is valid, eq. (5) is also maximum likelihood (ref. 3).

The variance of the OLS estimator is given by

$$\text{Var}(b) = \sigma^2(X'X)^{-1} \tag{6}$$

The variance of the estimator of a particular coefficient, b_j, is

$$\text{Var}(b_j) = C_{jj}\sigma^2 \tag{7}$$

while $$\text{Cov}(b_i, b_j) = C_{ij}\sigma^2$$

where $$C = [C_{ij}] = (X'X)^{-1}$$

Eqs. (6) and (7) play a central role in the discussion of multicollinearity. A complete discussion of the derivation of the least-squares estimator and its various numerical and statistical properties can be found in many standard texts on statistical analysis; special mention should be made of reference 2.

2.2 MULTICOLLINEARITY

It is very well established that the regressor weather variables in the Center for Climatogical and Environmental Assessment (CCEA) yield model are highly correlated (refs. 4 and 5).

Let us analyze the effects of multicollinearity on the OLS estimator. Multicollinearity is a form of ill-conditioning within the matrix, X, of regressor variables in which for some set of constants $a_1, a_2, \cdots, a_p$ not all zero, we have

$$\sum_{j=1}^{p} a_j \underline{X}_j \approx \underline{0} \tag{8}$$

If the relationship is exact, there is said to be an exact multicollinearity among the regressor variables. In this case, $(X'X)^{-1}$ does not exist because the rank of X will be less than p. This implies that there is not one solution to the normal equations, but infinitely many solutions. To obtain a unique solution for eq. (4) when the rank of X is less than p, the Moore-Penrose pseudoinverse of X'X, $(X'X)^{+}$, should be used (ref. 6) to obtain the solution

$$b^{+} = (X'X)^{+}X'Y$$

Of primary concern in this paper are the cases where eq. (8) only approximates zero. When this occurs, we say that multicollinearities exist among the regressor variables. Multicollinearity is explained in greater detail in references 7, 8, and 9.

Strong multicollinearities among the regressor variables produce the following problem with OLS estimation of the regression coefficients:

a. The estimates tend to be large in magnitude.

b. The signs of the estimates are greatly influenced by the multicollinearity, which can result in estimates having signs which disagree with known theoretical (agricultural) properties of the model.

c. Variances and covariances of the estimators tend to be extremely large, often causing the experimenter to delete variables incorrectly.

d. The coefficient estimates are very sensitive to the particular set of sample data, therefore the addition of a few more observations can cause large changes in the estimates.

These problems are due entirely to the presence of multicollinearities and occur regardless of the true values of the regression coefficients. The difficulty centers around the fact that multicollinearities among the regressor variables cause $X'X$ to be nearly singular. This, in turn, creates large values among the elements of $(X'X)^{-1}$.

To illustrate these properties, suppose a linear relationship of the form shown in eq. (8) holds for the first $k \leq p$ regressor variables with a_j nonzero. The diagonal elements of

$$C = (X'X)^{-1}$$

can be expressed as

$$C_{jj} = \left(1 - R_j^2\right)^{-1} \quad ; \quad j = 1, 2, \cdots, p$$

where R_j^2 is the coefficient of determination of the least-squares regression of x_j on the remaining $p - 1$ regressor variables. If $j \leq p$, x_j is involved in the multicollinearity and hence could be well estimated by the remaining regressor variables. This results in an R_j^2 which is very close to 1 and consequently a C_{jj} which is very large. Since

$$\mathrm{Var}(b_{jj}) = C_{jj}\sigma^2$$

the variance of the estimator of the regression coefficient of x_j is very large. The off-diagonal elements of $(X'X)^{-1}$ can be represented as

$$C_{ij} = -S_{ij.(p-2)} \Big/ \left\{ \left[1 - R_i^2 \right] \left[1 - R_{j.(p-2)}^2 \right] \right\}$$

where $S_{ij.(p-2)}$ is the partial covariance of x_i and x_j adjusted for the remaining $p - 2$ variables, and $R_{j.(p-2)}^2$ is the coefficient of determination for the regression of x_j on the remaining $p - 2$ variables, excluding x_i. Thus, if x_i and x_j are both involved in the multicollinearity, R_i^2 will be close to 1 while $S_{ij.(p-2)}$ generally will not be close to zero. Therefore,

$$\text{Cov}(b_i, b_j) = C_{ij}\sigma^2$$

typically will be large in magnitude.

The least-squares estimator of the individual regression coefficient can be written as

$$b_i = \sum_{j=1}^{p} C_{ij}(\underline{X}_j'\underline{y}) \qquad i = 1, 2, \cdots, p$$

Hence, if x_i is one of the variables involved in the multicollinearity, several of the C_{ij} will tend to be large in magnitude, in turn yielding a β_i which is large in magnitude. This is due primarily to the multicollinearity and does not necessarily reflect the true values of the regression parameters β_i.

If we let $\ell_1 \leq \ell_2 \leq \cdots \leq \ell_p$ be the latent roots or eigenvalues of $X'X$ as defined by the equation

$$|X'X - \ell_j I_p| = 0$$

and let $\underline{V}_1, \underline{V}_2, \cdots, \underline{V}_p$ be corresponding latent vectors or eigenvectors of $X'X$ as defined by the equation

$$(X'X)\underline{V}_j = \ell_j \underline{V}_j$$

subject to the constraints

$$\underline{V}_j'\underline{V}_j = 1$$

and

$$\underline{V}_i'\underline{V}_j = 0 \quad ; \quad i \neq j$$

then we can write

$$C = (X'X)^{-1} = \sum_{j=1}^{p} \ell_j^{-1} \underline{V}_j \underline{V}_j' \tag{9}$$

Equation (9) provides another way of illustrating the problems with least-squares estimation. The presence of multicollinearities means that $X'X$ will be near singular and hence one or more of the eigenvalues, ℓ_j, will be near zero. This creates the large elements in $(X'X)^{-1}$ mentioned above.

The problems associated with least-squares estimation motivate the need for alternative methods of estimation and analysis when confronted with multicollinear data. Several recently proposed alternatives are outlined in the next sections. Of necessity, all are biased estimators, but each will be seen to have several desirable as well as undesirable properties.

3. CHOOSING THE BEST REGRESSION

A major problem in regression analysis is that of deciding which regressor or predictor variables should be in the model. There are two conflicting criteria for selecting a subset of regressors. First, the model chosen should include as many of the X's as possible if reliable predictions are to be obtained from the fitted equation. Second, as discussed in section 2, the variance of the predictor increases with the number of regressors. A suitable compromise between these two extremes is usually called "selecting the best subset" or "selecting the best regression equation."

The CCEA model contains 23 predictor or regressor weather-related variables. This author was requested not to consider square or cross-product terms; such analysis should be done as part of the follow-on to this study.

It is recognized that individually the weather variables in the CCEA model provide little information but collectively they do reasonably well. Under these circumstances, it has been shown (ref. 10) that the all-subsets approach is much better than backward, forward, or stepwise regression when selecting a suitable subset of regressor variables.

3.1 ALL POSSIBLE REGRESSIONS

Algorithms have been described (refs. 11 and 12) for computing all possible regressions which are much superior to the naive approach involving the direct inversion of the moments matrix associated with each subset of independent variables. The number of operations per regression decreases from kp^3 to kp^2. If less output for each regression is satisfactory, further savings are possible. By computing the regression coefficients, their variances, and the residual sum of squares with a number of operations per regression, which is of order p, and if we are satisfied with only the residual sum of squares (RSS), the number of operations per regression can be reduced to slightly less than six (ref. 13). As there are two possibilities for each regressor, "in" or "out" of the equation, there are 2^p such regressions.

A systematic procedure for generating all possible regressions is given in references 11, 12, and 14. Garside (refs. 11 and 14) represents each regression by a K-digit binary number; for example, if K = 4, the binary code 1010 would represent the model $E[Y] = \beta_0 + \beta_1 X_1 + \beta_3 X_3$. For K = 3, we have 000, 100, 110, 010, 011, 111, 101, 001. These are the coordinates of the vertices of a K-dimensional hypercube; finding an efficient procedure is equivalent to finding a path along the edges of the hypercube which will pass through each vertex only once (a Hamiltonian walk).

3.1.1 FURNIVAL'S METHOD OF GENERATION

A Gaussian elimination method given by Furnival (refs. 13 and 15) is best described in terms of a "regression tree," as shown in figure 1. The Gaussian elimination operator is applied to each pivotal element just once in the order given by the binary tree. The full matrix is at the root of the tree, and at each interior node a submatrix is derived from the parent matrix by a series of pivots (solid lines) and deletions (dashed lines).

The regression tree can be traversed in any "biologically feasible" order; the only restraint is that a father be "born" before his son. By using horizontal, vertical, or hybrid searching techniques, Furnival obtains a number of regression sequences which he describes as natural, lexicographic, binary, and familial.

3.1.2 BASIC ASSUMPTION

A number of authors have described procedures for finding the best subset regressions without computing all possible regressions (refs. 16, 17, and 18). All of these methods are based on the fundamental inequality

$$RSS(A) \leq RSS(B)$$

where A is any set of independent variables; B is a subset of A; and RSS is root sum square. (See ref. 15 for more details.)

11

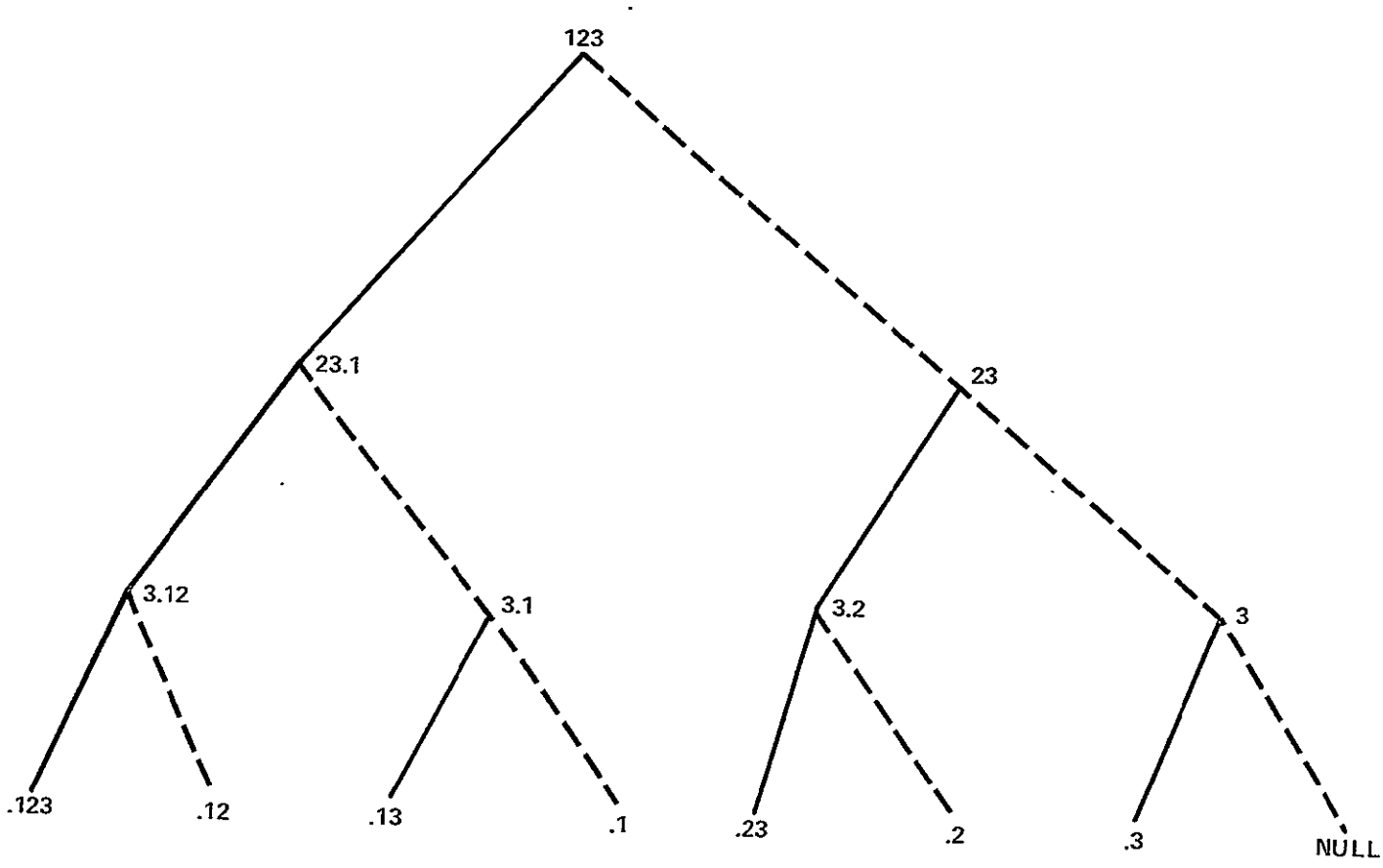


Figure 1.— The regression tree.

3.2 ADJUSTED R^2

One measure of goodness of fit of a regression model widely used in the past is the coefficient of determination

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

Since introducing an extra regressor increases R^2, the problem is not finding the subset with maximum R^2 (which in any case is the set of all p regressors) but rather that of finding a suitable subset with a high R^2.

The adjusted or corrected R^2 statistic is given by

$$\bar{R}^2_m = 1 - \left[1 - R^2_m\right]\left[\frac{n}{n - m}\right]$$

where m is the number of parameters in the model.

To see the effect on $\bar{R}^2$ of adding extra regressors to the equation, consider the F-statistic for testing the significance of q new additions (ref. 2):

$$F = \frac{R^2_{m+q} - R^2_m}{1 - R^2_{m+q}} \times \frac{n - m - q}{q}$$

It follows that

$$\bar{R}^2_{m+q} \geq \bar{R}^2_m$$

if and only if $F \geq 1$.

One criterion, therefore, for selecting the best regression is to choose the regression subset which maximizes $\bar{R}^2_m$.

3.3 MALLOW'S C_p STATISTIC

Consider a q-parameter model, $q < p$. If $N_q = E[\hat{Y}_q]$, then N_q will generally differ from $x'_q\beta_q$ because of possible bias in the q-parameter model.

Let $\theta = E[Y]$. Then, for a given future data point $\underline{X}$,

$$E\left[(\hat{Y}_q - \theta)^2\right] = Var\left[\hat{Y}_q\right] + (N_q - \theta)^2$$

$$= \sigma^2 \underline{x}_q'(\underline{x}_q'\underline{x}_q)^{-1}\underline{x}_q + (N_q - \theta)^2$$

As pointed out in reference 19, it is perhaps more appropriate to use the sum or the average, in some sense, over the future observations of interest.

Mallows (refs. 20-22) and others (ref. 23) suggested to minimize

$$\Delta_q = \frac{1}{\sigma^2} E\left[\sum_{i=1}^{n}(Y_{qi} - \theta_i)^2\right]$$

$$= q + \frac{SSB_q}{\sigma^2}$$

where SSB is the bias sum of squares, given by

$$SSB_q = \sum_{i=1}^{n} (N_{qi} - \theta_i)^2$$

It can be shown (ref. 20) that

$$C_q = \frac{RSS_q}{\hat{\sigma}^2} + 2q - n$$

is a suitable estimate of Δ_q (fig. 2).

Notice that adding $S = q_2 - q_1$ regressors to a model may reduce the bias term SSB, but at the expense of increasing the variance term from $q_1\sigma^2$ to $q_2\sigma^2$. If the equation is needed for prediction, it may be better to drop a few regressors and accept some bias in exchange for a smaller Δ_q and a simpler equation.

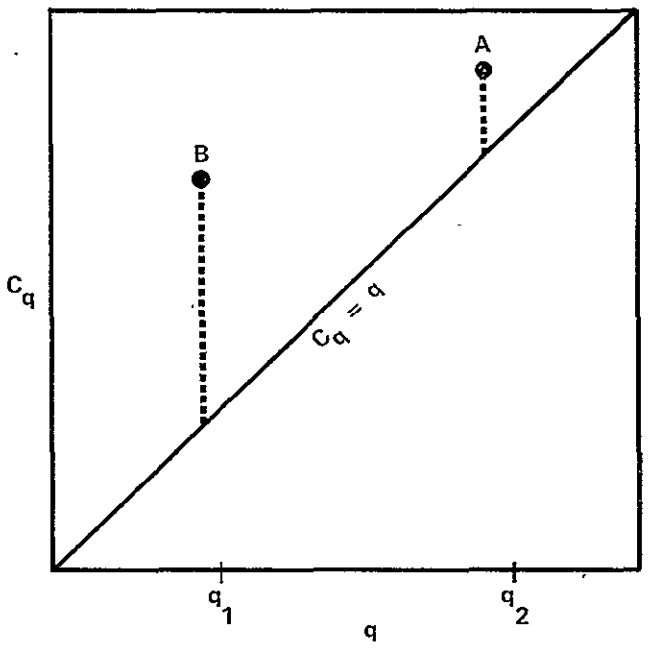


Figure 2.– C_p plot.

It can be shown (ref. 24) that

$$C_q = \frac{RSS_q(n - p - 1)}{RSS_{p+1}} - n + 2q$$

and

$$\frac{C_q - q}{n - q} = \frac{1 - \bar{R}_q^2}{1 - \bar{R}_{p+1}^2} - 1$$

4. ALTERNATIVES TO ORDINARY LEAST-SQUARES ESTIMATION

Some attention has recently been given to two aspects of regression analysis. The first aspect is the attempted improvement of point estimators where the criterion of goodness is the mean-square error (refs. 25-27). Sclove (ref. 27) discusses an estimation technique which guarantees that the sum of component-wise mean-square errors of the biased estimator is smaller than that of the ordinary unbiased least-squares estimator. He presents some further results under the restrictive condition that the independent variable of the model can be ordered in importance prior to analysis. These procedures can be somewhat difficult to implement, and very little is known about the distributional properties of the resulting estimators.

The second aspect of regression considered recently is the problem of point estimation where there is a high degree of multicollinearity among the predictor variables (refs. 28-33). Hoerl and Kennard (ref. 28) propose a class of biased estimators called ridge estimators; their criterion of goodness is mean-square error. The technique is relatively easy to use, and it may be shown that the class contains estimators which have smaller mean-square error than the least-squares estimator. However, they are not able to provide a well-defined and unique choice of estimators from this class, nor have they been able to prove that their suggested procedure actually chooses a member of the class which achieves smaller mean-square error. In fact, Newhouse and Oman (ref. 34) have reported some Monte Carlo simulation results which indicate that ridge estimators do not in general perform better than least-squares estimators.

LaMotte (ref. 35) presents some of the properties of best and Bayes linear estimators. The best estimator follows.

Let L_0 be an $n \times p$ matrix, then the linear estimator L_0y of β is best at (σ_0, β_0) if there exists a (σ_0, β_0) such that for any L,

$$TMSE_{L_0}(\sigma_0, \beta_0) \leq TMSE_{L}(\sigma_0, \beta_0)$$

where

$$TMSE_L(\sigma, \beta) = E[(L'y - \beta)'(L'y - \beta)]$$

$$= \text{trace}[\sigma^2 L'L + (X'L - I)'\beta\beta'(X'L - I)]$$

The Bayes linear estimator is a linear estimator with minimum average total mean-square error (TMSE), averaged over values of (σ, β). Basically, one substitutes $\gamma = E[\sigma^2]$ and $\phi = E[\beta\beta']$ in the above equation for the TMSE obtaining

$$E[TMSE_L(\sigma, \beta) | \gamma, \phi] = ETMSE_L(\gamma, \phi)$$

$$= \text{trace}[\gamma L'L + (X'L - I)'\phi(X'L - I)]$$

Marquardt (ref. 36) introduces a class of biased estimators called generalized inverse estimators. Mayer and Willke (ref. 30) discuss a number of classes of biased estimators called shrunken estimators. These classes contain members with smaller mean-square error than the least-squares estimator. It is not known how to choose such members, however; and there is very little known about the distributional properties of these estimators. Kendall (ref. 31) and Massy (ref. 32) discuss principal components regression, which was not introduced as a method of biased estimation but will be shown to provide biased estimators. The method is very closely related to Marquardt's and was introduced for use when there is multicollinearity.

In this chapter a method of unifying the treatment of these biased estimation methods and of unbiased least-squares estimation is considered. The presentation centers on a duality of the X'X matrix of the normal equations for unbiased least-squares estimation. The duality is in the sense that the spectral decomposition of X'X into its eigenspace representation has the property of describing how well the data points are spread out in the data space. A similar decomposition of $(X'X)^{-1}$ (or a generalized inverse of X'X if it is singular) has the property of describing how the distribution of the estimator b is spread out in the parameter space. We lean heavily on these decompositions to discuss the interrelationships of all these estimation methods and to describe the consequences of using them.

The near-singular case of X'X will be dealt with in this section. A measure of ill-conditioning of the X-matrix is its condition number C[X] which is defined as the ratio of the largest to the smallest nonzero singular value of X. The singular values of X are the positive square roots of the eigenvalues of X'X.

A more precise definition of ill-conditioning follows. A set of linear equations BX - c is said to be ill-conditioned if small errors or variations in the elements of B and c can have large effects on the exact solution X. For example, the difference dX between the solution of BX = c and that of

$$(B + dB)(X + dX) = c + dc$$

can be expressed as

$$dX = (B + dB)^{-1}(dc - dBX)$$

and its value depends critically on the inverse matrix. If B is near singular, that is, small changes in its elements can cause singularity, then dX could be very large. In the case of the normal least-squares equation, B = X'X and c = X'Y will contain roundoff errors because they must be computed from X and Y. Even if B could be computed exactly, it would not necessarily be stored exactly in the computer; all numbers are stored in binary mode, and a decimal number such as 0.1 is a nonterminating binary fraction. If X is ill-conditioned, small changes in the elements of X can cause large changes in $(X'X)^{-1}$; and if $b = (X'X)^{-1}X'Y$, then any errors in the formation of X'X could have a serious effect on the stability and accuracy of the solution. As an illustration, Searle (ref. 37) discusses a model for the weights of six rubber plants, three of which are normal, two of which are off-type, and one of which is an aberrant. The data are presented in the following matrices. The model considered is

$$y_{ij} = \mu + b_1X_1 + b_2X_2 + b_3X_3 + \varepsilon$$

where $X_i = 1$ if the plant is of the i*th* type; otherwise, $X_i = 0$.

Let

$$y = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \end{bmatrix} = \begin{bmatrix} 101 \\ 105 \\ 94 \\ 84 \\ 88 \\ 32 \end{bmatrix} \quad ; \; b = \begin{bmatrix} \mu \\ b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

In this example, we have

$$X'X = \begin{bmatrix} 6 & 3 & 2 & 1 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

and it is seen that X'X is singular and of rank 3. A generalized inverse of X'X is

$$G = (X'X)^{-} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and for this choice of G we have

$$H = GX'X = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

Hence, all estimable functions are of the form

$$w'Hb = (w_1 + w_2 + w_3)\mu + w_1b_1 + w_2b_2 + w_3b_3$$

and there are, at most, three linearly independent choices of w. The unbiased estimates of w'Hb are given by

$$w'GX'Y = w_1\bar{y}_{1.} + w_2\bar{y}_{2.} + w_3\bar{y}_{3.}$$

Three reasonable choices for independent estimable functions are:

$$b_1 - b_2$$

$$b_2 - b_3$$

and

$$\mu + 1/3(b_1 + b_2 + b_3)$$

Their corresponding estimators are

$$\bar{y}_{1.} - \bar{y}_{2.} = 14$$

$$\bar{y}_{2.} - \bar{y}_{3.} = 54$$

and

$$1/3(\bar{y}_{1.} + \bar{y}_{2.} + \bar{y}_{3.}) = 72\ 2/3$$

To consider what problems arise as the gap is slowly bridged from X'X nonsingular to singular, modify the previous example, barely removing it from singular setting, and perform a regression analysis. Performing an experiment to study the abrasion resistance of rubber as a function of the amount of three particular additives, let x_i denote the amount of pounds of the *ith* additive which is loaded with an approximately 1000-pound charge to the chemical reactor which produces the rubber. The proposed model is

$$y = X\beta + \varepsilon$$

where

$$X = \begin{bmatrix} 1 & 0.99 & 0 & 0 \\ 1 & 1.00 & 0 & 0 \\ 1 & 1.01 & 0 & 0 \\ 1 & 0 & 0.99 & 0 \\ 1 & 0 & 1.01 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

$$y' = [101, 105, 94, 84, 88, 32]$$

$$\beta' = [\mu, \beta_1, \beta_2, \beta_3]$$

and e is the random error. In this example,

$$X'X = \begin{bmatrix} 6 & 3 & 2 & 1 \\ 3 & 3.0002 & 0 & 0 \\ 2 & 0 & 2.0002 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

and it is evident that X'X is not singular, yet it is nearly so. Let ℓ_i denote the eigenvalues of X'X. For the matrix X'X above

$$\ell_1 = 8.41888$$
$$\ell_2 = 2.38695$$
$$\ell_3 = 1.19444$$
$$\ell_4 = 0.00010$$

Since $\ell_4 \approx 0.0$, X'X is nearly singular. The parameter estimates

$$\hat{\mu} = 168.002$$
$$b_1 = -68.0212$$
$$b_2 = -81.9742$$
$$b_3 = -136.002$$

were obtained with a covariance matrix of $(X'X)^{-1}\sigma^2$, where

$$(X'X)^{-1} = \begin{bmatrix} 2500.38 & -2500.21 & -2500.13 & -2500.38 \\ -2500.21 & 2500.38 & 2499.96 & 2500.21 \\ -2500.13 & & 2500.38 & 2500.13 \\ -2500.38 & & & 2501.38 \end{bmatrix}$$

It is evident that X'X is formally of rank four although it is essentially of rank three and that the resulting $(X'X)^{-1}$ matrix indicates a large variance in the parameter estimates. However, recall the estimable functions discussed in section 3.3, and compute

$$\left.\begin{aligned} b_1 - b_2 &= 13.9530 \\ b_2 - b_3 &= 54.0278 \\ \hat{\mu} + 1/3(b_1 + b_2 + b_3) &= 72.6695 \end{aligned}\right\} \qquad (10)$$

Allowing for the fact that X was slightly changed to provide nonsingularity, the agreement is admirable. Although the variances of the raw estimates are quite large, consider the variances of the linear combinations of parameters in eq. (10). The linear combinations are defined by K'b where

$$K = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 1/3 \\ -1 & 1 & 1/3 \\ 0 & -1 & 1/3 \end{bmatrix}$$

The covariance matrix of K'b is given by $K'(X'X)^{-1}K\sigma^2$. But

$$K'(X'X)^{-1}K = \begin{bmatrix} 0.82 & -0.33 & 0.05 \\ -0.33 & 1.50 & 0.17 \\ 0.05 & -0.17 & 0.53 \end{bmatrix}$$

It is thus evident that, even though the full parameter vector is quite ill-determined, the linear combinations of the parameters corresponding to the estimable functions of the previous example are well determined.

The normal equations matrix X'X plays the central role in linear model estimation and hypothesis testing. Note that X'X has a spectral decomposition (ref. 2) or representation as

$$X'X = \sum_{i=1}^{p} \ell_i V_i V_i' \tag{11}$$

where $\ell_1 \geq \cdots \geq \ell_p \geq 0$ are the eigenvalues of X'X and S_i are corresponding normalized eigenvectors of X'X. If r is the rank of X'X, then a similar decomposition (ref. 2) of $(X'X)^+$ is

$$(X'X)_r^{-1} = \sum_{i=1}^{r} \frac{1}{\ell_i} V_i V_i' \tag{12}$$

If r = p, then

$$(X'X)_p^{+} = (X'X)^{-1} = \sum_{i=1}^{p} \frac{1}{\ell_i} V_i V_i'$$

It is extremely important to note that the spectral representations of eqs. (11) and (12) are not invariant under linear transformations of X. Invariance may be attained by assuming the linear model is always considered in its correlation form. In the remainder of this report, it will be assumed that all diagonal elements of X'X equal unity. The spectral decomposition of X'X by eq. (11) indicates how and how well the variables' space is spanned by the experiment. If $\ell_i = 1.0$ for all i, then in a sense the variables' space is perfectly spanned. If $\ell_1 >> \ell_p$, then the variables' space is not well spanned. In fact, X_0V_1 represents the linear subspace (or linear combination) of predictor variables which is best spanned, and X_0V_p represents the linear combination of variables most poorly spanned. In fact, if $\ell_p = 0$, then X_0V_p is not spanned at all. These considerations are discussed by Kendall and Stuart (ref. 38). To illustrate the preceding paragraphs, consider the following two-dimensional example. Suppose the data points observed are as plotted on figure 3.

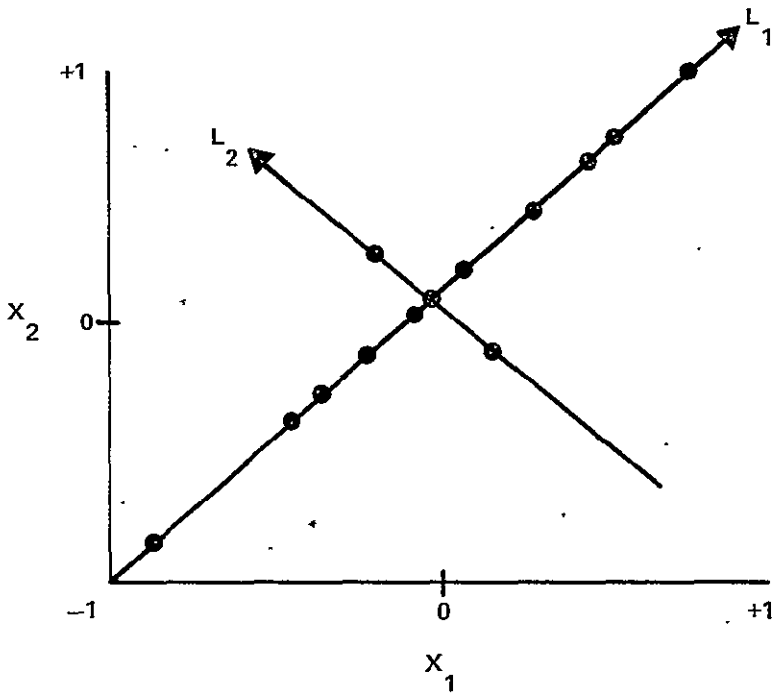


Figure 3.— Two-dimensional example.

Assume that the line L_1 is the line $X_1 - X_2 = 0$ in the variables' space and that L_2 is the line $X_1 + X_2 = 0$. Assume also that the two extreme points along L_2 are equally distant from $X_1 - X_2 = 0$. It is immediately seen that the observations are much more spread out along L_1 than along L_2. For such a situation, $\ell_1 > \ell_2$, and X_0V_1 is well spanned, while X_0V_2 is poorly spanned.

Considering the parameter space, it is well known that the least-squares estimator b (under normal distribution theory) follows the normal distribution $N[(X'X)^+(X'X)b, \sigma^2(X'X)_r^+]$. For a linear combination of the estimates $w'\hat{b}$ (ref. 37):

$$\mathrm{Var}(w'b) = w'(X'X)_r^+ w\sigma^2$$

It can be shown (ref. 39) that the choice of w which minimizes the variance of w'b is $w = S_1$ and that this variance is

$$\mathrm{Var}(V_1'b) = V_1'(X'X)^+V_1\sigma^2 = \frac{\sigma^2}{\ell_1}$$

The choice of w which maximizes the variance of w'b is $w = V_p$ and

$$\mathrm{Var}(V_p'b) = \frac{\sigma^2}{\ell_p} \text{ (assuming } p = r)$$

Thus, $V_1'b$ describes the most determined linear combination of the parameters, while $V_p'\hat{b}$ describes the least determined. In fact, if $\ell_p = 0$, then $V_p'b$ is nonestimable and hence not determined at all; an interpretation is that $V_p'b$ has infinite variance.

4.1 RIDGE AND GENERALIZED RIDGE

One recently proposed alternative to least-squares estimation of the regression parameters which has received considerable attention in the statistical literature is ridge regression (refs. 28 and 29). When multicollinearities exist among the regressor variables, the least-squares estimates b_j tend to be large in magnitude. This can result in b being far removed from β, in

terms of Euclidean distance, even though b is an unbiased estimator of β. If L_1 denotes the distance from b to β (ref. 28)

$$E[L_1^2] = \sigma^2 \sum_{i=1}^{p} \frac{1}{\ell_1}$$

where $\ell_1 \leq \ell_2 \leq \cdots \leq \ell_p$ are the eigenvalues of X'X as before. As pointed out in the previous section, the existence of multicollinearities means that X'X will have one or more small eigenvalues, thus the distance from b to β will generally be large.

Reviewing some of the most important properties of the ridge estimator, recall that the best linear unbiased estimator of β is

$$b = (X'X)^{-1}X'Y$$

Then X'X may be represented as X'X = PLP', where P is the orthogonal matrix whose columns are the normalized eigenvectors of X'X and L is the diagonal matrix of eigenvalues. If one considers the transformation to new predictor variables defined by

$$W = XP$$

and the model

$$y = Wa + e \tag{13}$$

then

$$a = P'b$$

$$W'W = L$$

$$a'a = b'b$$

The generalized ridge estimation procedure, defined by the family of estimators indexed by the parameters $k_i \geq 0$, is

$$a^* = (W'W + K)^{-1}W'y \tag{14}$$

where the matrix K is defined by

$$K = \text{diag}\{k_i\} \quad ; \quad i = 1, 2, \cdots, p$$

When all $k_i = 0$, a* is the OLS estimator and is unbiased. When any $k_i > 0$, the resulting estimator for a is biased, defining the mean-square error of a* as

$$M(K) = E[(a^* - a)'(a^* - a)]$$

It may be shown (ref. 28) that the choice of $k_i = \sigma^2/a_i^2$ will minimize M(K) among the class of estimators defined by eq. (14). Unfortunately, in order to utilize this optimal choice of k_i, one must know both σ^2 and a_i^2. To circumvent this seemingly hopeless situation, one must resort to the following iterative procedure (ref. 40).

1. Using OLS procedures on the canonical model, eq. (13), estimate the a_j's by computing

$$\hat{a} = (X'X)^{-1}Xy$$

and estimate σ^2 by s^2.

2. Use the value of s^2 and the $\hat{a}_j$'s from step 1 to compute

$$k_j = \frac{s^2}{\hat{a}_j^2} \quad ; \quad j = 1, 2, \cdots, p$$

3. Use the k_j's to solve the expression

$$a^* = (W'W + K)^{-1}W'y$$

and thus obtain initial estimates of the a_j^*'s. Next compute

$$a^{*\prime}a^* = \sum_{j=1}^{p} a_j^{*2}$$

4. Repeat steps 2 and 3 using the a_j^*'s from step 3 and again compute $a^{*\prime}a^*$.

5. Continue this iterative procedure and terminate when stability is achieved in $a^{*\prime}a^{*}$.

6. The generalized ridge regression coefficients are now given by

$$b = P'a^{*}$$

The above procedure has some intuitive appeal, but since the distributional properties of the resultant estimator are unknown, its validity as a statistical tool is subject to questioning.

Simulations (refs. 41, 42, and 43) have shown that ridge estimators of the form of eq. (14) do provide smaller mean-square errors than the OLS. In fact, reference 28 shows that if $k^{*} > 0$ and if $k_i = k^{*}$ $(i = 1, 2, \ldots, p)$ and hence

$$a^{*} = (W'W + k^{*}I_p)^{-1}X'y \tag{15}$$

then there exists a $k > 0$ such that the mean-square error $(M(k))$ of a^{*} is less than the mean-square error $M(0)$ of the least-squares estimator b, where

$$M(0) = E[(\beta - b)'(\beta - b)]$$

In practice, one must estimate k from the data. The properties of the estimator a^{*}, when k is estimated from the data, are unknown. An optimal method for selecting a suitable value of k has been the center of much recent discussion in the statistical literature. Various methods have been proposed (refs. 28, 29, 36, 41, 44, 45, and others). An excellent discussion of ridge regression and the various methods for choosing k is contained in reference 46.

Expressing the ridge estimator eq. (15) as

$$b(k) = (X'X + kI_p)^{-1}X'y \tag{16}$$

where $k \geq 0$ is nonstochastic.[1]

[1] In the literature, eq. (16) is referred to as the ordinary ridge estimator.

Obviously, if k = 0, eq. (16) is the OLS estimator; i.e.,

$$b(0) = (X'X)^{-1}X'y$$

It has been shown (ref. 28) that

$$M(k) = \sigma^2 \sum_{i=1}^{p} \frac{\ell_i}{(\ell_i + k)^2} + k^2 b'(X'X + kI)^{-2} b$$

and that $M'(0) < 0$. Observe the effect of k on the quantities $V_i'b(k)$. The equality

$$\begin{aligned} V_i'b(k) &= V_i'(X'X + kI)^{-1}X'y \\ &= V_i'\left(\sum_j \frac{1}{\ell_j + k} V_j V_j'\right) X'y \\ &= \frac{1}{\ell_i + k} V_i'X'y \end{aligned}$$

is immediately obtained; also

$$\begin{aligned} E\left[V_i'b(k)\right] &= \frac{1}{\ell_i + k} V_i'X'E[y] \\ &= \frac{\ell_i}{\ell_i + k} V_i'b \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}\left[V_i'b(k)\right] &= \frac{1}{(\ell_i + k)^2} V_i'X'(\sigma^2 I)XV_i \\ &= \frac{\ell_i \sigma^2}{(\ell_i + k)^2} \end{aligned}$$

Thus for any nonzero k, $V_1'b(k)$ is the least biased linear combination of the estimator and $V_p'b(k)$ is the most biased. Also, $V_1'b(k)$ has the least reduced variance, and $V_p'b(k)$ has the most reduced variance. Thus, the best determined linear combinations of the parameter estimates are the least modified, while

the least determined are the most modified. In effect, as k increases, those $V_i'b(k)$ corresponding to small ℓ_i are rapidly driven to zero.

Recall that the predicted regression function at X_0 is

$$y_0 = X_0 b(k)$$

and the mean-square error of this predicted regression is denoted by

$$\begin{aligned} M_p[y_0|b(k)] &= E\{[y_0 - X_0 b(k)]'[y_0 - X_0 b(k)]\} \\ &= E\{[X_0(X'X + kI)^{-1}X'\varepsilon + X_0(Z - I)b]'[X_0(X'X + kI)^{-1}X'\varepsilon \\ &\quad + X_0(Z - I)b]\} \\ &= E[\varepsilon'X(X'X + kI)^{-1}X_0'X_0(X'X + kI)^{-1}X'\varepsilon] + b'(Z - I)X_0'X_0(Z - I) \\ &= \gamma_1[b(k)] + \gamma_2[b(k)] \end{aligned}$$

where $\gamma_1[b(k)]$ corresponds to the variance and $\gamma_2[b(k)]$ corresponds to the bias squared.

THEOREM 1:

The variance function $\gamma_1[b(k)]$ is a monotonically decreasing function of k and $\gamma'[b(0)] < 0$ (ref. 47).

PROOF:

Note that if we assume $\varepsilon \sim N(0, \sigma^2 I)$ then γ_1 is the expectation of a quadratic form in assumption e, thus (ref. 37)

$$\gamma_1[b(k)] = \sigma^2 tr[X(X'X + kI)^{-1}X_0'X_0(X'X + kI)^{-1}X']$$

Note that

$$(X'X + kI)^{-1} = \sum_i \frac{1}{\ell_i + k} V_i V_i'$$

and hence

$$X(X'X + kI)^{-1}X_0 = \sum_i \frac{1}{\ell_i + k} XV_iV_i'X_0'$$

which is $n \times 1$. Thus,

$$\frac{\gamma_1[b(k)]}{\sigma^2} = tr\left[\left(\sum_i \frac{1}{\ell_i + k} XV_iV_i'X_0'\right)'\left(\sum_j \frac{1}{\ell_j + k} X_0V_jV_j'X\right)\right]$$

$$= \sum_i \sum_j \frac{1}{(\ell_i + k)(\ell_j + k)} tr\left[\left(XV_jV_j'X_0'\right)\left(X_0V_iV_i'X'\right)\right]$$

$$= \sum_i \sum_j \frac{1}{(\ell_i + k)(\ell_j + k)} \left(X_0V_iV_i'X'\right)\left(XV_jV_j'X_0\right)$$

$$= \sum_i \sum_j \frac{1}{(\ell_i + k)(\ell_j + k)} X_0\left[V_iV_i'\left(\sum_m \ell_mV_mV_m'\right)V_jV_j'\right]X_0$$

$$= \sum_i \frac{\ell_i}{(\ell_i + k)^2} X_0V_iV_i'X_0'$$

Note that

$$\gamma_1[b(0)] = \sigma^2X_0(X'X)^{-1}X_0'$$

$$\gamma_1[b(0)] = 0$$

and

$$\gamma_1'[b(k)] = -\sigma^2 \sum_i \frac{2\ell_i}{(\ell_i + k)^3} X_0V_iV_iX_0' < 0$$

Thus, γ_1' is a monotonically decreasing function of k, as was to be shown.

THEOREM 2:

The bias function $\gamma_2[b(k)]$ satisfies $\gamma_2[b(0)] = 0$ and $\gamma_2'[b(0)] = 0$.

PROOF:

Recall $\gamma_2[b(k)] = b'(Z - I)X_0'X_0(Z - I)b$. Since

$$Z - I = (X'X + kI)^{-1}X'X - I$$

$$= \sum_i \frac{-k}{\ell_i + k} V_i V_i'$$

we obtain

$$\gamma_2[b(k)] = \sum_i \sum_j \frac{k^2}{(\ell_i + k)(\ell_j + k)} b'V_i V_i' X_0'X_0 V_j V_j' b$$

Let

$$f_{ij}(k) = \frac{k^2}{(\ell_i + k)(\ell_j + k)}$$

then

$$f_{ij}'(k) = \frac{k^2(\ell_i + \ell_j) + 2k\ell_i\ell_j}{(\ell_i + k)^2(\ell_j + k)^2} \geq 0$$

Thus, it is easily seen that $\gamma[b(0)] = 0$ and $\gamma_2'[b(0)] = 0$.

One of the most important results of the ordinary ridge regression is the following theorem.

THEOREM 3:

$M_p[y_0|b(k)]$ is initially decreasing in k.

PROOF:

Since $M_p[y_0|b(k)] = \gamma_1[b(k)] + \gamma_2[b(k)]$ the result follows directly from theorems 1 and 2.

Hoerl and Kennard (ref. 28) discuss a general Bayesian interpretation for the ridge estimator. Marquardt (ref. 36) gives a more specific relationship to Bayesian estimation, as follows.

THEOREM 4:

The ridge estimator is equivalent to a least-squares estimator when the actual data are supplemented by a fictitious set of data points taken according to an orthogonal experiment H_k; the response y is set to zero for each of these supplementary data points.

PROOF:

Augmenting the X-matrix by H_k, the least-squares normal equations become

$$\left(X' \vdots H_k'\right)\begin{bmatrix} X \\ \cdots \\ H_k \end{bmatrix} b = \left(X' \vdots H_k'\right)\begin{bmatrix} y \\ \cdots \\ 0 \end{bmatrix}$$

or

$$\left(X'X + H_k'H_k\right)b = X'y \tag{17}$$

Since H_k is orthogonal, $H_k'H_k$ is a scalar multiple of I_p; for any value k, the matrix may always be scaled such that $H_k'H_k = kI_p$, and eq. (17) is identical to eq. (16).

To illustrate, possible choices for H_k are (a) $H_k = k^{1/2}I_p$, or (b) $H_k = 2^p$ factorial experiment with the variables at levels $-\alpha$ and $+\alpha$, where $\alpha = (k2^{-p})^{1/2}$. This theorem illustrates from another viewpoint the mechanism by which the regression coefficients are damped by the ridge estimator. The estimator is seen to be a type of weighted average between the actual data and other data (in Bayesian terms, the prior information) for which the response values are arbitrarily set to zero. (For nonstandardized variables, the response values for the fictitious data would be set equal to the mean response to the actual data if the model $y = X\beta + e$ contains a constant term.)

An excellent paper by Obenchain (ref. 48) develops the theoretical foundations to hypothesis testing and confidence regions for ordinary and generalized ridge regression estimators. He shows that any ridge estimator with strictly positive and nonstochastic "shrinkage factors" $k_1, k_2, \cdots, k_p$ yields the same exact F (or t) statistic for the test of any linear hypothesis as does least squares. It follows that the unbiased confidence region based upon the F (or t) distribution corresponding to any such ridge estimator is identical to the least-squares region of the same confidence.

4.2 MARQUARDT'S GENERALIZED INVERSE ESTIMATOR

Marquardt (ref. 36) discusses a method of applying generalized inverses to biased estimation. He also considers some relations among these estimators, ridge estimators, and nonlinear estimation. He considers the model

$$y = X\beta + e \tag{18}$$

where the X-matrix has been scaled so that X'X is in the correlation form. His family of estimators is indexed by a parameter h where $0 \leq h \leq p$. The family is defined by

$$b_m(h) = (X'X)_h^+ X'Y$$

The matrix $(X'X)_h^+$ is defined as follows: let $h^* = [h]$ denote the greatest integer in h and $dh = h - h^*$. Then $(X'X)_h^+$ is defined as

$$(X'X)_h^+ = \sum_{j=1}^{h^*} \frac{1}{\ell_j} V_j V_j' + \frac{dh}{\ell_{h^*+1}} V_{h^*+1} V_{h^*+1}' \tag{19}$$

$$= G_h$$

As the notation is meant to indicate, $(X'X)_h^+$ is closely related to a generalized inverse of X'X. In fact, if r = rank (X'X), the $(X'X)_r^+$ is the Moore-Penrose pseudoinverse of X'X and is unique (ref. 49). An important point to note is that the Moore-Penrose pseudoinverse yields the minimum-norm solution to the normal equations (ref. 50).

Marquardt's estimators thus provide a sort of minimum norm solution to the normal equations. He also shows that there always exists a $0 < h < p$ such that

$$M(h) = E\left\{[b_m'(h) - b]'[b_m(h) - b]\right\}$$

is minimized. It is also shown that $M'(p) > 0$ so that the mean-square error of $b_m(h)$ is initially decreasing as h decreases from p. As with the ridge estimators, no way is yet developed for determining the "best" h.

With X scaled so that X'X is in correlation form, Marquardt labels the diagonal elements of $(X'X)_h^+$ as variance inflation factors. His suggested analytical procedure is to consider several estimates $b_m(h)$ for h between p and 0. He suggests the rule of thumb that an acceptable value of h is one such that the maximum variance inflation factor should usually be larger than 1.0 but certainly not as large as 10.0. Marquardt has not been able to show that this procedure results in a reduction in M(h).

For these estimators,

$$V_i' b_m(h) = V_i'(X'X)_h^+ X'y = \begin{cases} 0 & h \leq i - 1 \\ \frac{dh}{\ell_i} V_i' X'y & i - 1 < h \leq i \\ \frac{1}{\ell_i} V_i' X'y & i < h \end{cases} \tag{20}$$

$$E\left[V_i' b_m(h)\right] = \begin{cases} 0 & h \leq i - 1 \\ dh V_i' b & i - 1 < h \leq i \\ V_i' b & i < 1 \end{cases} \tag{21}$$

and

$$\mathrm{Var}\left[V_i' b_m(h)\right] = \begin{cases} 0 & h \leq i - 1 \\ \dfrac{(dh)^2 \sigma^2}{\ell_i} & i - 1 < h \leq i \\ \dfrac{\sigma^2}{\ell_i} & i < h \end{cases} \qquad (22)$$

From eqs. (21) and (22), the following behavior is seen as h decreases from h = p: the $V_i' b_m(h)$ are successively set to zero in order of increasing ℓ_i. The best determined linear combinations of the parameter estimates are the last to be set to zero, while the least determined are the first to be set to zero.

THEOREM 5:

The estimate $b_m(h)$ is a linear transform of b, and the transform depends only on X and h.

PROOF:

Let A = X'X, then

$$S'AS = L$$

and

$$A^{+} = SL^{-1}S'$$

Suppose A is of rank r, so that the last (p - r) ordered elements of L are zero (or nearly so, if A is only "nearly singular"). Partition S as follows:

$$S = (S_r \vdots S_{p-r})$$

where S_r is [P × r]; S_{p-r} is [p × (p - r)]. Partition L similarly

$$L = \begin{bmatrix} L_r & \vdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \vdots & L_{p-r} \end{bmatrix}$$

where L_r is $[r \times r]$; L_{p-r} is $[(p - r) \times (p - r)]$.

Now, by supposition, L_{p-r} is zero, so that $L_{p-r}^{-1} = 0$ by definition. Thus, the psuedoinverse becomes

$$A_r^+ = S_r L_r^{-1} S_r'$$

Therefore

$$b_m(h) = S_r L_r^{-1} S_r' X'y$$

but

$$X'y = (X'X)b$$

and thus

$$b_m(h) = S_r L_r^{-1} S_r' (X'X)b$$
$$= Z_r b$$

It follows immediately that $b_m(h)$ is a biased estimator of β, if L_{p-r} is a nonnull matrix. If L_{p-r} is precisely a null matrix, $b_m(h)$ is conditionally unbiased relative to the constraints $E[b_m(h)] = Z_r b$ implied by the columns of S_{p-r}.

THEOREM 6:

The variance of $b_m(h)$ is

$$\text{Var}[b_m(h)] = \sigma^2 \left[S_r L_r^{-1} S_r'\right] (X'X) \left[S_r L_r^{-1} S_r'\right]$$

PROOF:

$$\text{Var}(b) = \sigma^2 (X'X)^{-1}$$

thus,

$$\text{Var}(Z_r b) = \sigma^2 Z_r (X'X)^{-1} Z_r'$$

Substituting Z_r, the result is immediate. It can also be shown that

$$\mathrm{Var}[b_m(h)] = \sigma^2 S_r L_r^{-1} S_r'$$

THEOREM 7:

The mean-square error of b(h) is

$$E\left[L_1^2\right] = \mathrm{tr}[\mathrm{Var}(b_m(h))] + \beta'(Z_r - I)'(Z_r - I)\beta$$

This may be proved in the same manner as in ridge regression (section 4.1). The second term on the right side of $E[L_1^2]$ is the square of the bias; it will be zero when $r = p$.

COROLLARY 1:

The variance term in $E[L_1^2]$ is an increasing function of r.

PROOF:

Employing eq. (19), we have

$$S_r L_r^{-1} S_r' = \sum_{j=1}^{r} \frac{1}{\ell_j} V_j V_j'$$

Hence,

$$\mathrm{tr}\left[S_r L_r^{-1} S_r'\right] = \sum_{j=1}^{r} \frac{1}{\ell_j} \mathrm{tr}[V_j V_j']$$

But $\mathrm{tr}[V_j V_j'] = ||V_j|| = 1.0$, since S is an orthonormal rotation. Thus,

$$\mathrm{tr}[S_r L^{-1} S_r] = \sum_{j=1}^{r} \frac{1}{\ell_j}$$

and

$$\mathrm{tr}\left[\mathrm{Var}\left(b_m(h)\right)\right] = \sigma^2 \left(\sum_{j=1}^{r} \frac{1}{\ell_j}\right) \tag{23}$$

Since $\ell_j \geq 0$ for all j, eq. (23) increases monotonically with r. In the special case where the data are orthogonal, i.e., all $\ell_j = 1.0$, and where $r = p$, this result becomes $p\sigma^2$, as expected.

COROLLARY 2:

The bias term in $E[L_1^2]$ is a monotonic decreasing function of r.

PROOF:

The bias term is

$$(\text{Bias})^2 = \beta'(Z_r - I)'(Z_r - I)\beta$$

$$= \beta'\left[S_r L_r^{-1} S_r(X'X) - I\right]'\left[S_r L_r^{-1} S_r(X'X) - I\right]\beta$$

Partitioning the several matrices in $(Z_r - I)$, and simplifying, results in

$$(Z_r - I) = S_r S_r' - I_p$$

$$= -S_{p-r} S_{p-r}'$$

and therefore,

$$(\text{Bias})^2 = \beta'\left[S_{p-r}\ S_{p-r}'\right]'\left[S_{p-r} S_{p-r}'\right]\beta$$

$$= \beta' S_{p-r} S_{p-r}' S_{p-r} S_{p-r}' \beta$$

But

$$S_{p-r}' S_{p-r} = I_{p-r}$$

and therefore,

$$(\text{Bias})^2 = \beta' S_{p-r} S_{p-r}' \beta$$

Now $T_{p-r} = S_{p-r}'\beta$ is the $(p - r)$-element vector of projections of β onto the subspace spanned by S_{p-r}. In this notation

$$(\text{Bias})^2 = T_{p-r}' T_{p-r}$$

$$= \sum_{j=r+1}^{p} t_j^2 \qquad (24)$$

Since t_j does not depend on r, we have the result that $(\text{Bias})^2$ is a monotonic decreasing function of r. Furthermore, $(\text{Bias})^2$ has the limiting value 0.0 for $r = p$ and the value $\beta'\beta$ for $r = 0$, since $T_p' T_p = \beta'\beta$.

THEOREM 8:

A sufficient condition for the mean-square error $E[L_1^2]$ to be less than the least-squares variance is

$$\sum_{j=r+1}^{p} \frac{1}{\ell_j} > \frac{1}{\sigma^2} (\beta'\beta) \tag{25}$$

PROOF:

Using eqs. (23) and (24), $E[L_1^2]$ is given by

$$E[L_1^2] = \sigma^2\left(\sum_{j=1}^{r} \frac{1}{\ell_j}\right) + \sum_{j=r+1}^{p} t_j^2$$

while the least-squares variance is

$$\text{Var}(b) = \sigma^2\left(\sum_{j=1}^{p} \frac{1}{\ell_j}\right)$$

Thus, a neccesary and sufficient condition for

$$E\left[L_1^2\right] < \text{Var}(b)$$

is that

$$\sigma^2\left(\sum_{j=r+1}^{p} \frac{1}{\ell_j}\right) > \sum_{j=r+1}^{p} t_j^2$$

or

$$\sum_{j=r+1}^{p}\left(\frac{\sigma^2}{\ell_j} - t_j^2\right) > 0$$

Thus, a sufficient condition is that

$$\frac{\sigma^2}{\ell_j} > t_j^2 \quad ; \quad j > r \tag{26}$$

Since eq. (26) would be difficult to apply in practice, a less stringent but more useful sufficient condition can be obtained by noting that

$$\sum_{j=1}^{p} \beta_j^2 \geq \sum_{j=r+1}^{p} t_j^2$$

for any $r \leq p$. Thus, the inequality can be written as in eq. (25). This result corresponds to the following theorem.

THEOREM 9:

If $\beta'\beta$ is bounded, then there exists a $k > 0$ such that the mean-square error of the ordinary ridge regression $b(k)$ is less than the mean-square error of the least-squares estimator (ref. 28). An important theorem due to Marquardt (ref. 36) with some interesting interpretation follows.

THEOREM 10:

Let γ_r be the angle between $b_m(r)$ and $g = X'y$. Then $\gamma_r \geq \gamma_{r-1}$ (r an integer) whenever ℓ_r and ℓ_{r-1} satisfy the inequalities $(0 < \ell_r)$, $(\ell_r/\ell_{r-1} \ll 1)$, and $(\ell_{r-1}/\ell_{r-2} \ll 1)$. Since g is independent of $b_m(r)$, it follows that $b_m(r)$ rotates toward g as r is decreased under these conditions.

4.3 SHRUNKEN ESTIMATORS

Mayer and Willke (ref. 30) discuss several families of biased estimators which may be labeled shrunken estimators. They consider the model $y = X\beta + e$, but do not require that $X'X$ be in correlation form. Each family of estimators is indexed by a parameter c, $0 \leq c \leq 1$ and defined by

$$b_s(c) = c(X'X)^{-1}X'y = cb$$

where b is the ordinary unbiased least-squares estimator. If the constant c is a scalar fixed in advance of the analysis, then $b_s(c)$ is called a deterministically shrunken estimator. If c is a scalar function of the least-squares estimator, then $b_s(c)$ is called a stochastically shrunken estimator.

Hoerl and Kennard (ref. 28) justify the use of the ridge estimator in non-orthogonal problems in two ways: (1) They show that, for a fixed k, b(k) corresponds to the point on a fixed ellipse of concentration of b which has minimum Euclidean length and (2) they show that in any given problem the class of ridge estimators satisfy the following admissibility condition: A class of estimators E will be called (mean square) admissible if for every problem there is an e in E such that $M(e) < M(b) = Var(b)$.

Although the shrunken estimator $b_s(c)$, with shrinkage factor c, may seem a rather simplistic alteration of b, the following proposition proved in reference 30 shows that these satisfy the admissibility condition presented above.

PROPOSITION 1: For every β there exists a fixed c in [0, 1] such that $M[b_s(c)] < M(b)$ and thus the subclass of deterministically shrunken estimators is admissible.

Consider the stochastically shrunken estimator $b_s(c)$, where

$$c = [1 - qS^2(b'b)^{-1}]$$

$$S^2 = y'y - b'(X'X)^{-1}b \quad ; \quad p \geq 3$$

and

$$0 < q < 2(p - 2)(n - p + 2)^{-1}$$

This estimator is one discussed by Sclove (ref. 27). Defining

$$W[b_s(c)] = E\left\{[b_s(c) - b]'[b_s(c) - b]\right\}$$

it was shown by Sclove that

$$q = q_0 = \frac{p - 2}{n - p + 2}$$

minimizes $W[b_s(c)]$. This is the only biased estimator known to this author for which a choice of biased estimator can be explicitly given which guarantees a reduction in mean-square error.

Let C denote the class of linear transforms of b, and let t = Ab for some p × p matrix A. Note that if we let t(A) = Ab for fixed A, then

$$E[t(A)] = A\beta$$

$$Var[t(A)] = \sigma^2 A'S^{-1}A$$

$$M[t(A)] = \sigma^2 trA'S^{-1}A + \beta'(A - I)'(A - I)\beta$$

and the sum-of-squares loss associated with t(A) is

$$\begin{aligned} L(A) &= [y - Xt(A)]'[y - Xt(A)] \\ &= (y - Xb)'(y - Xb) + b'(A - I)'S(A - I)b \\ &= L(b) + L^*(A) \end{aligned}$$

Since $L^*(I) = 0$, L(A) is minimized by letting A = I, which yields the least-squares estimator. However, if $L^*(A) > 0$, then the mapping from the space of p × p matrices to the real line defined by $\gamma(A) = L^*(A)$ maps an entire class of matrices to the same value. The preimage of any fixed constant r_0 consists of all p × p matrices satisfying

$$b'(A - I)'S(A - I)b = r_0$$

Let $C(r_0)$ denote the subclass of C such that $t(A_0)$ is in $C(r_0)$ if and only if $L^*(A_0) = r_0$. $C(r_0)$ is actually an equivalence class, the equivalence being defined with respect to the sum-of-squares loss function. It can be shown that both ridge estimators and the deterministically shrunken estimators can be characterized as minimum normal estimators in the class C (ref. 30). Suppose the criterion for selecting an estimator from an equivalence class is to choose the estimator which has minimum Euclidean length (normal). Let

$$m(A) = t'(A)t(A) = b'A'Ab$$

denote the squared Euclidean length of t(A). Mayer and Willke (ref. 30) have proved two propositions that link the ordinary ridge estimators and the shrunken estimators.

PROPOSITION 2: If $A_0 = (kS + I)^{-1}$ for some k and $t(A_0)$ is in $C(r_0)$, then

$$m(A_0) = \min_{C(r_0)} m(A)$$

This proposition states that within its equivalence class the ridge estimator is the shortest estimator, provided m(A) is the norm used to measure length. Now consider the design dependent norm

$$m_d(A) = t'(A)St(A) = b'A'SAb$$

and suppose the optimal estimator in an equivalence class is defined to be the estimator with minimum length as measured by $m_d(A)$.

PROPOSITION 3: If $A_1 = cI$ for some c in [0, 1] and $b_s(c)$ belongs to $C(r_0)$, then

$$m_d(cI) = \min_{C(r_0)} m_d(A)$$

Since $t(A_1) = cb = b_s(c)$ we have shown that both ridge estimators and the shrunken estimators are minimum length estimators with respect to the appropriate norms.

For the choice of deterministically shrunken estimator $b_s(c) = cb$ we have

$$V_i'b_s(c) = cV_i'b = \frac{c}{\ell_i} V_i'X'Y$$

$$E[V_i'b_s(c)] = cV_i'b \tag{27}$$

and

$$\mathrm{Var}[V_i'b_s(c)] = \frac{c^2\sigma^2}{\ell_i} \tag{28}$$

From eqs. (27) and (28) we observe that, unlike the ridge and generalized inverse estimators, all linear combinations of the parameter estimates are driven toward zero proportionately and that the variances are also proportionately reduced.

The mean-square error of the estimated regression function is given (ref. 47) by

$$M_p[y_0|b_s(c)] = E[(y_0 - X_0b)'(y_0 - X_0b)]$$

$$= E\left\{\left[(c-1)X_0b + cX_0(X'X)^{-1}X'\varepsilon\right]'\left[(c-1)X_0b + cX_0(X'X)^{-1}X'\varepsilon\right]\right\}$$

$$= E\left[c^2\varepsilon'X(X'X)^{-1}X_0'X_0(X'X)^{-1}X'\varepsilon\right] + 2E\left[(c-1)cb'X_0'X_0(X'X)^{-1}X'\varepsilon\right] + E\left[(c-1)^2b'X_0'X_0b\right]$$

For stochastically shrunken estimators, these expectations may be somewhat difficult. For a deterministically shrunken estimator, c is a constant and $M_p[b_s(c)]$ is easily found to be

$$M_p[y_0|b_s(c)] = c^2E\left[\varepsilon'X(X'X)^{-1}X_0'X_0(X'X)^{-1}X'e\right] + (c-1)^2b'X_0'X_0'b$$

$$= \gamma_1[b_s(c)] + \gamma_2[b_s(c)]$$

The following three theorems are due to Sidik (ref. 47).

THEOREM 11:

The variance function $\gamma_1[b_s(c)]$ is a monotonically increasing function of $c > 0$ and $\gamma_1'[b_s(1)] > 0$.

THEOREM 12:

The bias function $\gamma_2[b_s(c)]$ is a monotonically decreasing function of c for $0 \leq c \leq 1$.

THEOREM 13:

$M_p[y_o|b_s(c)]$ is initially decreasing as c decreases from c = 1, and there is a unique minimum for some $0 < c < 1$.

Theorem 13 states that an optimal choice of c exists. However, this optimal value of c will be a function of σ^2 and b.

Several authors (refs. 51 and 52) have considered different ways of unifying the study of biased estimators in an effort to determine their relative merits. Obenchain (ref. 52) has considered the problem of testing whether ridge analysis may be useful. He defines the shrunken statistic that is used to decide if ridge analysis should be used or not.

4.4 PRINCIPAL COMPONENTS REGRESSION

A particular type of Marquardt's generalized inverse estimator is the principal components estimator, which involves an orthogonal reparameterization of the values of the regressor variables through the following procedure.

Let S be the orthogonal matrix whose columns are the eigenvectors of X'X and let L be a diagonal matrix whose diagonal elements are the eigenvalues of X'X. If we also let Z = XS, then the j*th* column of Z, z_j, is called the j*th* principal component of X for j = 1, 2, ···, p.

The response variable is now regressed on the principal components z_j, rather than on the original variables x_j. In place of the usual regression model

$$y = X\beta + e$$

now we have

$$y = Z\gamma + e$$

where

$$\gamma = S'\beta$$

Using least squares, we obtain

$$g = (Z'Z)^{-1}Z'y = L^{-1}Z'y \tag{29}$$

If all components are retained in the model, the estimates of the regression coefficients when transformed from g back to b through b = Sg will be identical to the least-squares estimates.

Use of the procedures discussed above would hardly be necessary when the beta vector could be estimated directly by classical methods. At least two situations arise, however, in which ordinary least-squares is not appropriate (ref. 32): (1) when the independent variables are collinear with one another, making inversion of the correlation matrix impossible and the elements of beta indeterminate; and (2) when, because of high (but not complete) collinearity or for some other reason, it is desirable to collapse the independent variable space by deleting one or more principal components from the regression relationship. We are mostly concerned with the second case.

To overcome the effects of multicollinearity on the least-squares estimates, the procedure in principal components regression is to delete from the analysis those components corresponding to small eigenvalues of X'X. The regression analysis is then performed using least squares on the remaining components. If s ($1 \leq s \leq p$) components are deleted, we can partition

$$S = [S_t \vdots S_s], \gamma' = [\gamma_t' \vdots \gamma_s']$$

and

$$L = [L_t \vdots L_s]$$

where $t = p - s$. From eq. (29) we have

$$z_t = L_t^{-1} S_t' X' y \tag{30}$$

or, in terms of estimates of the original coefficients,

$$b_{pc} = S_t L_t^{-1} S_t X' y$$

By inserting $(X'X)(X'X)^{-1}$ into eq. (30) we have

$$b_{pc} = S_t L_t^{-1} X' X b$$

Marquardt (ref. 36) has shown that mean-square error $(MSE)(b_{pc}) < MSE(b)$ if and only if

$$\sum_{j=t+1}^{p} \ell_j^{-1} > \frac{1}{\sigma^2} \beta' S_s S_s' \beta$$

so that, as with $b(k)$ and $b_s(c)$, there is potential for improvement in MSE when compared with the least-squares estimator.

A major problem with the use of principal components regression is deciding which components to delete. Two criteria are usually considered:

a. Delete components associated with small eigenvalues

b. Delete components which are relatively unimportant as predictors of the response variable y.

Mansfield (ref. 53) has shown that the F-statistic used for measuring the predictiveness of a component associated with a small eigenvalue is unreliable and can lead to poor results. Mansfield recommends deleting all components associated with small eigenvalues, and he also provides a method of variable selection following principal components regression.

Marquardt (ref. 36) points out the assumption of an integral number of zero eigenvalues of X'X may be overly restrictive (see section 4.2). He

notes that in the case where X'X is actually of rank t, eq. (30) is the Moore-Penrose generalized inverse solution to the normal equations. In the case where X'X has full rank p but has several small eigenvalues, Marquardt suggested the concept of fractional rank of X, that is, we assume X to have rank f where $t < f < t + 1$ and use the generalized inverse

$$(X'X)^{+} = S_t L_t^{-1} S_t' + \frac{p - t}{\ell_{t+1}} S_{t+1} S_{t+1}'$$

The principal components estimators depend upon the particular method used for determining the significance of the coefficients. The MSE of the predicted regression function is not considered in this paper. Two different procedures for subset regression in the principal components case were considered in references 54 and 55.

The method of principal components regression is further discussed in references 31, 32, 56, and 57.

4.5 LATENT ROOT REGRESSION

One of the most important issues (ref. 32) in principal components regression is the criteria to be used in choosing a subset. There are at least two alternative criteria for deleting components:

a. Delete the components that are relatively unimportant as predictors of the original independent variables in the problem; i.e., the components having the smallest eigenvalues should be dropped.

b. Delete the components that are relatively unimportant as predictors of the dependent variable y in the problem. In this case, the components having the smallest values of the correlation between the components and y should be dropped.

Hotelling (ref. 58) has noted that in general there is no reason why components that are important as far as the independent variables of a problem are concerned will be highly correlated with the dependent

variable in a regression, so criteria a and b above are likely to lead to different results. Furthermore, it is easily shown that y need not be highly correlated with components having large eigenvalues in order for the explanatory power of the complete principal component regression to be high.

The choice of criteria must rest with the purpose of the analysis, as well as the degree to which the principal components results can be interpreted in terms of the structure of the process underlying the data for the independent variables. If the first few principal components can be related to something "real," as is hopefully the case in factor analysis, for example, then it may make sense to retain them as explanatory variables in a principal components-regression analysis, regardless of their correlation with the dependent variable. Massy (ref. 32) claims that components with large eigenvalues are usually the ones most likely to yield natural interpretations. Conversely, if the emphasis is on finding the correlates of y rather than testing its relation to any particular structural concepts, it would seem to make more sense to adopt criterion b and retain those components with the highest values of the correlation coefficients between the components and the vector y. This is often the case in purely exploratory studies.

Latent root regression is a procedure for implementing principal components regression by using criterion b above; this analysis was first suggested by Massy (ref. 32) and developed independently by Hawkins (ref. 59) and Webster, Gurst, and Mason (ref. 60). It is a modified least-squares procedure which uses the eigenvalues (latent roots) and eigenvector (latent vector) of the correlation matrix of response and regressor variables.

Analysis of these eigenvalues and vectors will enable the experimenter to

a. Identify multicollinearities among the regressor variables

b. Determine whether the multicollinearities have value in predicting the response variable

c. Obtain modified least-squares estimates of the regression coefficients through a procedure which adjusts for nonpredictive multicollinearities.

A stepwise backward elimination of variables was developed (ref. 60), using ordinary least squares or the modified procedure. Using the model $y = X\beta + e$ where the vector y and the matrix X have been standardized, let

$$t^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

and define the matrix $A = [y \vdots X]$; i.e., the $(n \times p + 1)$ matrix of standardized dependent and independent variables. A'A is the extended correlation of dependent and independent variables and has eigenvalues and eigenvectors defined by $|A'A - \lambda_j I| = 0$ and $(A'A - \ell_j I)V_j' = 0$; $j = 0, 1, \cdots, p$. Denote the elements of the j*th* eigenvector by

$$V_j' = (S_{0j}, S_{1j}, \cdots, S_{pj})$$

and let

$$V_j^{0'} = (S_{1j}, S_{2j}, \cdots, S_{pj})$$

Also let

$$S = (V_0, V_1, \cdots, V_p)$$

and

$$L = \text{diag}(\ell_j) \quad ; \quad j = 0, 1, 2, \cdots, p$$

where $\ell_0 \leq \ell_1 \leq \cdots \leq \ell_p$. Hence

$$S'(A'A)S = L$$

and

$$A'A = SLS'$$

Now note that the j*th* column of AS can be expressed as

$$AV_j = \begin{bmatrix} y_1 S_{0j} + \sum_{k=1}^{p} x_{1k} S_{kj} \\ y_2 S_{0j} + \sum_{k=1}^{p} x_{2k} S_{kj} \\ \vdots \\ y_n S_{0j} + \sum_{k=1}^{p} x_{nk} S_{kj} \end{bmatrix}$$

We also note that the j*th* eigenvalue of A'A can be expressed as

$$\ell_j = V_j'(A'A)V_j = (AV_j)'(AV_j)$$

$$= \sum_{i=1}^{n} (y_i S_{0j} + \sum_{k=1}^{p} x_{ik} S_{kj})^2 \qquad (31)$$

Thus ℓ_j is the sum of squares of the j*th* set of linear combinations of response and regressor variables which is provided by the j*th* column of AS.

If $\ell_j = 0$ for any $j = 0, 1, \cdots, p$ then each term in eq. (31) is equal to zero and an exact linear relationship exists among some or all of the columns of A. If the corresponding $S_{0j} \neq 0$, a perfect predictor exists of the form

$$\hat{y}_i = \bar{y} - tS_{0j}^{-1} \sum_{k=1}^{p} x_{ik} S_{kj}$$

If $\ell_j = 0$ and $S_{0j} = 0$, we see from eq. (31) that an exact linear dependence (exact multicollinearity) exists among the columns of X, the relationship being

$$\sum_{k=1}^{p} x_{ik}S_{kj} = 0 \quad ; \quad i = 1, 2, \cdots, n$$

In general, none of the eigenvalues will be zero, but some may be quite small. Small but nonzero eigenvalues indicate near singularities. Notice from eq. (31) that if we have $\ell_j \approx 0$ then each term in the sum must be near zero and we will have

$$S_{0j}y_i + \sum_{k=1}^{p} S_{kj}x_{ik} \approx 0 \quad ; \quad i = 1, 2, \cdots, n \tag{32}$$

If in addition $S_{0j} \approx 0$, we have a multicollinearity involving only the regressor variables and not the response variable, the relationship being

$$\sum_{k=1}^{p} S_{kj}x_{ik} \approx 0 \quad ; \quad i = 1, 2, \cdots, n$$

Since this relationship does not involve the response variable, it would be of little value for prediction.

Let us see a geometrical interpretation of a nonpredictive multicollinearity. Consider the n data points $(y_i, x_{i1}, x_{i2}, \cdots, x_{ip})$ $i = 1, 2, \cdots, n$ as n points in the p + 1 dimensional Euclidean space defined by the mutually orthogonal axes $Y, X_1, \cdots, X_p$. The eigenvectors of A'A define a second set of mutually orthogonal axes $Z_0, Z_1, \cdots, Z_p$, where Z_i is the axis defined by V_i, $i = 0, 1, \cdots, p$. The direction of axis Z_j relative to the original axes is given by the vector sum

$$\sum_{k=0}^{p} S_{kj}e_k \quad ; \quad j = 0, 1, \cdots, p$$

where $e_0, e_1, \cdots, e_p$ are unit length vectors from the origin in the direction axes $Y, X_1, X_2, \cdots, X_p$. The first element of V_j represents the cosine of the angle between axes Y and Z_j, while S_{kj} $(k = 1, 2, \cdots, p)$ represents the

cosines of the angle between axes X_j and Z_j. Assuming the eigenvectors are normalized and the eigenvalues are distinct, V_j is uniquely determined apart from a multiple of -1.

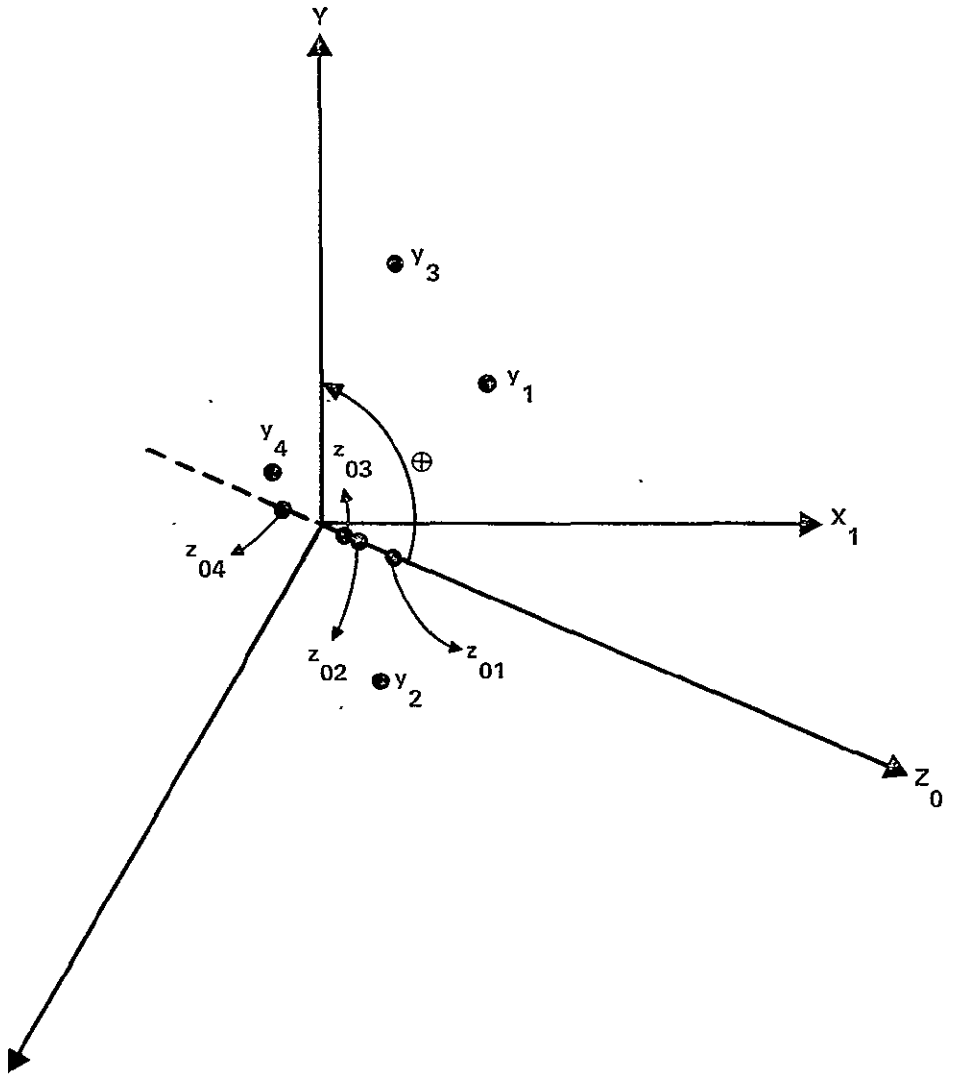
The eigenvalue corresponding to a particular eigenvector measures the spread of the n data points in the direction defined by the eigenvector. In other words, ℓ_j is the sum of squares of the projections of the n data points on the Z_j axis. A small value of ℓ_j indicates that there is little variability in the Z_j direction, i.e.,

$$z_{ij} = y_i S_{0i} + \sum_{k=1}^{p} x_{ik} S_{kj}$$

is near zero for i = 1, 2, ..., n. If S_{0j} is near zero, the axis Z_j is nearly orthogonal to the Y-axis. Hence, if both ℓ_j and S_{0j} are small, the eigenvector V_j reveals a nonpredictive near singularity; a strong linear dependence only among the independent variables which produces little or no change in the dependent variable. The situation where ℓ_j is small but S_{0j} is not small, so that the response variable is involved in the relationship, is termed predictive multicollinearity. The ability to detect the presence of predictive and nonpredictive multicollinearity and to determine the nature of the relationships through eq. (32) is one of the key features of the latent root regression procedure. This feature is not shared by any of the other procedures outlined in the previous sections.

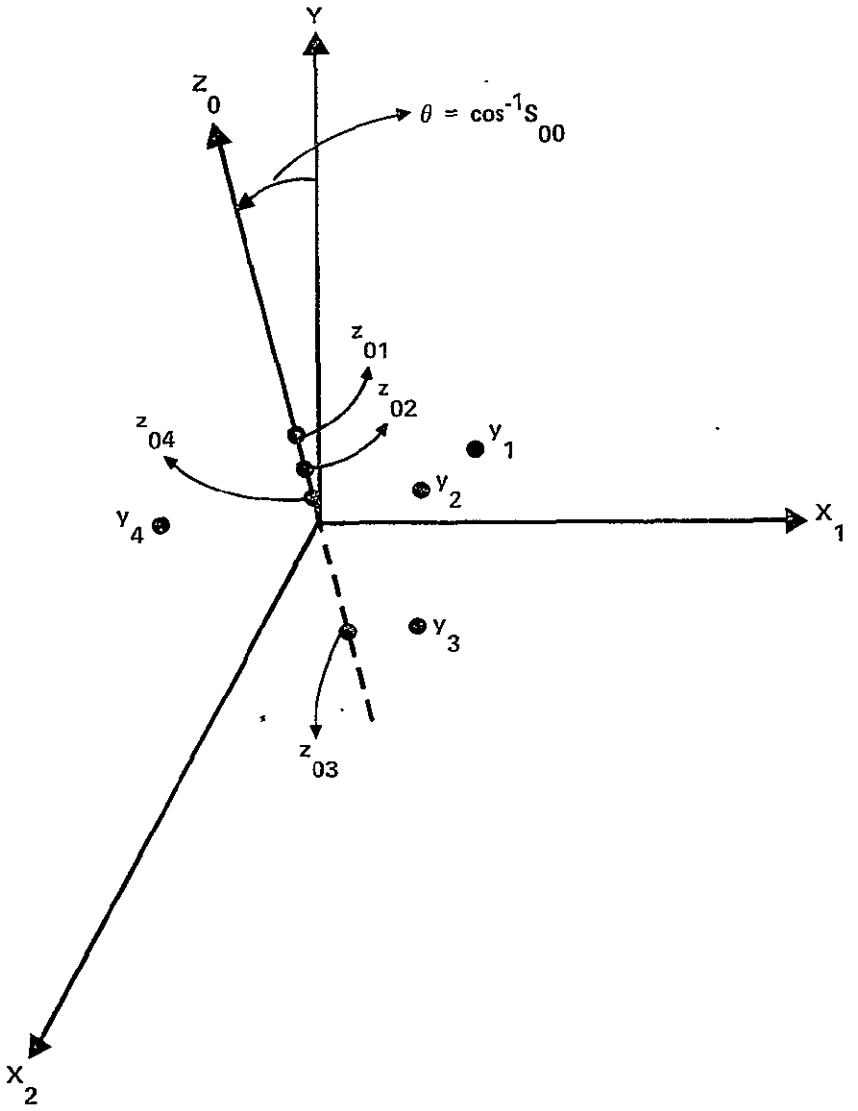
The least-squares estimator is a linear combination of all p + 1 eigenvectors, including eigenvectors corresponding to nonpredictive near singularities. The modified least squares (ref. 60) utilize only linear combinations of the eigenvectors not having both ℓ_j and S_{0j} small. In this fashion, the estimates of the regression coefficients are adjusted for the effect of nonpredictive near singularities.

Figures 4 and 5 illustrate, for three dimensions and four data points, the cases of predictive and nonpredictive multicollinearity. Nonpredictive



$$\left[S_{00} \approx 0 \text{ and } \ell_0 = \sum_{i=1}^{4} z_{i0}^2 \approx 0 \right]$$

Figure 4.– Nonpredictive multicollinearity.



$$\left[\ell_0 = \sum_{i=1}^{4} z_{0i}^2 \text{ (but } S_{00} \text{ large)} \right]$$

Figure 5.– Predictive multicollinearity.

multicollinearity characterized by a small ℓ_0 and small $|S_{00}|$ is shown in figure 4. The case in which ℓ_0 is small but $|S_{00}|$ is large is illustrated by figure 5.

Hawkins (ref. 59) arrives at similar conclusions but from a different point of view. Let $(y_i, x_{i1}, \cdots, x_{ip})$ be the i*th* data point on the p + 1 dimensional space spanned by $y, x_1, \cdots, x_p$ and let

$$y - b_1x_1 - b_2x_2 - \cdots - b_px_p = 0 \tag{33}$$

be a fitted hyperplane. Hawkins considers measuring deviations of the n data points from the hyperplane (eq. (33)) in the direction of the normal to the hyperplane rather than in the Y-direction.

If we let

λ = mean-squared deviation between fitted and observed responses in the direction of the normal line

$$\lambda = \frac{1}{n}\sum_{i=1}^{n}(\lambda_i - \bar{\lambda})^2$$

where λ_i is the deviation of the i*th* data point in the direction normal to the fitted plane, and

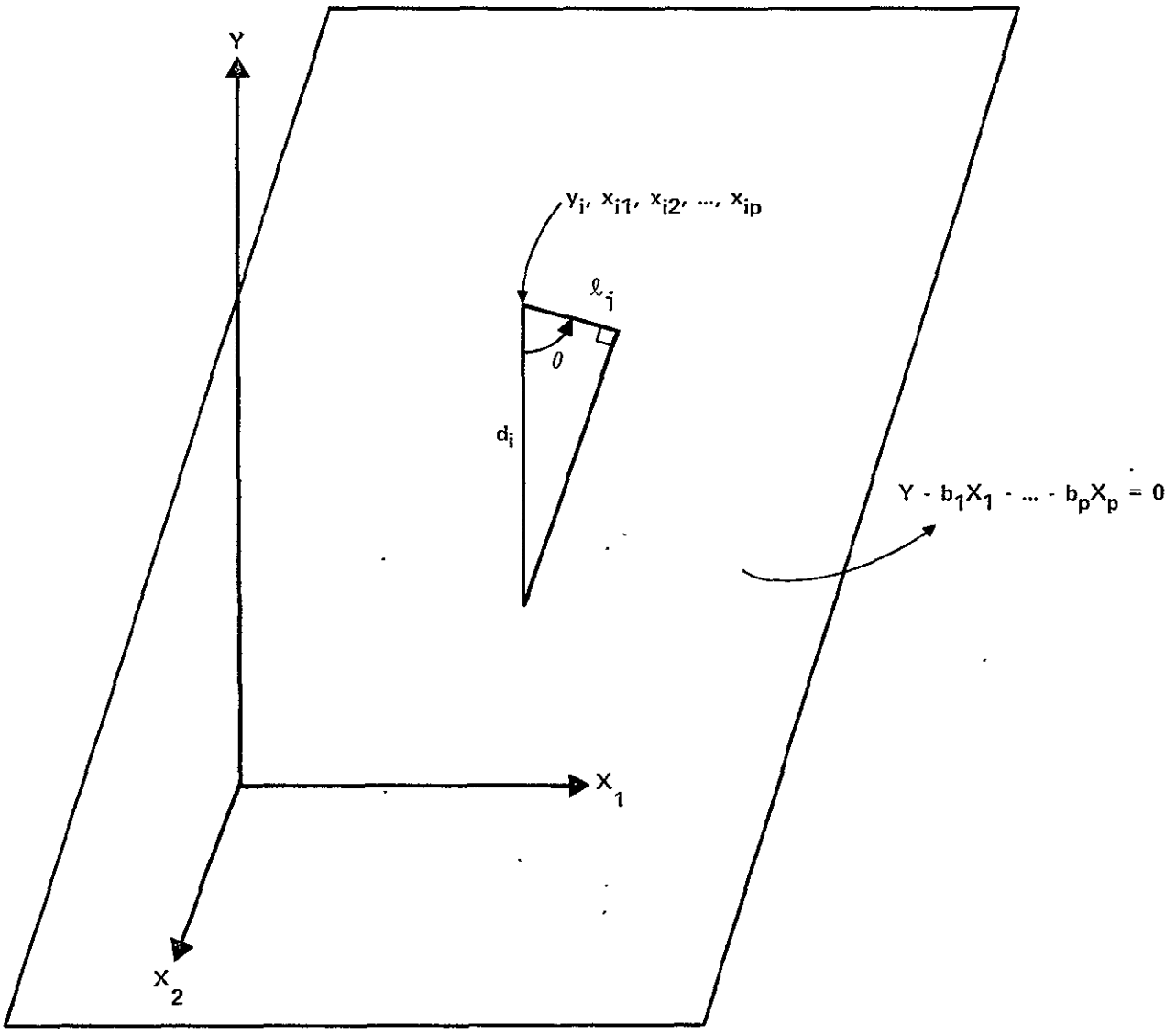
d^2 = mean-squared deviation between fitted and observed responses in the direction of the Y-axis

$$d^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(d_i - \bar{d})^2$$

we then have

$$\lambda = (\cos^2\theta)d^2 \tag{34}$$

where θ is the angle between the normal to the hyperplane and the Y-axis. Hawkins calls λ the vertical norm and d^2 the Y-norm (see fig. 6). He proposes λ as an alternate measure of the fit of the hyperplane (eq. (33)).



$$\left[\begin{array}{l} \lambda_i = i\text{th deviation in direction normal to hyperplane} \\ d_i = i\text{th deviation in direction of Y-axis} \end{array}\right]$$

Figure 6.— Vertical norm.

From eq. (34) we see that

a. If d^2 is small, indicating the hyperplane provides a good fit to the given points, then the vertical norm λ is also small.

b. If λ is small, d^2 is not necessarily small as cos θ may be small. This would correspond to a multicollinearity among the regressor variables.

Hawkins also notes that, for a hyperplane chosen to minimize d^2, the vertical norm will be equal to ℓ_0, the smallest eigenvalue of A'A, and will be in the direction defined by the corresponding eigenvector, V_0. Thus we see that the Z_0 axis (ref. 60) will be in the direction normal to the fitted hyperplane and that cos $\theta = S_{00}$, where S_{00} is the first element of V_0. Thus the nonpredictive multicollinearity characterized by a small ℓ_0 and a small S_{00} is the same as characterized (ref. 59) by a small vertical norm by a large Y-norm. This is illustrated in figure 6. If a second vector is chosen so as to be orthogonal to V_0 and to minimize the vertical norm, that vector will be V_1 and the vertical norm is now ℓ_1, the second smallest eigenvalue of A'A.

Now consider the problem of estimation. If all $S_{0j} \neq 0$, then eq. (31) will provide p + 1 prediction equations of the form

$$y^j = \overline{y}\underline{1} - tS_{0j}^{-1}XV_j^0 \quad ; \quad j = 0, 1, \cdots, p \tag{35}$$

where $\underline{1}$ is an m × 1 vector of 1's.

Normally, none of the individual equations in eq. (35) will by itself be a good predictor. Linear combinations of these predictors, therefore, will be used to obtain estimates of the parameters of the model. Consider the following arbitrary linear combination of the predictors (eq. 35)):

$$\hat{y} = \sum_{j=0}^{p} a_j S_{0j} y^j$$

Imposing the restriction

$$\sum_{j=0}^{p} a_j S_{0j} = 1$$

yields

$$\hat{y} = \bar{y}\underline{1} - tX\left(\sum_{j=0}^{p} a_j V_j^0\right) \tag{36}$$

The residual sum of squares using this predictor is

$$(y - \hat{y})'(y - \hat{y}) = t^2 a' L_a = t^2 \sum_{j=0}^{p} a_j^2 \ell_j$$

where $a' = (a_0, a_1, \cdots, a_p)$.

If a is now chosen to minimize the residual sum of squares subject to the above restriction, eq. (32) will yield the least-squares estimator. Thus we wish to minimize

$$p(a) = t^2 \sum_{j=0}^{p} a_j^2 \ell_j - 2\mu\left(\sum_{j=0}^{p} a_j S_{0j} - 1\right) \tag{37}$$

where -2μ is a Lagrangian multiplier. The solution is (ref. 60)

$$a_j = S_{0j} \ell_j^{-1} \left(\sum_{k=0}^{p} \ell_k^{*-1}\right)^{-1} \quad ; \quad j = 0, 1, \cdots, p \tag{38}$$

where $\ell_j^* = \ell_j / S_{0j}^2$.

From eqs. (36) and (38) the least-squares estimator of the regression coefficients is then given by

$$\left.\begin{aligned} b &= -t\left(\sum_{k=0}^{p} \ell_k^{*-1}\right)^{-1} \sum_{j=0}^{p} S_{0j}\ell_j^{-1}V_j^0 \\ &= -t \sum_{j=0}^{p} a_j V_j^0 \end{aligned}\right\} \qquad (39)$$

with residual sum of squares error (SSE)

$$SSE = t^2\left(\sum_{k=0}^{p} \ell_k^{*-1}\right)^{-1}$$

Suppose now that eigenvectors $V_0, V_1, \cdots, V_{k-1}$ correspond to nonpredictive near singularities. An obvious modification of the above procedure is to take a linear combination of the predictors in eq. (35) except for those which correspond to nonpredictive multicollinearities. We should then expect to obtain improved estimates of the regression coefficients without losing very much of the ability to predict the response variable y.

The above least-squares estimator can be adjusted by setting $a_0 = a_1 = \cdots = a_{k-1} = 0$. Then minimizing eq. (37) yields

$$a_j = S_{0j}\ell_j^{-1}\left(\sum_{r=k}^{p} \ell_r^{*-1}\right)^{-1} \quad ; \quad j = k, k+1, \cdots, p$$

so that the modified least-squares coefficients are

$$b_{LR} = -t\left(\sum_{r=k}^{p} \ell_r^{*-1}\right)^{-1} \sum_{j=k}^{p} S_{0j}\ell^{-1}V_j^0 \qquad (40)$$

with residual sum of squares

$$SSE_{LR} = t^2\left(\sum_{r=k}^{p} \ell_r^{-1}\right)^{-1}$$

Note that if X is not of full column rank, i.e., X'X is singular, this same procedure can be applied and minimization of eq. (37) will yield results identical with eq. (40). This follows from the fact that a singular matrix X'X implies some of the ℓ_j and corresponding S_{0j} of A'A will be zero-equivalent to setting the appropriate a_j in eq. (37) to zero. Hence solutions to the normal equations can be obtained from this procedure regardless of whether X is of full column rank.

The estimates obtained from eqs. (39) and (40) are often strikingly different when X is near singular. One reason for this is that the a_j corresponding to eigenvectors revealing nonpredictive near singularities are often large relative to the remaining a_j. When this occurs, the terms $a_j V_j^0$, $j = 0, 1, \cdots, k-1$ can dominate b. Removing these dominating terms will then yield more accurate estimates of the true parameters β. The latent root estimator is then a linear combination of vectors essentially orthogonal to the subspace defined by the nonpredictive multicollinearities and hence may yield more accurate estimates, depending on the orientation of β relative to that subspace. (See refs. 33 and 60 for examples.)

In this example, with n = 12 and p = 6, $\ell_0 = 0.001$ and $\ell_1 = 0.0287$ while the remaining eigenvalues are larger than 0.3. The corresponding S_{0j} are 0.0339 and 0.6987 so that V_0 indicates a nonpredictive multicollinearity while V_1 does not. Since ℓ_1 is small and S_{01} is the largest S_{0j}, V_1 provides more information about the underlying model than any of the other eigenvectors. In this example, the latent root estimator is formed by removing V_0 from the analysis. The a_j for least squares (LS) and for latent root (LR) are then:

	a_0	a_1	a_2	a_3	a_4	a_5	a_6
LS	1.760	1.317	0.012	0.002	0.017	0.015	0.004
LR	0	1.404	0.013	0.002	0.018	0.017	0.004

Notice that the least-squares procedure gives 56 percent of the total weight to V_0 and only 42 percent is given to V_1, the vector providing the most information about the model.

The latent root procedure gives zero weight to V_0 and 96 percent of the total weight to V_1. The least squares estimates of the parameters of model $y = X\beta + e$ are given in the first row of table I with the true values of the parameters in the third row. The matrix A'A, the extended correlation matrix, is given in table II. The eigenvectors of A'A are given in table III. Note that the four estimates of parameters involved in the near singularity are moderately large negative values. The fact that these estimates are similar despite the differences in their true values of the parameters is indicative of the effect of the near singularity. Using the modified least-squares procedure by computing a linear combination of all the eigenvectors except V_0 yields the estimates in the second row of table I. The absolute values of the first elements of the eigenvectors and the eigenvalues are given in table IV.

TABLE I.— LEAST SQUARES AND MODIFIED LEAST SQUARES

	b_1	b_2	b_3	b_4	b_5	b_6	$\hat{\sigma}^2$
LS	-6.0378	-8.472	-10.1435	-11.7271	4.0967	9.4506	1.2762
LR	2.5447	-0.3982	0.2416	-0.7348	4.2125	9.4914	1.3575
True values	2.000	1.000	0.2000	-2.000	3.000	10.000	1.000

TABLE II.— A'A, THE EXTENDED CORRELATION MATRIX

y	X_1	X_2	X_3	X_4	X_5	X_6
1.000	0.252	-0.099	0.217	-0.339	0.364	0.811
	1.000	0.052	-0.343	-0.498	0.417	-0.192
		1.000	-0.432	-0.371	0.485	-0.317
			1.000	-0.355	-0.505	0.494
				1.000	-0.215	-0.087
					1.000	-0.123
						1.000

TABLE III.— EIGENVECTORS OF A'A

j	S_{0j}	S_{1j}	S_{2j}	S_{3j}	S_{4j}	S_{5j}	S_{6j}
6	0.1653	-0.3300	-0.4471	0.5165	0.1009	-0.4370	0.4427
5	.6006	.3444	.0925	.1436	-.4785	.3294	.3925
4	.3406	-.1134	-.1886	-.4556	.6518	.3303	.3068
3	.0388	-.6944	.6712	.0937	-.0417	.1427	.1869
2	.0713	.2344	.3550	-.4505	-.0128	-.7033	.3410
1	.6987	-.1694	.0242	-.0091	.0453	-.2766	-.6355
0	.0339	.4402	.4229	.5416	.5763	-.0071	-.0276

TABLE IV.— INDEXES FOR STANDARDIZED PREDICTION EQUATIONS

j:	0	1	2	3	4	5	6
ℓ_j:	0.0010	0.0287	0.3115	0.9178	1.1150	2.1816	2.444
$\lvert S_{0j} \rvert$:	.0399	.6987	.0713	.0388	.3406	.6006	.1653
$A_j t$:	19.1496	14.3352	.1347	.0249	.1798	.1620	.0398

White (ref. 61) studies the problem of deciding whether an eigenvector of A'A should be removed from the analysis. Upon first consideration, it appears that the problem centers on deciding when $|S_{0j}|$ and ℓ_j are small enough to indicate the presence of nonpredictive multicollinearities. White proposes that a more crucial consideration is the orientation of the true coefficient vector, β, in the p-dimensional subspace spanned by the eigenvectors of X'X, the correlation matrix. His proposal is plausible if we are willing to accept that

$$V_j^0 \approx V_{j+1} \quad ; \quad j = 0, 1, \cdots, k - 1$$

where

V_j; $j = 1, 2, \cdots, p$ = the eigenvectors of X'X

$k - 1$ = the number of nonpredictive multicollinearities

5. APPLICATIONS

The author used the procedure of combining some of the good features of the several biased techniques (section 4) and the unbiasness property of the ordinary least-squares estimator, using weather data for Oklahoma and Texas. In addition, trend data were available for Oklahoma. The weather variables are the following:

Variable	Name	Type
X_1	January	Precipitation for current year
X_2	February	
X_3	March	
X_4	April	
X_5	May	
X_6	June	
X_7	August	Precipitation for previous year
X_8	September	
X_9	October	
X_{10}	November	
X_{11}	December	
X_{12}	January	Mean temperature
X_{13}	February	
X_{14}	March	
X_{15}	April	
X_{16}	May	
X_{17}	June	

Variable	Name	Type
X_{18}	January	Percent evapotranspiration
X_{19}	February	
X_{20}	March	
X_{21}	April	
X_{22}	May	
X_{23}	June	
X_{24}	Trend	Trend
y	Yield	Yield

The model postulated is

$$y = \beta_0 + \sum_{i=1}^{24} \beta_i X_i + e$$

and if the matrix is standardized, then

$$\underline{y} = X\beta + e$$

The 45 data points consist of weather information from 1932 to 1976, inclusive.

The first task is to reduce the number of variables in a meaningful way. The all-possible regressions procedure (see section 3.1) was used to analyze all possible subsets of variables. The optimum number of variables that should be kept in the model was determined by the adjusted R^2 (upper bound) and the Mallows' C_p (lower bound). (See sections 3.2 and 3.3.)

The following results were obtained from the all-possible regressions approach.

OKLAHOMA

Ten variables were selected by using the adjusted R^2 as a criterion of goodness of fit, X_1, X_3, X_5, X_6, X_9, X_{11}, X_{14}, X_{17}, X_{20}, and X_{24}.

The Mallows' C_p criterion selected seven variables: X_3, X_5, X_6, X_{14}, X_{17}, X_{20}, and X_{24}. The highest R^2 possible (using all 24 variables) is 91.65.

Ten Variable Results

Two small eigenvalues of the extended correlation matrix, $\ell_1 = 0.004389$ and $\ell_2 = 0.073157$, were obtained, indicating two multicollinearities. Their respective values of $|S_{01}|$ and $|S_{02}|$ are 0.091927 and 0.71778, indicating that the first eigenvectors correspond to a nonpredictive near singularity. In fact, the second eigenvector provides the most information about yield. This situation is very similar to the example given in section 4. The second eigenvector is

$$V_2' = [-0.04957, -0.16844, 0.29297, 0.21722, -0.07449, 0.06980,$$
$$0.05778, 0.17312, 0.2233, -0.47684, 0.71778]$$

so the following equation holds.

$$-0.04957X_1 - 0.16844X_2 + 0.29297X_3 + 0.21722X_4 - 0.07449X_5$$
$$+ 0.06980X_6 + 0.05778X_7 + 0.17312X_8 + 0.2233X_9 - 0.47684X_{10}$$
$$+ 0.71778Y = 0.073157$$

Notice that yield, y, is heavily involved in the multicollinearity.

The computed value of R^2 is 89.22 and the value of the determinant of the correlation matrix, $|R|$, is 0.003211. $1/|R|$ is called the generalized variance (ref. 56). The size of $|R|$ indicates possible instability in the estimates of the parameters.

Thus far, latent root regression techniques have been used to determine the source and type of the multicollinearities present in the data; now the problem is to remove the multicollinearity by deleting one or more variables. Looking at the first eigenvector (which corresponds to the nonpredictive near singularity), observe that two components are larger than the other; i.e.,

$$V_1' = [0.019389,\ 0.004192,\ -0.01999,\ -0.025515,\ 0.018872,$$
$$-0.004851,\ -0.716423,\ -0.009117,\ 0.687389,\ 0.062269,\ -0.091927]$$

Again, the interpretation is similar as the equation above, but here the variable yield is not involved in the multicollinearity; only variables X_{14} and X_{20} are involved in the nonpredictive near singularity. The weights given to X_{14} and X_{20} are -0.716423 and 0.687389, which are much larger than the other components in V_1.

We delete X_{20} rather than X_{14} because X_{20} is more correlated with yield than X_{14}. Alternatively, the variable that least decreases R^2 could be deleted.

Nine Variable Results

With variable X_{20} deleted, we have only one small eigenvalue of A'A, $\ell_1 = 0.066375$. Since $|S_{01}| = 0.70785$, the eigenvector V_1 corresponds to a predictive near singularity. This eigenvector is

$$V_1' = [0.05068,\ 0.161875,\ -0.28525,\ -0.21834,\ 0.073801,\ -0.069854,$$
$$-0.291657,\ -0.169026,\ -0.466021,\ -0.707851]$$

The interpretation of this eigenvector is as before. The computed value of R^2 is 87.33, so the net loss in goodness of fit is 1.89. The value of $|R|$ is 0.2967, which is much higher than the previous value of 0.003211. This new value of $|R|$ is an indication of stability of the parameters estimated; i.e., the variances of the estimates are not too large.

Let us see if the current results can be improved by introducing some bias to the estimator (generalized ridge procedure). The determinant of R went up to 0.3518 and the estimated R^2 is now 87.02, so a loss of 0.308 in R^2 gives an improvement of 0.011. This result seems to be adequate as an initial start in the modeling of wheat yield. The next step should be to consider interaction and square terms. The resulting values of the parameters are:

$$b_1 = 0.01005$$
$$b_2 = 0.02819$$
$$b_5 = -0.02465$$
$$b_6 = -0.02182$$
$$b_9 = 0.01048$$
$$b_{11} = -0.01537$$
$$b_{14} = -0.60445$$
$$b_{17} = -0.51368$$
$$b_{24} = 0.26523$$

The value of b_0, the intercept, needs to be calculated by using the sample means of the X's and Y.

TEXAS

Ten variables were selected by using the adjusted R^2 as a criterion of goodness of fit. They are: X_5, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{16}, X_{18}, X_{19}, and X_{22}. Trend, X_{24}, is not available. Mallows' C_p criterion selected five variables: X_3, X_5, X_7, X_{14}, and X_{18}. If all 23 variables are used, the computed value of R^2 is 53.458.

Ten Variable Results

The A'A matrix has three small eigenvalues: $\ell_1 = 0.000332$, $\ell_2 = 0.005177$, and $\ell_3 = 0.060468$, so we have three near singularities. The respective values of $|S_{0i}|$ are: $|S_{01}| = 0.004814$, $|S_{02}| = 0.025483$, and $|S_{03}| = 0.082601$, indicating that we have three nonpredictive near singularities. The determinant of R, $|R|$, is 0.00000054 and the computed value of R^2 is 49.45. The vector V_1 is

$$V_1' = [0.005817, -0.008772, 0.002736, -0.012504, 0.014695,$$
$$-0.00475, -0.706582, 0.015953, -0.020851,$$
$$0.706764, 0.004814]$$

By observing the eigenvector V_1, we see that the weights for X_{16} and X_{22}, -0.706582 are 0.706764, are clearly larger than the other weights. Therefore,

X_{16} and X_{20} are candidates for deletion. We delete X_{16} because the correlation of X_{16} and yield is lower than the correlation of X_{20} and yield.

Nine Variable Results

Delete X_{16}, the A'A has two small eigenvalues: $\ell_1 = 0.005177$ and $\ell_2 = 0.060408$. Their respective values of $|S_{0i}|$ are: $|S_{01}| = 0.005177$ and $|S_{02}| = 0.060408$. The value of $|R|$ is 0.00079055 and the value of R^2 is 47.714. The candidates for deletion were X_{13} and X_{19}; X_{19} was deleted by the same reasons as before.

Eight Variable Results

Delete X_{19}, and A'A has only one small eigenvalue, $\ell_1 = 0.06354$, which corresponds to one nonpredictive near singularity. $|R|$ is 0.0714075 and R^2 is 43.88. The candidates for deletion are X_{12} and X_{18}. Variable X_{12} was deleted.

Seven Variable Results

Delete X_{12}, and A'A has no small eigenvalues, $|R|$ is 0.5797, and R^2 is 42.72. This value of $|R|$ is excellent and the loss in R^2 has not been too great.

In this case, there is no need to enter the biased estimation procedure; therefore, the unbiased least-squares estimator is used to estimate the parameters of the yield model as

$$b_5 = -0.05516$$
$$b_{10} = 0.07096$$
$$b_{11} = -0.06655$$
$$b_{13} = -0.57036$$
$$b_{14} = -0.70483$$
$$b_{18} = -0.43672$$
$$b_{22} = -0.1944$$

6. CONCLUSIONS AND RECOMMENDATIONS

It was observed that OLS is not adequate as an estimation procedure when the independent or regressor variables are involved in multicollinearities. This was shown to cause the presence of small eigenvalues of the extended correlation matrix A'A. It has been demonstrated that the biased estimation techniques and the all-possible subset regression can help in finding a suitable model for predicting yield.

Latent root regression is an excellent tool that allows us to find how many predictive an nonpredictive multicollinearities we have, and it also tells us exactly what variables are involved in the multicollinearities. Thus, we can decide what variables to drop from the model to remove the multicollinearities and hence obtain estimates with small variances.

It is recommended that the procedures discussed in this memorandum be made available in the Earth Observations Division Laboratory for Applications of Remote Sensing classification system. The author has made available to NASA/JSC personnel the necessary programs to implement these techniques.

The results presented in this memorandum are the initial attempts to find a yield model for wheat. Additional research should be conducted to estimate interaction terms and other ways of measuring trend.

7. REFERENCES

1. Bloomfield, P.; and Watson, G.: The Inefficiency of Least Squares. Biometrika, vol. 62, 1975.

2. Seber, G. A. F.: Linear Regression Analysis. John Wiley and Sons (New York), 1977.

3. Graybill, F.: Theory and Application of the Linear Model. Duxbury Press (Mass.), 1976.

4. Stuff, R. G.; and Wilcox, D. D.: A Comparison of the Effect of 10- and 30-Day Weather Data Intervals on the Correlation of North Dakota Wheat Yields With Temperature and Precipitation. LEC-4984, 1974.

5. Thompson, L.: Weather and Technology in the Production of Wheat in the United States. J. Soil and Water Conservation, vol. 24, 1969, pp. 221-224.

6. Boullion, T. L.; and Odell, P. L.: Generalized Inverse Matrices. John Wiley and Sons (New York), 1971.

7. Johnston, J.: Econometric Methods. Second ed., McGraw-Hill Book Co. (New York), 1972.

8. Farrar, D. E.; and Glauber, R. R.: Multicollinearity in Regression Analysis: The Problem Revisited. Review of Economics and Statistics, vol. 49, 1967, pp. 92-107.

9. Silvey, S. D.: Multicollinearity and Imprecise Estimation. JRSS, ver. B, vol. 31, 1969, pp. 539-552.

10. Berk, Kenneth N.: Comparing Subset Regression Procedures. Technometrics, vol. 20, 1978, pp. 1-6.

11. Garside, M. J.: The Best Subset in Multiple Regression Analysis. Appl. Stat., vol. 14, 1965, pp. 196-200.

12. Schatzoff, M.; Tsao, R.; and Fienberg, S.: Efficient Calculations of All Possible Regressions. Technometrics, vol. 10, 1968, pp. 769-779.

13. Furnival, G. M.: All Possible Regressions With Less Computation. Technometrics, vol. 13, 1971, pp. 403-408.

14. Garside, M. J.: Some Computational Procedures for the Best Subset Problem. Appl. Stat., vol. 20, 1971, pp. 8-15.

15. Furnival, G. M.; and Wilson, R. W. M.: Regressions by Leaps and Bounds. Technometrics, vol. 16, 1974, pp. 499-511.

16. Hocking, R. R.; and Leslie, R. M.: Selection of the Best Subset in Regression Analysis. Technometrics, vol. 9, 1967, pp. 531-540.

17. Beale, E. M. L.; Kendall, M. G.; and Mann, D. W.: The Discarding of Variables in Multivariate Analysis. Biometrika, vol. 54, 1967, pp. 357-366.

18. LaMotte, L. R.; and Hocking, R. R.: Computational Efficiency in the Selection of Regression Variables. Technometrics, vol. 12, 1970, pp. 83-93.

19. Hocking, R. R.: Criteria for Selection of a Subset Regression: Which One Should be Used? Technometrics, vol. 14, 1972, pp. 967-970.

20. Mallows, C. L.: Choosing Variables in a Linear Regression: A Graphical Aid. Presented at the Central Regional Meeting of the Institute of Math. Stat., Manhattan, Kansas, 1964.

21. Mallows, C. L.: Choosing a Subset. Presented at the Joint Stat. Meeting, Los Angeles, Calif., 1966.

22. Mallows, C. L.: Some Comments on C_p. Technometrics, vol. 15, 1973, pp. 661-675.

23. Gorman, J. W.; and Toman, R. J.: Selection of Variables for Fitting Equations to Data. Technometrics, vol. 8, 1966, pp. 27-51.

24. Kennard, R.: A Note on the C_p Statistic. Technometrics, vol. 13, no. 4, Nov. 1971, pp. 899-900.

25. Stein, C.: Multiple Regression, Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. I. Olkin, ed., Stanford Univ. Press, 1960, pp. 424-443.

26. James, W.; and Stein, C.: Estimation with Quadratic Loss. Proc. of the Fourth Berkeley Symp. on Math. Stat. and Prob., vol. I, Univ. Calif. Press, 1961, pp. 361-379.

27. Sclove, S. L.: Improved Estimators for Coefficients in Linear Regression. J. American Stat. Assoc., vol. 63, 1968, pp. 596-606.

28. Hoerl, A. E,; and Kennard, R. W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, vol. 12, 1970, pp. 55-68.

29. Hoerl, A. E.; and Kennard, R. W.: Ridge Regression: Applications to Nonorthogonal Problems. Technometrics, vol. 12, 1970, pp. 69-82.

30. Mayer, Lawrence S.; and Willke, Thomas A.: On Biased Estimation in Linear Models. Technometrics, vol. 15, no. 3, Aug. 1973, pp. 497-508.

31. Kendall, M. G.: A Course in Multivariate Analysis. Hafner Pub. Co. (New York), 1957.

32. Massy, W. F.: Principal Components Regression in Exploratory Statistical Research. J. American Stat. Assoc., vol. 60, 1965, pp. 234-246.

33. Gunst, R. F.; and Mason, R. L.: Advantages of Examining Multicollinearities in Regression Analysis. Biometrics, vol. 33, 1977, pp. 249-260.

34. Newhouse, Joseph P.; and Oman, Samuel D.: An Evaluation of Ridge Estimators. R-716-PR, Rand Corp. (AD-723626), 1971.

35. LaMotte, L. R.: Best, Bayes, and Ridge Linear Estimation. University of Houston, 1976.

36. Marquardt, D. W.: Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. Technometrics, vol. 12, 1970, pp. 591-612.

37. Searle, S. R.: Linear Models. John Wiley and Sons, Inc. (New York) 1971.

38. Kendall, M.; and Stuart, A.: The Advanced Theory of Statistics, vol. 3, Griffin (London) 1968.

39. Rao, Colyompudi R.: Linear Statistical Inference and Its Applications. John Wiley and Sons, Inc. (New York), 1965.

40. Walpole, R.; and Myers, R.: Probability and Statistics for Engineers and Scientists. McMillan, 1978.

41. Hoerl, A.; Kennard, R.; and Baldwin, Kent: Ridge Regression: Some Simulations. Commun. in Stat., vol. 4, 1975, p. 2.

42. Lawless, J. F.; and Wang, P.: A Simulation Study of Ridge and Other Regression Estimators. Comm. in Stat., A, Theory and Methods, vol. 1, 1976, pp. 307-323.

43. McDonald, G. C.; and Galarneau, D. I.: A Monte Carlo Evaluation of Some Ridge-Type Estimators. J. American Stat. Assoc., vol. 70, 1975, pp. 407-416.

44. Farebrother, R. W.: Minimum Mean Square Error Linear Estimator and Ridge Regression. Technometrics, vol. 17, 1975, pp. 127-128.

45. Hoerl, A.; and Kennard, R.: Ridge Regression: Iterative Estimation of the Biasing Parameter. Comm. in Stat., A5(1), 1976, pp. 77-88.

46. Hocking, R. R.: The Analysis and Selection of Variables in Linear Regression. Biometrics, vol. 32, 1976, pp. 1-50.

47. Sidik, S. M.: Comparison of Some Biased Estimation Methods in the Linear Model. NASA TN D-7932, 1975.

48. Obenchain, R. L.: Classical F-tests and Confidence Regions for Ridge Regression. Technometrics, vol. 19:4, 1977.

49. Albert, A.: Regression and the Moore-Penrose Pseudoinverse. Academic Press (New York), 1972.

50. Rao, C. R.; and Mitra, S. K.: Generalized Inverse of Matrices and its Applications. John Wiley and Sons, Inc. (New York), 1971.

51. Hocking, R.; Speed, F.; and Lynn, M.: A Class of Biased Estimation in Linear Regression. Technometrics, vol. 18:4, 1976.

52. Obenchain, R. L.: Ridge Analysis Following A Preliminary Test of The Shrunken Hypothesis. Technometrics, vol. 17:4, 1975.

53. Mansfield, E. R.; Webster, J. T.; and Gunst, R. F.: An Analytic Variable Selection Technique for Principal Component Regression. Applied Statistics, vol. 26, 1977, pp. 34-40.

54. Kennedy, W. J.; and Bancroft, T. A.: Model Building for Prediction in Regression Based Upon Repeated Significance Tests. Ann. Math. Stat., vol. 42, 1971, pp. 1273-1284.

55. Holms, Arthur G.: "Chain Pooling" to Minimize Errors in Subset Regression. NASA TM X-71645, 1974.

56. Anderson, T. W.: An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, Inc. (New York), 1958.

57. Rao, C. R.: The Use and Interpretation of Principle Components Analysis in Applied Research. Sankhya. A(26), 1965, pp. 329-358.

58. Hotelling, H.: The Relations of the Newer Multivariate Statistical Methods to Factor Analysis. Brit. J. Stat. Psychol., vol. 10, 1957, pp. 69-79.

59. Hawkins, D. M.: On the Investigation of Alternative Regressions by Principal Component Analysis. Applied Statistics, vol. 22, 1973, pp. 257-286.

60. Webster, J. T.; Gunst, R. F.; and Mason, R. L.: Latent Root Regression Analysis. Technometrics, vol. 16, 1974, pp. 513-522.

61. White, J.: Inference Procedures for Latent Root Regression Analysis. Ph.D. Dissertation, Department of Statistics, Southern Methodist University, 1976.


APR 1979
RECEIVED
NASA STI FACILITY
ACCESS DEPT.