BIASED ESTIMATION TECHNIQUES FOR MULTIPLE LINEAR REGRESSION

by

PHILLIP DEAN WITTMER

B. S., Kansas State University, 1975

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1976

Approved by:

George A. Milliken
Major Professor

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my Major Professor, Dr. George Milliken, for his guidance and encouragement in this endeavor. He always had time for my problems and questions. My thanks also to Dr. Krishna Akkina and Dr. Holly Fryer for their additional support and suggestions.

Finally, a special "Thank You" to my wife Gloria, not only for her help with typing and drawings, but also for her neverending patience and devotion throughout my graduate studies.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Multiple linear regression is one of the most widely used of all statistical methods. Statisticians and non-statisticians alike have found it to be a useful tool for modeling the response of a dependent variable as it is influenced by independent variables.

When dealing with experimental data, one often encounters the problem of near multicollinearity of the data vectors. Usually, this is due to high correlations between two or more of the explanatory variables. When this occurs, the least squares estimators often contain values which are useless in the sense that they are extremely large or even of the wrong sign. (Wrong in the sense that they deviate from the accepted theory for the related field.) Hoerl and Kennard [6] refer to estimates of that type as unstable estimates.

This report investigates techniques for estimating the parameters of the linear model when near multicollinearity exists between the independent variables. The linear model is:

$$Y = X\beta + \varepsilon \tag{1.1}$$

where; (1) Y is an nxl vector of observations on the dependent variable.

(2) X is an nxp matrix of observations on the independent variables such that $\rho(X)=p$ where $p \leq n$.

(3) ε is an nxl vector of unobservable random errors such that

(a) $E(\varepsilon)=0$ and
(b) $E[\varepsilon\varepsilon']= \sigma^2 I$

Unless otherwise specified, X'X is assumed to be in the form of a pxp correlation matrix. See Appendix I for details of transforming the matrix to this form and obtaining the estimate of β.

The Ordinary Least Squares (OLS) estimator $\hat{\beta}$, of β is;

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

That estimator is unbiased for β and, by the Gauss-Markc Theorem, has minimum variance among the class of linear unbiased estimators of β. Computationally, the least squares procedure is good if X'X is well conditioned, i.e., not singular or near singular. If the matrix is ill-conditioned, the analyst will be tempted to delete variables in an attempt to remove the multicolinearities. This is hardly satisfactory when the model is correct as specified. We therefore look for more useful estimators, which are biased, but have smaller mean square error (MSE).

Before we discuss biased estimation techniques, we need to examine the characteristics of OLS estimators when the X'X matrix is ill-conditioned. The covariance matrix of the least squares estimator is

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

Let the distance between the OLS estimator and the true but unknown value of β be denoted by $L_1 = ||\hat{\beta}-\beta||$. Then,

(1) $L_1{}^2 = (\hat{\beta}-\beta)'(\hat{\beta}-\beta)$ (1.2)

(2) $E(L_1{}^2) = \sigma^2 \mathrm{tr}(X'X)^{-1}$ (1.3)

(3) $E[\hat{\beta}'\hat{\beta}] = \beta'\beta + \sigma^2 \mathrm{tr}(X'X)^{-1}$ (1.4)

and if $\epsilon \sim N(0,\sigma^2 I)$, then

(4) $\mathrm{Var}(L_1{}^2) = 2\sigma^4 \mathrm{tr}(X'X)^{-2}$

Denote the characteristic roots of X'X by;

$$\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p = \lambda_{min} > 0,$$

then (1.3) can be rewritten as

$$E(L_1{}^2) = \sigma^2 \sum_{i=1}^{p}(1/\lambda_i) \qquad (1.5)$$

and if the errors are normal,

$$\mathrm{Var}(L_1^2) = 2\sigma^4\ \Sigma(1/\lambda_i)^2 \qquad (1.6)$$

Hence, lower bounds for $E(L_1^2)$ and $\mathrm{Var}(L_1^2)$ are $\sigma^2/\lambda_{\min}$ and $2\sigma^4/\lambda^2{}_{\min}$ respectively.

When one or more of the λ_i are small, this indicates a linear dependence of the ith column vector on the other p - 1 vectors, and we say X'X is ill-conditioned. If this occurs, the distance from $\hat{\beta}$ to β is large, as indicated by coefficients, $\hat{\beta}_i$, large in absolute value or with reversed signs, as alluded to earlier. By definition, the least squares estimate is that value of β which minimizes

$$\Phi(\beta) = (Y - X\beta)'(Y - X\beta) \qquad (1.7)$$

The λ_i measure the sensitivity of the solution to (1.7) and thus should be utilized to construct "better estimates". The criterion for determining which estimators are "better" differs with the biased technique. Most authors strive to minimize the MSE.

The biased estimation techniques outlined in the remainder of the report all utilize the small λ_i's in one way or another to aid in the estimation of β. The ways in which these near-singularities are utilized to predictive advantage are the basic differences in the techniques. They all achieve a reduction in the length of the vector of estimated coefficients $(\hat{\beta}'\hat{\beta})$ when compared with the length of the vector of OLS coefficients.

Another view of the problem that ill-conditioning creates in the OLS estimates is given by Webster, Gunst and Mason [12]. They partition X as $X = [x_j \vdots X_j^*]$ where x_j is the vector of observations for the jth independent variable, and X_j^* is the remaining p-1 observation vectors. If c_{jj} is the jth diagonal element of $(X'X)^{-1}$,

$$\begin{aligned} c_{jj} &= [x_j'x_j - x_j'X_j^{*\prime}(X_j^{*\prime}X_j^{*})^{-1}X_j^{*}x_j]^{-1} \\ &= [1 - R_j^2]^{-1} \end{aligned} \tag{1.8}$$

where R_j is the multiple correlation coefficient between the jth independent variable and the other p-1 independent variables. Thus,

$$\mathrm{Var}(\hat{\beta}_j) = \sigma^2\, c_{jj} = \sigma^2/(1 - R_j^2)$$

Clearly, if a great amount of correlation exists between x_j and some of the columns of X_j^*, R_j^2 will be close to 1 (some of the characteristic roots will be close to zero) and $\mathrm{Var}(\hat{\beta}_j)$ will be large. This again shows how an ill-conditioned X'X matrix results in unstable coefficient estimates.

Four biased estimation techniques are explored in the succeeding four chapters. For ease of reference, these will be designated by the following names;

Chapter 2	Ridge Regression (RR)
Chapter 3	Generalized Inverses (g-inverse)
Chapter 4	Shrunken Estimators
Chapter 5	Latent Root Regression Analysis (LRRA)

Other techniques are documented in various journals, although most are highly theoretical in nature and not extremely useful in a practical context. Hence, the scope of this report will be limited to the above four techniques.

Each chapter begins with a general but straightforward description of the methods involved in implementing the technique. Following the methodology in each chapter is a section summarizing the theory backing the method. The four techniques are compared and contrasted in Chapter 6.

In Chapter 7, a set of nonorthogonal data is analyzed by each method in order to illustrate the computations and analysis involved. It is also the aim here to depict with real data what happens to the various parameter estimates.

It is not the intent of this report to critique or formulate opinions as to the "correctness" or usefulness of the techniques. The reader may draw his own conclusions from the presentation and examples which follow, in addition to an examination of the resource materials.

CHAPTER 2

RIDGE REGRESSION

2.1 - Method

The form of the ridge estimator proposed by Hoerl and Kennard in [6] is;

$$\begin{aligned}\hat{\beta}^* &= [X'X + kI]^{-1}X'Y \qquad ;k \geq 0 \qquad (2.1)\\ &= WX'Y\end{aligned}$$

For an estimate $\hat{\beta}^*$ the residual sums of squares is

$$\begin{aligned}\Phi^*(k) &= (Y - X\hat{\beta}^*)'(Y - X\hat{\beta}^*)\\ &= Y'Y - \hat{\beta}^{*\prime}X'Y - k(\hat{\beta}^*)'(\hat{\beta}^*) \qquad (2.2)\end{aligned}$$

where $\Phi^*(k)$ is the total sums of squares less the "regression" sums of squares for $\hat{\beta}^*$ with a modification depending upon the squared length of $\hat{\beta}^*$. In practice, it is most useful to select values of k in the interval [0,1], as the system usually stabilizes very quickly in this region.

The Ridge Trace is a useful tool in determining which value of k seems to cause the system to stabilize. This is nothing more than a two-dimensional plot of the functions $\Phi^*(k)$ and $(\hat{\beta}^*)$, for values of k between zero and +1. (See Figure IV.) This portrays graphically the effects of the factor (independent variable) correlations and makes possible assessments that cannot be made even if all 2^p regressions are computed. Inherent in the analysis, therefore, is a method for selecting a "best" subset of predictor variables. Namely, those factors which contribute the most toward explaining variability in Y without being highly correlated among themselves. An illustration of the method, and the ridge trace appear in Chapter 7.

Note that the method simply involves adding a small constant k to the diagonal elements of the X'X matrix to correct for the near-singularity of the matrix, thereby making it easier to invert. The magnitude of k required to stabilize $\hat{\beta}^*$ will depend on the degree of ill-conditioning in X'X.

The recommended procedure is as follows;

(i) Calculate 11 - 20 regressions by substituting different values of k into equation (2.1). It is most helpful to begin with k=0 (OLS estimates), and use small increments (say .02) up to k=.1. Then use increments of 0.1 up to 1.0. The small increments up to k=0.1 enable one to see the rapid stabilization in the estimates which generally takes place in the early stages of the analysis. This, of course, depends on the actual problem and data at hand.

(ii) Plot the ridge trace by plotting the $\hat{\beta}_i^*$ and Φ_i^* obtained as a function of k. Take special note of those $\hat{\beta}_i^*$'s which rapidly go to zero, or even cross the zero line, taking on different signs. The analyst should also pay close attention to the squared length of the coefficient vector $\hat{\beta}^{*\prime}\hat{\beta}^*$ at those points where the system seems to stabilize and compare them to $\hat{\beta}'\hat{\beta}$. A significant reduction in the squared length of the coefficient vector corresponding to a small increase in residual sum of squares is the researcher's goal.

(iii) From the ridge trace, select a value of k for which the system seems to stabilize. When selecting k, do not allow the residual sum of squares to inflate unreasonably. Obatin the coefficients $\hat{\beta}_1^*$, $\hat{\beta}_2^*$, . . ., $\hat{\beta}_p^*$ for this value of k. Then view the system as one of p controlled factors with these coefficients as the "best" estimates.

Factors with small effects have small coefficients. To "discard" a factor, simply set its coefficient equal to zero. Factors to be eliminated should be those whose coefficients are quickly driven toward zero with the addition of $k > 0$, and hence cannot retain their predicting power.

Do not delete factors and then reestimate coefficients for the remaining factors, as it is likely that all the instabilities and over-estimation will still be present, perhaps even amplified. Hoerl and Kennard [7] give an example for which the length of the coefficient vector increases after deletion. If the discarded factors contribute the least, their estimation should not be affected by their associations with other factors.

An iterative procedure to estimate k_i is given in [5] and [6]. This is presented in concise form in Appendix II.

Hemmerle [5] gives an explicit solution for the limiting $\hat{\beta}_j^*$ values, so that the above iteration is not necessary. The basic steps to this procedure are in Appendix III.

2.2 - Theory of Ridge Regression

The relationship of a ridge estimate to a least squares estimate is given by;

$$\hat{\beta}^* = [I_p + k(X'X)^{-1}]^{-1}\hat{\beta} \qquad (2.3)$$
$$= Z\hat{\beta}$$
$$= Z(X'X)^{-1}X'Y \qquad (2.4)$$

Thus, the ridge estimator is a linear transformation of $\hat{\beta}$. Clearly, $\hat{\beta}^*$ is biased,

$$E(\hat{\beta}^*) = E[Z(X'X)^{-1}X'Y] = Z(X'X)^{-1}X'E(Y)$$
$$= Z(X'X)^{-1}X'X\beta$$
$$= Z\beta \qquad (2.5)$$

The variance of $\hat{\beta}^*$ is given by;

$$\mathrm{Var}(\hat{\beta}^*) = Z(X'X)^{-1}X'\mathrm{Var}[Y]X(X'X)^{-1}Z'$$
$$= \sigma^2 Z(X'X)^{-1}Z' \qquad (2.6)$$

Other important relationships given by Hoerl and Kennard [6] are;

(i) $$\xi_i(W) = 1/(\lambda_i + k) \qquad (2.7)$$
$$\xi_i(Z) = \lambda_i/(\lambda_i + k) \qquad (2.8)$$

where $\xi_i(W)$ and $\xi_i(Z)$ are the ith characteristic roots of W and Z, respectively, and the λ_i are the characteristic roots of X'X.

(ii) $$(\hat{\beta}^*)'(\hat{\beta}^*) < \hat{\beta}'\hat{\beta} \qquad (2.9)$$

ie. $\hat{\beta}^*$ is shorter than $\hat{\beta}$ for $k \neq 0$. This result is based on equation (2.5) and the fact that Z is symmetric positive definite.

(iii) Let B be any estimate of β. Then the residual sums of squares becomes;

$$\begin{aligned}\Phi &= (Y - XB)'(Y - XB)\\ &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (B - \hat{\beta})'X'X(B - \hat{\beta})\\ &= \Phi(\hat{\beta}) + \Phi(B) \qquad (2.10)\end{aligned}$$

Clearly (2.10) is the residual sums of squares for $\hat{\beta}$, when $B = \hat{\beta}$. But the expected value of the squared length of $\hat{\beta}$ is too long when $X'X$ is close to being singular. The worse the conditioning of $X'X$, the farther we can get from $\hat{\beta}$ without an appreciable increase in the residual sum of squares. Therefore, by bounding the length of the estimate B (this works well in practice since $\beta'\beta$ never becomes infinite), the estimate of β that minimizes the sums of squares is $\hat{\beta}^*$.

This is shown by minimizing $B'B$ subject to $(B - \hat{\beta})X'X(B - \hat{\beta}) = \phi_0$ (2.11)

Solving this with LaGrange Multipliers yeilds the ridge estimator $\hat{\beta}^*$ defined in (2.1).

The most important theoretical aspect of the ridge estimator as far as its advantages over least squares is the Mean Square Error. Define the MSE as;

$$\begin{aligned}E[L_1^2(k)] &= E[(\hat{\beta}^* - \hat{\beta})'(\beta^* - \hat{\beta})] \qquad (2.12)\\ &= E[(\hat{\beta} - \beta)'Z'Z(\hat{\beta} - \beta)] + (Z\beta - \beta)'(Z\beta - \beta)\\ &= \sigma^2 \text{trace}(X'X)^{-1}Z'Z + \beta'(Z - I)'(Z - I)\beta\\ &= \sigma^2[\text{trace}(X'X + kI)^{-1} - k\ \text{trace}(X'X + kI)^{-2}]\\ &\qquad + k^2\beta'(X'X + kI)^{-2}\beta\\ &= \sigma^2 \sum_{i=1}^{p} \lambda_i/(\lambda_i + k)^2 + k^2\beta'(X'X + kI)^{-2}\beta \qquad (2.13)\\ &= f_1(k) + f_2(k) \qquad (2.14)\end{aligned}$$

If this mean square error is to be better when $k>0$, it must be smaller than (1.2). This is in fact possible, but first let us establish the properties and meanings of $f_1(k)$ and $f_2(k)$ individually.

A. Properties of $f_1(k)$

$f_1(k)$ is the sum of the variances (total variance) of the parameter estimates, whether biased or unbiased. Some of the attributes of this function are (all theorems and corollaries in this chapter are taken from [6]);

Theorem 2.1 The total variance $f_1(k)$ is a continuous, monotonically decreasing function of k.

Corollary 2.1.1 The first derivative with respect to k of the total variance, $f_1'(k)$, approaches $-\infty$ as $k \to 0^+$ and some $\lambda_i \to 0$.

Note that as the X'X matrix becomes singular, at least one of the $\lambda_i \to 0$. From equation (2.13), one can see this causes the total variance to increase without bound. The theorem and corollary above say, however, that $f_1(k)$ is a decreasing function of k. Moreover, values of k close to zero make it a sharply decreasing function. Hence, by adding a small positive constant to the diagonal of X'X, we will drastically reduce the variance of the parameter estimates.

B. Properties of $f_2(k)$

$f_2(k) = 0$ when $k=0$ and is a positive quantity for values of $k>0$. Thus, $f_2(k)$ is the square of the bias injected into the system by using $\hat{\beta}^*$ instead of $\hat{\beta}$. The important properties of $f_2(k)$ are [6];

Theorem 2.2 The squared bias $f_2(k)$ is a continuous, monotonically increasing function of k.

Corollary 2.2.1 The upper limit of $f_2(k)$ is $\beta'\beta$.

Corollary 2.2.2 The slope of $f_2(k)$ approaches zero as $k \to 0^+$.

Corollary 2.2.2 is very important since it says that the slope of $f_2(k)$ in the neighborhood of the origin is essentially zero. Coupling this with corollary 2.1.1, it becomes clear that by introducing a little bias thus greatly decreasing the variance, we should be able to find, for small values of k, corresponding values of the mean square error which are less than for unbiased estimates.

C. Properties of MSE

The foregoing analysis in A. and B. is linked together in the following theorem.

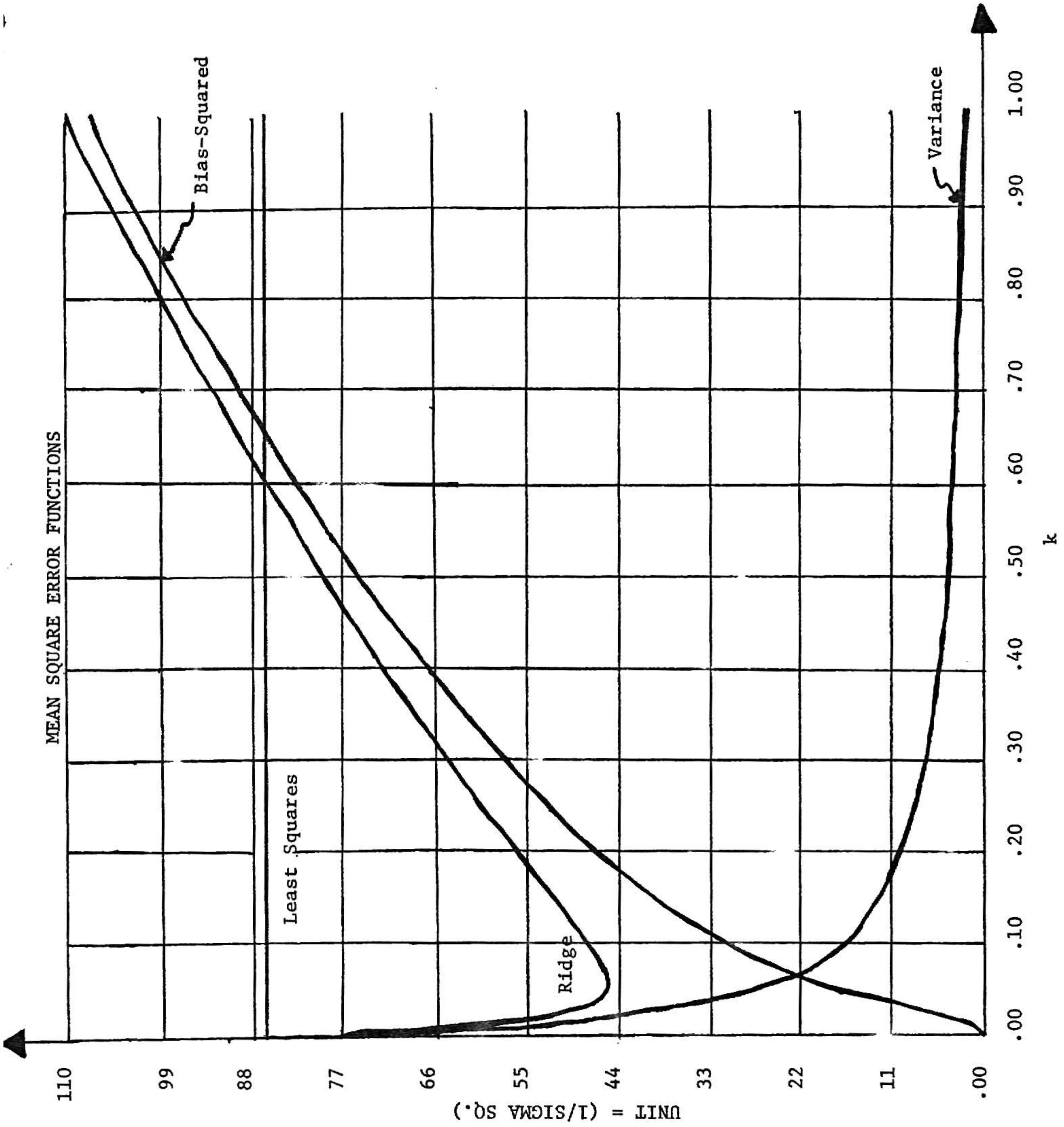
Theorem 2.3 (Existence Theorem) There always exists a $k>0$ such that

$$E[L_1^2(k)] \leq E[L_1^2(0)] = \sigma^2 \Sigma_{i=1}^{p}(1/\lambda_i) \qquad (2.15)$$

Therefore, $E[L_1^2(k)]$ will indeed go through a minimum. It is possible to find a value(s) for k that will yield a $\hat{\beta}^*$ closer to β than $\hat{\beta}$.

Hoerl and Kennard [6] graph the functional form of $f_1(k)$, $f_2(k)$ and $E[L_1^2(k)]$ presented in Figure I. Notice how $f_1(k)$ and $f_2(k)$ decrease and increase monotonically, respectively. The sum of $f_1(k)$ and $f_2(k)$ yields the mean square error, which is less than that for the least squares estimate when $0 \leq k \leq 0.6$. If the mean square error is to be the decision criterion for choosing the best estimates of β, this rules out the least squares estimates.

Figure I


MEAN SQUARE ERROR FUNCTIONS
Bias-Squared
Variance
Least Squares
Ridge
k
.00
.10
.20
.30
.40
.50
.60
.70
.80
.90
1.00
UNIT = (1/SIGMA SQ.)
110
99
88
77
66
55
44
33
22
11
.00

Another theorem important to the theory of ridge estimators is [8];

<u>Theorem 2.4</u> Let $(X'X + kI)\hat{\beta}^* = X'Y = g$ (2.16) and γ_k be the angle between $\hat{\beta}^*$ and g (note that g is the gradient vector of $\Phi(\beta)$). Then;

(i) γ_k is a continuous monotone decreasing function of k, such that as $k \to \infty$, $\gamma_k \to 0$. Since g is independent of k, it follows that,

(ii) $\hat{\beta}^*$ rotates toward g as $k \to \infty$.

A final property of the ridge estimator (useful primarily for comparison with other estimators) is given in the following theorem [6];

<u>Theorem 2.5</u> The ridge estimator is equivalent to a least squares estimator when the X matrix is augmented by H_k, where H_k consists of an orthogonal set of fictitious data points. The response Y is set to zero for each of these supplementary data points.

Proof: Augment X with H_k, the least squares normal equations are;

$$(X' \vdots H_k') \begin{bmatrix} X \\ \cdots \\ H_k \end{bmatrix} \hat{\beta}^* = (X' \vdots H_k') \begin{bmatrix} Y \\ \cdots \\ \underline{0} \end{bmatrix}$$

implies $(X'X + H_k'H_k)\hat{\beta}^* = X'Y$ (2.17)

Since H_k is orthogonal, $H_k'H_k$ is a scalar multiple of I_p; for any value k, the matrix can always be scaled such that

$H_k'H_k = kI_p$. This implies that (2.17) is equivalent to (2.16), and the proof is completed.

More will be said about this in Chapter 6.

CHAPTER 3

GENERALIZED INVERSE

3.1 - Method

The form of the generalized inverse estimator proposed by Marquardt [8] is;

$$\hat{\beta}^{+} = T_r^{+}X'Y \qquad (3.1)$$

where; (i) $T_r^{+} = \sum_{j=1}^{p}(1/\lambda_j)S_jS_j'$ (3.2)

(ii) $T = X'X$ (3.3)

(iii) $S'TS = D$ is an orthogonal transformation diagonalizing T. (3.4)

(iv) The matrix D has diagonal elements $\lambda_1 \geq \lambda_2 \quad . \; . \; . \geq \lambda_p$, which are the characteristic roots of T.

(v) S_j is the characteristic vector of T corresponding to λ_j.

(vi) r is the rank of T which we assume does not necessarily have to be an integer.

Marquardt [8] groups the characteristic roots of X'X into three types;

(a) substantially greater than zero,

(b) slightly greater than zero, and

(c) precisely zero (except for rounding error).

Computationally, it is often difficult to distinguish between the three types. An entire range of values, from large to near-zero or zero, may be observed.

With X'X near-singular, it is tempting to invert X'X by means of a generalized inverse. Since analysis of the characteristic roots suggests there is no one "rank" that can clearly be assigned to X'X, rather a range of reasonable choices, the generalized inverse is computed for various assigned ranks, r, in this range. It can be shown that T_r^{+} is indeed a generalized inverse of rank r if A is of rank r. The researcher attempts to find a value of r that is "optimum" by some criterion, such as small mean square error. MSE is defined $E[(B - \beta)'(B - \beta)]$, where B is any estimator of the true parameter β. The actual value of β is unknown, making the MSE a value which cannot be calculated.

Since MSE is not computable, some other criterion must be established for actual data analysis. Most biased estimation procedures utilize small coefficient vector length in conjunction with small residual sums of squares for this criterion. Marquardt [8] uses the "Variance Inflation Factors". These are given by the diagonal elements of $(X'X)^{-1}$ when X is scaled such that X'X is in the form of a correlation matrix, as we assume. Since these elements are proportional to the variances of the parameter estimates (see (1.8)), they are the factors by which the variances of the respective parameters are increased, due only to the correlation among the independent variables. In problems where $p>2$, attention is focused on the largest parameter variance inflation factor.

Hence the recommended procedure is;

(i) Find the characteristic root - $\lambda_1, \lambda_2, \ldots, \lambda_p$ of X'X. From these, determine feasible limits for the range of r, i.e., the rank of X'X.

(ii) Calculate the following values and tabulate for various incremented values of r in the interval selected in (i) above (reasonable increments for r depend on the researcher's time and cost considerations. Increments of the order of 0.1 - 0.5 are generally useful.);

a. $\hat{\beta}^+$, the vector of parameter estimates as given in (3.1) and (3.2). To use (3.2),include all terms for which j is less than or equal to the integer part of r, plus that fraction of the next term by which r exceeds its integer part.

b. $||\hat{\beta}^{+}||$, the length of $\hat{\beta}^{+}$.

c. $$\Phi = \sum_{i=1}^{p}(y_i - \hat{y}_i)^2 \qquad (3.5)$$
the residual sums of squares for the generalized inverse estimate.

d. The largest variance inflation factor (the largest diagonal element of T_r^{+}).

(iii) Examine the behavior of $||\hat{\beta}^{+}||$, Φ, and the variance inflation factor for the different values of r. Selection of the "best" set of estimates should involve the following considerations;

a. $||\hat{\beta}^{+}||$ should decrease significantly.

b. Small increases in Φ will be tolerated, but in some cases a point is reached where a decrease in r creates a much larger jump in Φ than preceding increments of r. This is a good point to consider for the estimate.

c. A rule of thumb [8] for choosing the amount of bias to allow with ill-conditioned data, is that the largest variance inflation factor usually should be larqer than 1.0 but certainly not as large as 10. Maximum variance inflation factors less than 1.0 tend to indicate a bit too much suppression of the parameter variance, and a somewhat larger value of r is desirable.

3.2 - Theory of Generalized Inverses

As noted in Section 3.1, the formulation of the generalized inverse estimator depends greatly on the assumption that a range of possible ranks r for X'X exist; and that a different g-inverse of X'X can be computed for each r, (T_r^+), and hence a different estimate for each r. A sketch of the background for these ideas is presented, followed by a look at the properties of the g-inverse estimator.

A. Development

Since S'S = I, equation (3.4) can be rewritten

$$T^{-1} = SD^{-1}S' \tag{3.6}$$

If rank(T) = r, and the last (p-r) ordered elements of D are zero (or close to zero, when T is "near-singular"), S and D can be partitioned;

$$S = (S_r : S_{p-r}) \tag{3.7}$$

where S_r is [pxr] and S_{p-r} is [px(p-r)]

$$D = \begin{bmatrix} D_r & : & \underline{0} \\ \cdots & & \cdots \\ \underline{0} & : & D_{p-r} \end{bmatrix} \tag{3.8}$$

where D_r is [rxr] and D_{p-r} is[(p-r)x(p-r)].

We are supposing D_{p-r} is zero (or close to it), so that $D_{p-r}^{-1} = 0$ by definition. Thus the inverse is

$$T_r^+ = S_r D_r^{-1} S_r' \tag{3.9}$$

which can be rewritten in the form given in equation (3.2).

Marquardt [8] notes that the trace is preserved by the orthonormal transformation (3.4), and that λ_j represents the sum of squares of projections of the points depicted by the rows of X onto the characteristic vector associated with the jth (ordered) characteristic root. This leads to the suggestion of a criterion that the assigned rank r include "substantially all" of the variation in the points of X. The criterion is that one select the smallest r such that

$$\left| \frac{\sum_{j=p}^{r} \lambda_j}{\text{trace } D} \right| < \omega \qquad (3.10)$$

where ω is selected in the interval $10^{-7} \leq \omega \leq 10^{-1}$, usually 10^{-5}. Note the summation of the λ_j in reverse order, which minimizes rounding error. The inverse given by (3.9) spans only the subspace spanned by S_r.

B. Properties of the g-inverse estimator

The estimate $\hat{\beta}^+$ is a linear transformation of $\hat{\beta}$, depending only on X and r.

$$\begin{aligned} \hat{\beta}^+ &= S_r D_r^{-1} S_r' X' Y \\ &= S_r D_r^{-1} S_r' (X'X) \hat{\beta} \qquad (3.11) \\ &= Z_r \hat{\beta} \end{aligned}$$

It follows that $\hat{\beta}^+$ is a biased estimator of β. The variance of $\hat{\beta}^+$ is given by

$$\begin{aligned} \mathrm{Var}(\hat{\beta}^+) &= \sigma^2 [S_r D_r^{-1} S_r'](X'X)[S D_{rr}^{-1} S_r'] \\ &= \sigma^2 Z_r (X'X)^{-1} Z_r' \qquad (3.12) \\ &= \sigma^2 S_r D_r^{-1} S_r' \end{aligned}$$

The theorems and discussion remaining in this section from [8] shed more light on the properties of $\hat{\beta}^+$.

Theorem 3.1 $||\hat{\beta}^+||^2$ is a stepwise (or piecewise continuous) increasing function of r.

In other words, $\hat{\beta}^+$ is shorter than $\hat{\beta}$ for values of $r \leq p$. This is best seen from the following equality;

$$\begin{aligned} ||\hat{\beta}^+||^2 &= \hat{\beta}^{+\prime}\hat{\beta}^+ \\ &= \sum_{u=1}^{r}\lambda_u^{-2}[\sum_{j=1}^{p}(g_j S_{ju})]^2 \end{aligned} \tag{3.13}$$

The uth term in this summation is the increase in the squared length of $\hat{\beta}^+$ due to including the uth characteristic vector dimension. Since the λ_u is raised to a negative power, the length of $\hat{\beta}^+$ increases greatly when dimensions are included for which λ_u is small.

Theorem 3.2 Let $\hat{\beta}^+$ be the estimator arrived at by equations (3.1) and (3.2). Then $\hat{\beta}^+$ minimizes the residual sum of squares

$$\Phi(\hat{\beta}) = (Y - X\hat{\beta})'(Y - X\hat{\beta}) \tag{3.14}$$

for all $\hat{\beta}$ within the r-dimensional subspace spanned by S_r.

Theorem 3.3 The mean square error of $\hat{\beta}^+$ is

$$E(L_1^2) = tr[Var(\hat{\beta}^+)] + \beta'(Z_r - I)'(Z_r - I)\beta \tag{3.15}$$

The second term on the right side of (3.15)is the squared bias. It is zero when $r=p$.

Corollary 3.3.1 The variance term in (3.15) is an increasing function of r.

Corollary 3.3.2 The bias term in (3.15) is a monotonic decreasing function of r.

In practice, we will decrease r according to the extent of ill-conditioning in the X'X matrix. The above corollaries indicate that as we decrease r, the variance of $\hat{\beta}^+$ will decrease, while the amount of bias increases. We would like to be able to find a value for r such that the mean square error $E(L_1^2)$ is less than the least squares mean square error.

As in ridge regression, we will tolerate some bias in the estimator in hopes of reducing the mean square error.

Theorem 3.4 A sufficient condition for the mean square error $E(L_1^2)$ to be less than the OLS mean square error is

$$\sum_{j=r=1}^{p} 1/\lambda_j > (1/\sigma^2)(\beta'\beta) \tag{3.16}$$

Hence it is possible to find an r that will yield a $\hat{\beta}^+$ closer to β than $\hat{\beta}$ Note that β must be bounded, which we said presents no problem in actual data analysis.

Theorem 3.5 $\hat{\beta}^+$ is equivalent to $\hat{\beta}$ if one of the following is true;

(i) D_{p-r} is a null matrix.

(ii) D_{p-r} is not precisely a null matrix.

In (ii), $\hat{\beta}^+$ is equal to $\hat{\beta}$ when the actual data are supplemented by a fictitious set of data points taken according to an experiment H_r; with the response Y equal to zero for each supplementary data point. The proof of (ii) is shown by finding H^r such that

$$(X'X + H_r'H_r)^{-1} = S_r D_r^{-1} S_r' \tag{3.17}$$

The matrix H_r plays the role of prior information from the standpoint of a Bayesian interpretation.

The following theorem aids in the geometric interpretation of the g-inverse technique.

Theorem 3.6 Let γ_r be the angle between $\hat{\beta}^+$ and g. If r is an integer and λ_r and λ_{r-1} satisfy;

(i) $0 < \lambda_r$

(ii) $\lambda_r/\lambda_{r-1} << 1$

(iii) $\lambda_{r-1}/\lambda_{r-2} << 1$

then $\gamma_r \geq \gamma_{r-1}$.

Since g is indpendent of $\hat{\beta}^+$, it follows that $\hat{\beta}^+$ rotates toward g as r is decreased.

CHAPTER 4

SHRUNKEN ESTIMATORS

4.1 - Method

Mayer and Willke [9] view the ridge estimators as a subclass of the class of linear transforms of the least squares estimator $\hat{\beta}$. They propose the shrunken estimators as an alternative to either the least squares or ridge estimators in the case of an ill-conditioned X'X matrix.

The general form of the estimator is given by

$$\underline{c}_\alpha = \alpha[(X'X)^{-1}X'Y] = \alpha\hat{\beta} \qquad ; \alpha=[0,\infty) \tag{4.1}$$

where α is called the shrinkage factor. If α is a fixed scalar, $\underline{c}_\alpha$ is called a deterministically shrunken estimator. But if α is a scalar function of $\hat{\beta}'\hat{\beta}$, (i.e. $\alpha = f(\hat{\beta}'\hat{\beta})$), then $\underline{c}_\alpha$ is called a stochastically shrunken estimator.

Thus, the shrunken estimator simply involves multiplying the least squares estimator by some shrinkage factor. The problem arises in choosing the proper shrinkage factor. Three methods of selecting a factor are presented below.

A. The Deterministically Shrunken Estimator $\underline{c}_\alpha$

(i) Calculate the least squares estimator $\hat{\beta}$.

(ii) Compute $\underline{c}_\alpha$ from formula (4.1) using different values for α in the interval [0,1], possibly in increments of 0.1.

(iii) Plotting each $c_{\alpha i}$ as a function of α yields a straight line from the origin to the least squares estimator, hence the plot does not stabilize as in the ridge trace. Due to this lack of stabilization, Mayer and Willke [9] recommend one of the other two estimators below if a shrunken estimator is to be used.

B. The Stochastically Shrunken Estimator $\underline{d}_\delta$

(i) Calculate $\hat{\beta}$.

(ii) Evaluate $\underline{d}_\delta = \delta\hat{\beta}\hat{\beta}'(I + \delta\hat{\beta}\hat{\beta}')^{-1}\hat{\beta}$ (4.2)

for values of δ in the interval [0,1], selected at say, increments of 0.1.

(iii) Plot the $d_{\delta i}$ as a function of δ as in (ii) for ridge regression. This plot will stabilize, and hence δ can be chosen in the same manner k is chosen for the ridge estimator.

C. The Stochastically Shrunken Estimator $\underline{e}_\xi$

(i) Calculate $\hat{\beta}$.

(ii) Compute $\xi_0 = (p - 2)(n - p + 2)^{-1}$ (4.3)

(iii) Evaluate $\underline{e}_{\xi_0} = [1 - \xi_0 s^2(\hat{\beta}'\hat{\beta})^{-1}]\hat{\beta}$ (4.4)

where $s^2 = Y'Y - \hat{\beta}'X'X\hat{\beta}$

(iv) Each $e_{\xi_{0i}}$ is a constant value, so that a plot sheds no new light on the situation.

4.2 - Theory of Shrunken Estimators

The justification for considering shrunken estimators as an alternative to either a least squares estimator or any of the preceeding techniques is four-fold. All Definitions and Propositions in this chapter are taken from [9].

(i) The shrunken estimators satisfy an existence condition of small mean square error similar to that given in Theorem 2.3 to justify the ridge estimator. Consider the following proposition

<u>Proposition 1</u> For every β there exists a fixed α in $[0,1]$ such that $E[L_1^2(\underline{c}_\alpha)] < f_1(\hat{\beta})$, where $E[L_1^2]$ is the mean square error and f_1 is the total variance. (i.e. $f_1(\hat{\beta}) = tr[Var(\hat{\beta})]$)

Thus, it is possible to choose an α such that the mean square error of the shrunken estimator is smaller than the total variance of the least squares estimator.

(ii) Begin with two definitions;

<u>Definition 1</u> Let C denote the class of linear transforms of $\hat{\beta}$ such that if $\underline{b} \in C$ then $b = G\hat{\beta}$ for some pxp matrix G.

<u>Definition 2</u> Let $C(\tau)$ denote the subclass of C such that $b(G) \in C(\tau)$ iff $\hat{\beta}'(G - I)(X'X)(G - I)\hat{\beta} = \tau$.

Proposition 2 gives the second justification for the use of shrunken estimators.

<u>Proposition 2</u> The shrunken estimator has the shortest length of all estimators in the equivalence class $C(\tau)$, provided $m_d(G)$ is the norm used to measure length, where

$$m_d(G) = \hat{\beta}'G'(X'X)G\hat{\beta} \tag{4.5}$$

Mayer and Willke refer to $m_d(G)$ as the design-dependent norm, and note that it imposes constraints on the bias of the estimator based on the observed (ill-conditioned) data. Hence if the scientist feels that $m_d(G)$ is a poor norm to use, Proposition 2 is little evidence for use of the shrunken estimator. This is dealt with further in Chapter 6.

(iii) <u>Proposition 3</u> Let $G_1 = \delta\hat{\beta}\hat{\beta}'(I + \delta\hat{\beta}\hat{\beta}')^{-1}$ for some δ, if $b(G_1) \in C(\tau)$

$$f_1(b(G_1)) = \min_{C(\tau)} f_1(b(G))$$

So the shrunken estimators d_δ are minimum total variance estimators within the equivalence class $C(\tau)$.

(iv) <u>Definition 3</u> Let $W(B) = E[(B - \beta)'(X'X)(B - \beta)]$ denote the weighted total mean square error of estimator B.

<u>Proposition 4</u> If $p \geq 3$ and $0 < \xi < 2(p-2)(n-p+2)^{-1}$, then

$W(\underline{e}_\xi) < W(\hat{\beta})$ and if $\xi_o = (p-2)(n-p+2)^{-1}$,

then $W(\underline{e}_{\xi o}) = \min_{\xi} W(\underline{e}_\xi)$.

Therefore the stochastically shrunken estimator $\underline{e}_{\xi o}$ calculated as in (4.4) has smaller weighted mean square error than $\hat{\beta}$.

The moments for the deterministically shrunken estimator $\underline{c}_\alpha$ are given [9] to be;

$$\mathrm{Var}(\underline{c}_\alpha) = \alpha^2\sigma^2(X'X)^{-1} \tag{4.6}$$

$$f_1(\underline{c}_\alpha) = \alpha^2\sigma^2\mathrm{tr}[(X'X)^{-1}] \tag{4.7}$$

$$E[L_1^2(\underline{c}_\alpha)] = (1 - \alpha)^2\beta'\beta \tag{4.8}$$

Mayer and Willke have had little success in determining a general form of the moments for the stochastically shrunken estimators $\underline{d}_\delta$ and $\underline{e}_\xi$. It is also not clear how to choose the particular α, δ, or ξ which minimizes the total mean square error of the estimator.

CHAPTER 5

LATENT ROOT REGRESSION ANALYSIS

5.1 - Method

This technique, as presented by Webster, Gunst and Mason [12], is a modified least squares procedure utilizing the latent (characteristic) roots and latent vectors of the correlation matrix including both the dependent and independent variables. The reason the dependent variable is included in the matrix is that a geometric interpretation of the first element of each latent vector provides a measure of that latent vector's predictive value. If a latent root and the magnitude of the first element of the corresponding latent vector are both small, the latent vector is said to reveal a "non-predictive near singularity". This geometric interpretation will be pursued in greater detail in the following section.

After the presence of ill-conditioning is ascertained, prediction equations are obtained from linear combinations of the latent vectors. If non-predictive near-singularities are not present,a linear combination of all the latent vectors can be used to obtain the OLS predictor. If non-predictive near singularities are found, the latent vectors which reveal them are removed from the estimation procedure and the linear combination of the remaining latent vectors which minimizes the residual sums of squares is used as a predictor.

The form of the estimator is [12];

$$b^* = -\eta(\sum_{i=t}^{p}\lambda_i^{*-1})^{-1}\sum_{j=t}^{p}\gamma_{0j}\lambda_j^{-1}\underline{\gamma}_j^0 \qquad (5.1)$$

The terms in the above equation will be defined as we move through the procedure Latent root regression analysis can be applied as follows;

(i) Form the matrix $A = [Y^* \vdots X]$ where $Y^* = (Y_i - \bar{Y}) / \eta$

and $\eta^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$. Thus, A'A is the [(p+1)x(p+1)] "correlation matrix" of dependent and independent variables.

(ii) Find the latent roots and latent vectors of A'A. Denote the jth latent root by λj and the jth latent vector by

$$\underline{\gamma}_j' = [\gamma_{0j}, \gamma_{1j}, \cdot \cdot \cdot, \gamma_{pj}] \quad (5.2)$$

Let $\underline{\gamma}_j^{0'} = [\gamma_{1j}, \gamma_{2j}, \cdot \cdot \cdot, \gamma_{pj}]$ (5.3)

contain all the elements of $\underline{\gamma}_j$ except the first one. Finally,

$$\lambda_j^* = \lambda_j / \gamma_{oj}^2 \quad (5.4)$$

(iii) Construct a table containing all the λ_j and the $|\gamma_{oj}|$. Recall that the criterion for identifying non-predictive near singularities is small values of both λ_j and $|\gamma_{oj}|$. Now we must decide what to call small. Webster finds latent roots of size 0.05 or smaller to be fairly reliable indicators of a near singularity. On the other hand, latent roots as large as 0.10 or 0.20 and larger are seldom reliable. A cutoff value which works well for $|\gamma_{oj}|$ is $|\gamma_{oj}| < 0.10$.

For purposes of notation, assume the λ_j and $|\gamma_{oj}|$ have been arranged as follows;

$$\lambda_o \leq \lambda_1 \leq \cdot \cdot \cdot \leq \lambda_{t-1} \leq \lambda_t \leq \cdot \cdot \cdot \leq \lambda_p$$

$$|\gamma_{oo}| \leq |\gamma_{o1}| \leq \cdot \cdot \cdot \leq |\gamma_{ot-1}| \leq |\gamma_{ot}| \leq \cdot \cdot \cdot \leq |\gamma_{op}|$$

Where t-1 is the subscript of λ and $|\gamma_o|$ such that both have been determined "small" according to some criterion similar to the above.

(iv) Upon deciding which latent vectors contain non-predictive near singularities,eliminate these latent vectors and corresponding latent roots from the analysis, and use (5.1) to find b^*.

Note that using all latent roots and latent vectors gives

$$b = - \eta(\sum_{i=0}^{p} \lambda_i^{*-1})^{-1} \sum_{j=o}^{p} \gamma_{oj} \lambda_j^{-1} \underline{\gamma}_j^o \qquad (5.5)$$

which is the LS estimator.

Webster, Gunst and Mason utilize this analysis to formulate two backward elimination procedures. One such procedure is based on the idea that by using only latent vectors not indicating non-predictive near singularities, the true influences of the independent variable on the dependent var.ables are more clearly represented. In many cases several independent variable may be eliminated at the first stage when the computations are easiest. The details of these procedures are beyond the scope of this report, but may be found in [12].

5.2 - Theory of Latent Root Regression Analysis

Part A. of this section traces the main steps in the development of the latent root estimator [12]. Part B. deals with the geometrical interpretation of the analysis. The ensuing analysis in greater detail is in [12].

A. Development

An estimator of the form

$$Y = b_o\underline{1} + X\underline{b} \tag{5.6}$$

is desired, where b_o and $\underline{b}$ estimate β_o and $\underline{\beta}$ respectively. As in all the methods, we let $b_o = \bar{Y}$, since the X - matrix contains standardized independent variables.

Begin with

$$A\underline{\gamma}_j = \begin{bmatrix} Y_1^*\gamma_{oj} + \sum_{r=1}^{p} X_{1r}\gamma_{rj} \\ Y_2^*\gamma_{oj} + \sum_{r=1}^{p} X_{2r}\gamma_{rj} \\ \vdots \\ Y_n^*\gamma_{oj} + \sum_{r=1}^{p} X_{nr}\gamma_{rj} \end{bmatrix} \tag{5.7}$$

From (5.7), define p+1 prediction equations

$$\hat{\underline{Y}}_j = \bar{Y}\underline{1} - \eta\gamma_{oj}^{-1}X\underline{\gamma}_j^{o} \qquad ; \; j = 0,1, \ldots, p \tag{5.8}$$

where all the $\gamma_{oj} \neq 0$. This is clearly of the form (5.6) with

$$\underline{b} = -\eta\gamma_{oj}^{-1}\underline{\gamma}_j^{o}.$$

Usually none of the prediction equations alone yield a good predictor. Hence, linear combinations of these predictors are used to obtain estimates of the parameters of the model. Webster intoduces an arbitrary linear combination of (5.8)

$$\hat{Y} = \sum_{j=0}^{p} a_j \gamma_{oj} \hat{Y}_j \tag{5.9}$$

with the restriction $\sum_{j=0}^{p} a_j \gamma_{oj} = 1$. This implies

$$\hat{Y} = \bar{Y}\underline{1} - \eta X(\sum_{j=0}^{p} a_j \underline{\gamma}_j^o) \tag{5.10}$$

which is also of the form (5.6) with

$$\underline{b} = -\eta \sum_{j=0}^{p} a_j \underline{\gamma}_j^o \tag{5.11}$$

The residual sum of squares is

$$\Phi = (Y - \hat{Y})'(Y - \hat{Y}) = \eta^2 \underline{a}'\Lambda\underline{a} = \eta^2 \sum_{j=0}^{p} a_j^2 \lambda_j \tag{5.12}$$

where $\underline{a}' = (a_o, a_1, . . ., a_p)$ and $\Lambda = \text{diag.}(\lambda_o, \lambda_1, . . ., \lambda_p)$.

The problem is now one of minimizing the residual sum of squares subject to the constraint $\sum_{j=0}^{p} a_j\gamma_{oj} = 1$, and solving for $\underline{a}$. This yields the OLS predictor. Hence, we minimize

$$f(\underline{a}) = \eta^2 \sum_{j=o}^{p} a_j^2 \lambda_j - 2\mu_o(\sum_{j=o}^{p} a_j \gamma_{oj} - 1) \tag{5.13}$$

where $-2\mu_o$ is the LaGrangian multiplier. Solving for $\underline{a}$ in (5.12) yields

$$a_j = \gamma_{oj} \lambda_j^{-1}(\sum_{i=0}^{p} \lambda_i^{*-1})^{-1} \quad ; \; j=0, 1, . . ., p \tag{5.14}$$

where λ_i^* is given in (5.4). Substituting this in (5.11) gives the OLS coefficients

$$\underline{b} = -\eta(\sum_{i=0}^{p} \lambda_i^{*-1})^{-1} \sum_{j=o}^{p} \gamma_{oj} \lambda_j^{-1} \underline{\gamma}_j^o \tag{5.15}$$

which is the same as (5.5) with residual sums of squares

$$\Phi = \eta^2 (\sum_{i=0}^{p} \lambda_i^{*-1})^{-1} \tag{5.16}$$

When the latent vectors $\underline{\gamma}_o$, $\underline{\gamma}_1, . . ., \underline{\gamma}_{t-1}$ correspond to non-predictive near singularities, the latent root estimator is arrived at by setting $a_o = a_1 = . . . = a_{t-1} = 0$ and minimizing (5.13). The solution is

$$a_j = \gamma_{oj} \lambda_j^{-1} (\sum_{i=t}^{p} \lambda_i^{*-1})^{-1} \quad ; \quad j = t, t+1, . . ., p \qquad (5.17)$$

and the modified least squares coefficients are as given in (5.1), with residual sum of squares

$$\Phi = \eta^2 (\sum_{i=t}^{p} \lambda_i^{*-1})^{-1} \qquad (5.18)$$

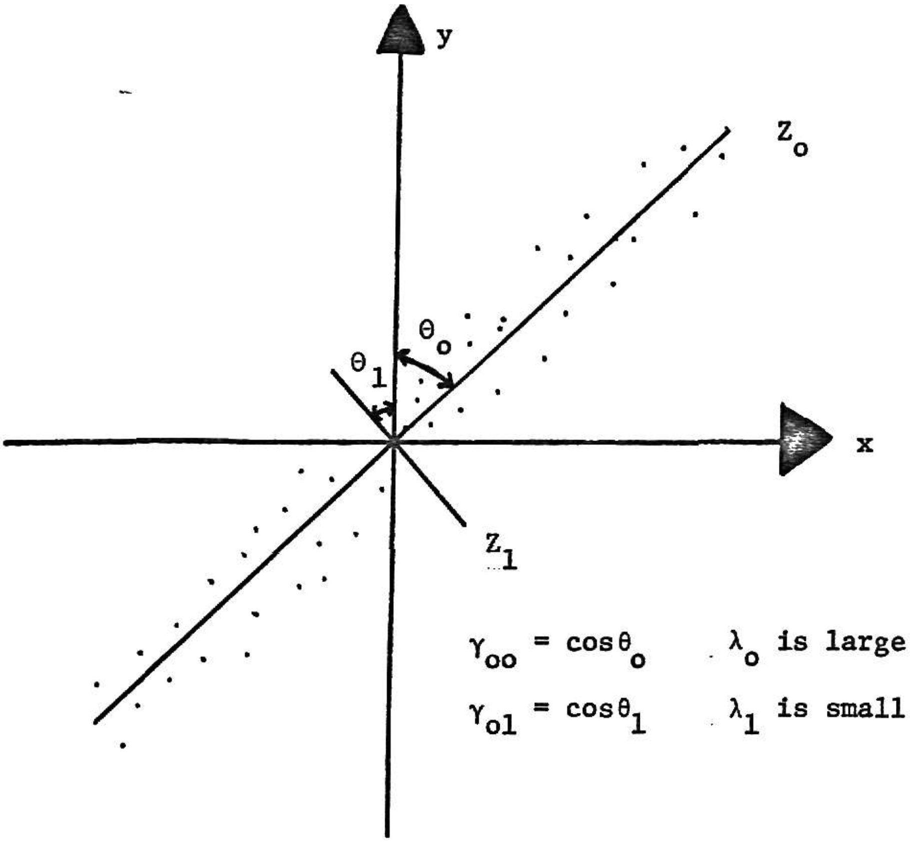
Solutions to the normal equations can be obtained from this procecure even if X'X is singular. This arises from the fact that a singular X'X matrix implies some of the λ_j and corresponding γ_{oj} are zero, the same as setting the appropriate $a_j = 0$ in (5.13).

B. Geometrical Interpretation

We are interested in determining whether the near-singularities in A'A contain information about the underlying model (1.1). Consider the n data points $(Y_i^*, X_{i1}, X_{i2}, . . ., X_{ip})$ i = 1, 2, . . ., n as n points in the p+1 dimensional Euclidean space defined by the mutually orthogonal axes Y^*, $X_1, X_2, . . ., X_p$. A new set of axes $Z_o, Z_1, . . ., Z_p$ are defined by $\underline{\gamma}_o, \underline{\gamma}_1, . . ., \underline{\gamma}_p$; the latent vectors of A'A. The first element, γ_{oj}, in each of these latent vectors represents the cosine of the angle between axes Y^* and Z_j. The other γ_{rj} (j = 1,2,...,p) in each latent vector represent the cosine of the angle between X_r and Z_j. A zero γ_{oj} means Y^* and Z_j are orthogonal. (cos 0° = 1 and cos 90° = 0)

Each λ_j represents the spread of the n data points in the direction of Z_j. ie. the latent root is the sum of squares of the projections of

FIGURE II



the n data points on the Z_j axis. Hence, a small λ_j indicates the data points are clustered tightly along the Z_j axis, little spread in that direction.

Therefore, if both λ_j and $|\gamma_{oj}|$ are small the latent vector $\underline{\gamma}_j$ reveals a non-predictive near-singularity - very little spread in a direction orthogonal to the dependent variable, which explains little or none of the variability in the dependent variable. On the other hand, a large λ_j and $|\gamma_{oj}|$ indicate much spread in a direction almost parallel to Y, and therefore a good vector for prediction.

Figure II illustrates the geometric ideas just given for the simple linear model

$$Y = \beta_o + \beta X + \varepsilon$$

CHAPTER 6

COMPARISON AND CONTRAST

The theory in Chapters 2 and 3 was presented in such a manner that the reader will note a close parallel in the characteristics and justification for both ridge regression (RR) and generalized inverses (g-inverses). For increasing values of k in RR and decreasing values of r in the case of g-inverses;

(i) the length of the vector of coefficients decreases.

(ii) $\hat{\beta}^*$ and $\hat{\beta}^+$ rotate toward g.

(iii) the total variance terms [$f_1(k)$and $f_1(r)$] decrease.

(iv) the bias terms [$f_2(k)$ and $f_2(r)$] increase.

As shown in theorems 2.5 and 3.5, both techniques correspond to Allen's[1] "Data Augmentation", a somewhat Bayesian approach to the problem of ill-conditioning.

Although the RR and g-inverse estimators share many desirable properties, the RR estimator is not a g-inverse estimator. For example the ridge inverse $[X'X + kI]^{-1}$ does not obey

$$T\,T^+\,T = T \qquad (6.1)$$

Marquardt views the ridge inverse as an approximate g-inverse.

RR, g-inverse and shrunken estimators all fall in a more general class of linear transforms of the least squares estimators. The arguments for all three of these estimators are based on the following two main points;

(1) a reduction in the length of the vector of estimated coefficients when compared with the length of the vector of LS coefficients, and

(2) existence theorems (2.3) and (3.4) and proposition 1 revealing classes of estimators with smaller total variance than the LS estimator.

Both the ridge estimators and the shrunken estimators are minimum length estimators with respect to the appropriate norms in the class C. The norm for the shrunken estimators is $m_d(G)$ given in (4.5) and the norm for ridge estimators is

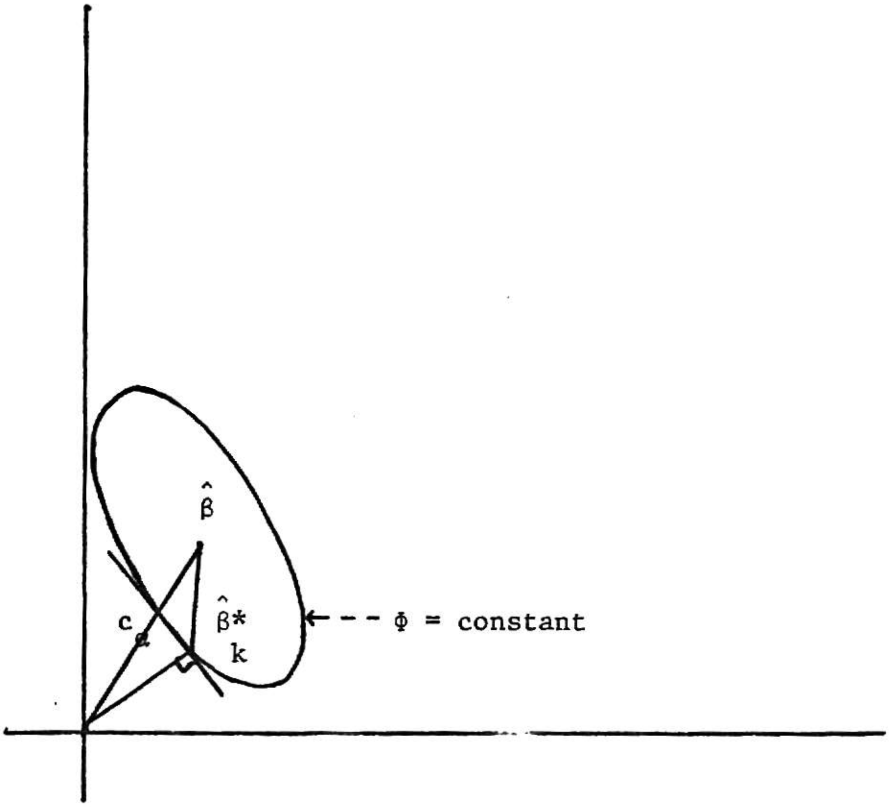
$$m(G) = \underline{\hat{\beta}}' \ G' \ G \ \underline{\hat{\beta}} \tag{6.2}$$

Mayer and Willke observe that the ridge estimator has minimum squared length among all estimators with a given sum of squares loss, but that the estimator $\underline{d}_\delta$ has minimum variance among those estimators with a given sum of squares loss which are in class C. (Proposition 3). They also contend that the estimator $\underline{e}_\xi$ is superior to the ridge estimators or deterministically shrunken estimators $\underline{c}_\alpha$ since a value of ξ can be determined which will guarantee an estimator with smaller weighted mean square error than $\underline{\hat{\beta}}$. (Proposition 4). Sclove [10] gives estimators guaranteed to have total MSE smaller than the total variance of $\underline{\hat{\beta}}$, but which are rather complex.

The difference between the ridge estimator and shrunken estimator can be seen by observing Figure III.[9] The shrunken estimator $\underline{c}_\alpha$ or $\underline{d}_\delta$ in a given equivalence class corresponds to the point on the elipse which falls on the line from the origin to the OLS estimator $\underline{\hat{\beta}}$. The ridge estimator,$\underline{\hat{\beta}}^*_k$, in a given equivalence class corresponds to the point on the ellipse which is closest to the origin in the Euclidean sense. Note that both of these estimators in a given equivalence class are shorter than the OLS estimator. The importance of this is seen upon recalling that the OLS estimator tends to exceed the actual parameter vector length when dealing with ill-condintioned data.

The estimators obtained from latent root regression analysis (LRRA) do not fall in the class of linear transforms of the LS estimator. The major difference between the OLS estimator and the LRRA estimator is the term containing $\underline{\gamma}_j^o$ for j = 0,1, . . ., t - 1 (See equations (5.1) and (^{5}.5)). Gunst, Webster and Mason[3] compare the relative merits of OLS and LRRA with respect to both estimation and variable selection with multicollinear data. They conclude that unless the parameter vector is parallel to the latent vector corresponding to the smallest latent root of X'X, LRRA is preferable to OLS both for estimation of parameters and variable selection. An argument for the usefulness of LRRA is that it provides measures [λ_j and $|\gamma_{oj}|$] for determining when a biased estimator should be used.

FIGURE III



Hawkins [4] compares LRRA with RR. Recall the form of the LRRA estimator

$$b^* = [-\eta_j \sum_{j=t}^{p} \gamma_{oj} \, \underline{\gamma}_j^o / \lambda_j] \; / \; [\sum_{j=t}^{p} \gamma_{oj}^2 / \lambda j] \qquad (6.3)$$

Hawkins shows that the RR estimator can be written

$$\hat{\beta}^* = [-\eta_j \sum_{j=t}^{p} \gamma_{oj} \underline{\gamma}_j^o / (\lambda_j + k)] \; / \; [\sum_{j=t}^{p} \gamma_{oj}^2 / (\lambda_j + k)] \quad (6.4)$$

Thus RR is a rescaling of the summation terms in (6.3) by a factor of $\lambda_j/(\lambda_j + k)$. (This is E(Z), where $\hat{\beta}^* = Z\,\hat{\beta}$.) Recall that an ill-conditioned X'X matrix is characterized by a few small λ_j which dominate the OLS estimator. These c. vectors are the ones which will most be affected by the diagonal magnification, and whose contribution to the estimation will fall away most rapidly. This is analogous to the relationship between RR and g-inverses.

The succeeding chapter provides further opportunities for comparisons and contrasts, as one set of ill-conditioned data is analyzed by each of the four biased techniques and by least squares.

CHAPTER 7

A NUMERICAL EXAMPLE

7.1 - Least Squares Solution

The data used for this illustration contains a great deal of multicolinearity among the independent variables. It does not pertain to any particular scientific phenomena. There are 15 observations on 9 independent variables X_1, X_2,. . ., X_9; and the dependent variable Y.

Table I gives the X'X matrix in its correlation form. The last row contains the correlations with Y and the independent variables used for the LRRA computations. Note the great amount of correlation among X_1, X_2, X_3 and X_4.

The least squares solution is;

$$\hat{\beta} = \begin{bmatrix} -3.3140 \\ -0.1952 \\ -0.1431 \\ 3.9319 \\ 0.0947 \\ 1.5639 \\ 11.4223 \\ 0.1272 \\ -0.0040 \end{bmatrix} \quad \text{with } \Phi(\hat{\beta}) = 20.5821 \quad \text{and } ||\hat{\beta}||^2 = 159.444$$

A stepwise deletion procedure, deleting the variable with the highest α-level greater than 0.05 leaves only variable 6 in the model with $\Phi = 41.6684$ and $R^2 = 0.65$.

TABLE I

Correlation Matrix

X_1	1									
X_2	0.91	1								
X_3	0.80	0.94	1							
X_4	0.98	0.94	0.88	1						
X_5	0.05	0.30	0.44	0.16	1					
X_6	-0.44	-0.66	-0.77	-0.54	-0.68	1				
X_7	0.07	-0.01	-0.13	0.01	-0.17	0.43	1			
X_8	-0.47	-0.63	-0.74	-0.59	-0.69	0.76	-0.02	1		
X_9	-0.43	-0.35	-0.25	-0.41	0.02	0.14	0.18	-0.11	1	
Y	-0.25	-0.44	-0.56	-0.32	-0.65	0.81	0.27	0.74	-0.19	1
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Y

7.2 - Ridge Regression Solution

The multicolinearity in the data is also revealed in the characteristic roots of X'X (ordered);

$\lambda_1 = 5.038$

$\lambda_2 = 1.732$

$\lambda_3 = 1.236$

$\lambda_4 = 0.607$

$\lambda_5 = 0.145$

$\lambda_6 = 0.131$

$\lambda_7 = 0.074$

$\lambda_8 = 0.031$

$\lambda_9 = 0.005$

The sum of the reciprocals of the characteristic roots is $\Sigma(1/\lambda_i) =$ 265.451. Thus (1.5) shows that the expected squared distance of the coefficient estimate, $\hat{\beta}$, from β is $265.451\sigma^2$, which is more than 26 times what it would be for a system with no near-singularities present.

Since the smallest characteristic root λ_9 is not zero, the factors define a 9-dimensional space. The first four characteristic roots total 8.614 so that most of the variation can probably be explained in four dimensions.

Figure IV is the ridge trace for this problem. This trace was constructed by computing a total of 16 regressions using the APL program in Appendix IV. The following observations can be made from the ridge trace:

(i) The coefficients from OLS are undoubtedly overestimated. They are not stable as a group. Moving a short distance from the least squares point k=0 shows a rapid decrease in absolute value in the coefficients of X_1, X_4 and X_7. When k=0.04, $||\hat{\beta}^*||^2$ is only 13.0% of its original value.

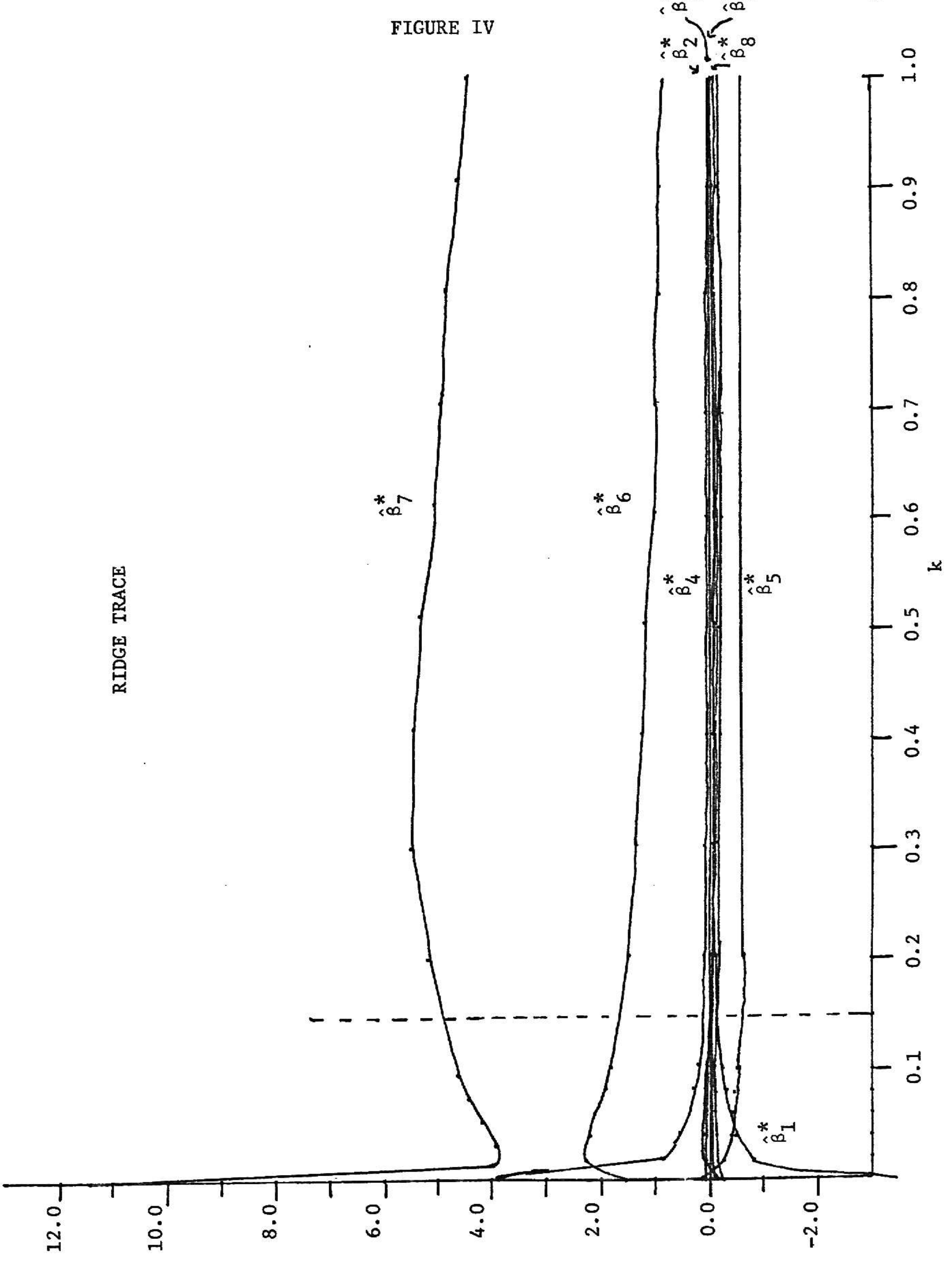
(ii) X_4 is the second largest positive factor and X_1 is the largest negative factor. Both are quickly driven to zero with the addition of k>0.

(iii)The correlations with other factors causes X_5 to be underestimated. At k=0, it is the smallest positive factor. The addition of k>0 drives the negative coefficients to zero and X_5 becomes the most important negative factor.

(iv) Factors X_1. X_2, X_3, X_4, X_8 and X_9 all appear to be overestimated and are driven toward zero.

(v) At a value of k in the interval (0.1,0.2) the system has stabilized and coefficients for k=0.15 (say) will most likely be closer to β and more stable for prediction than the OLS estimates.

FIGURE IV



RIDGE TRACE

The estimates at k=0.15 are;

$$\hat{\beta}^*(0.15) = \begin{bmatrix} -0.1467 \\ -0.0479 \\ -0.0005 \\ 0.1465 \\ -0.4680 \\ 1.6326 \\ 5.0233 \\ 0.0389 \\ -0.0094 \end{bmatrix} \qquad \text{with } \Phi(\hat{\beta}^*) = 35.4907 \quad \text{and } ||\hat{\beta}^*||^2 = 28.165$$

This represents a 72.4% increase in residual sum of squares over the least squares estimate and an 82.33% decrease in the squared length of the coefficient vector. The analyst would probably want to drop all variables except X_5, X_6, and X_7 from the model, using the coefficients obtained from k=0.15 fcr the the variables remaining in the model.

The Explicit Solution (Hemmerle [5]) in Appendix III was also used to calculate a Ridge Regression Solution. The results differ from the previous ridge solution. This procedure begins by checking the convergence/divergence criterion;

$$\hat{\mu} = \Lambda^{-1}SX'Y = \begin{bmatrix} -2.8918 \\ 0.4325 \\ 7.6527 \\ 35.3947 \\ -4.9067 \\ 2.0918 \\ -0.5569 \\ 3.0973 \\ -1.8643 \end{bmatrix} \quad \text{and} \quad \begin{aligned} q_{1(o)} &= 0.09770 \\ q_{2(o)} &= 709.76 \\ q_{3(o)} &= 0.04087 \\ q_{4(o)} &= 0.67058 \\ q_{5(o)} &= 0.09872 \\ q_{6(o)} &= 7.1592 \\ q_{7(o)} &= 10.736 \\ q_{8(o)} &= 2.9572 \\ q_{9(o)} &= 1.95089 \end{aligned}$$

$q_{1(o)}$ and $q_{5(o)}$ are both < 1/4, so they were used to compute $q_1^* = .123337$ and $q_5^* = .124949$.

Next, $\hat{\mu}_1^* = \dfrac{-2.8918}{1 + .123337} = -2.57425$

$$\hat{\mu}_5^* = \frac{-4.9067}{1 + .124949} = -4.3617$$

Finally,

$$S'\hat{\alpha}^* = \hat{\beta}^* = \begin{bmatrix} 0.0995 \\ -0.0428 \\ -0.0710 \\ 0.0246 \\ -0.8850 \\ 0.8388 \\ 6.5840 \\ 0.0350 \\ -0.0049 \end{bmatrix}$$

with $\Phi(\hat{\beta}^*) = 44.8278$
(a 117.8% increase over OLS)

and $||\hat{\beta}^*||^2 = 44.854$
(a 71.9% decrease from OLS)

This is a rather large increase in the residual sum of squares, which can be improved by performing additional modifications given in [5].

7.3 - Generalized Inverse Solution

The results of the generalized inverse method, obtained using VGINV in Appendix IV are presented in Table II. Twenty regressions were calculated in all. Note the rapid decrease in magnitude of both $||\hat{\beta}^+||^2$ and the variance inflation factor, with very little corresponding increase in the residual sum of squares. Using the rule of thumb (p. 17) for deciding how much bias to allow narrows the range of selection for r to the interval [4.0,7.4].

A good selection is r = 6.9 with a variance inflation factor of 6.17, well within the interval [1,10]. Thus, just under seven dimensions are

TABLE II

r	$\|\|\hat{\beta}^{+}\|\|^2$	$\Phi(\hat{\beta}^{+})$	Variance Inflation Factor
9	159.444	20.5821	125.67
8.5	47.449	23.6670	63.55
8	9.518	26.7689	21.28
7.5	9.557	26.7718	11.26
7.4	9.569	26.7724	9.51
7.3	9.582	26.7730	8.78
7.2	9.596	26.7736	8.05
7.1	9.612	26.7742	7.32
7	9.629	26.7747	6.59
6.9	9.489	27.2083	6.17
6.8	9.698	27.6418	5.84
6.75	9.934	27.8586	5.68
6.7	10.256	28.0753	5.52
6.6	11.163	28.5089	5.20
6.5	12.419	28.9424	4.87
6	23.930	31.1100	3.95
5.5	41.349	31.3974	3.26
5	64.436	31.6848	2.69
4.5	77.377	32.3810	1.68
4	91.806	33.0773	1.06

adequate to describe the model. For this effective rank,

$$\hat{\beta}^{+} = \begin{bmatrix} -0.2801 \\ -0.0221 \\ 0.1831 \\ 0.0561 \\ -0.3411 \\ 2.8565 \\ 1.0471 \\ 0.0280 \\ -0.0120 \end{bmatrix}$$

This estimate achieves a 94.0% decrease in the squared length of the coefficient vector relative to OLS with a 32.2% increase in Φ over OLS.

7.4 - Shrunken Estimator Solutions

A. The program ∇SHRINK C in Appendix IV was used to obtain the results in Table III.

TABLE III

α	$\Phi(\underline{c}_\alpha)$	$\|\|\underline{c}_\alpha\|\|^2$
.80	40.3457	102.044
.81	39.3575	104.611
.82	38.3693	107.210
.83	37.3811	109.841
.84	36.3930	112.504
.85	35.4048	115.198
.86	34.4166	117.925
.87	33.4284	120.683
.88	32.4403	123.473
.89	31.4521	126.296
.90	30.4639	129.150

The proper value for α and hence for the coefficient vector $\underline{c}_\alpha$ is chosen by selecting the value of $\Phi(\underline{c}_\alpha)$ the analyst deems reasonable. A possible selection here is $\alpha = .85$,

$$\underline{c}_{.85} = \begin{bmatrix} -2.8169 \\ -0.1659 \\ -0.1216 \\ 3.3422 \\ 0.0804 \\ 1.3293 \\ 9.7091 \\ 0.1081 \\ -0.0034 \end{bmatrix}$$

This solution allows a 72.0% increase in the residual sum of squares for a 27.7% decrease in the squared length of the coefficient vector relative to OLS.

B. The program ∇SHRINK DELTA in Appendix IV yielded the results in Table IV and Figures V and VI.

TABLE IV

δ	$\Phi(\underline{d}_\delta)$	$\|\|\underline{d}_\delta\|\|^2$
.001	62.4825	52.897
.002	47.1690	85.189
.003	40.0529	102.802
.004	35.9418	113.730
.0045	34.4751	117.763
.005	33.2641	121.145
.006	31.3814	126.499
.007	29.9855	130.543
.008	28.9091	133.705
.009	28.0538	136.244
.01	27.3579	138.328
.05	22.0159	154.851
.1	21.3041	157.121
.5	20.7272	158.977
1.0	20.6547	159.209

Clearly, most of the stabilization takes place in the interval (0,.01). A possible selection for δ is .0045.

$$\underline{d}_\delta = \begin{bmatrix} -2.8481 \\ -0.1678 \\ -0.1230 \\ 3.3792 \\ 0.0813 \\ 1.3441 \\ 9.8165 \\ 0.1093 \\ -0.0034 \end{bmatrix}$$

This solution yields a 67.5% increase in Φ and a 26.1% decrease in $||\underline{d}_\delta||^2$ relative to the OLS solution.

FIGURE V

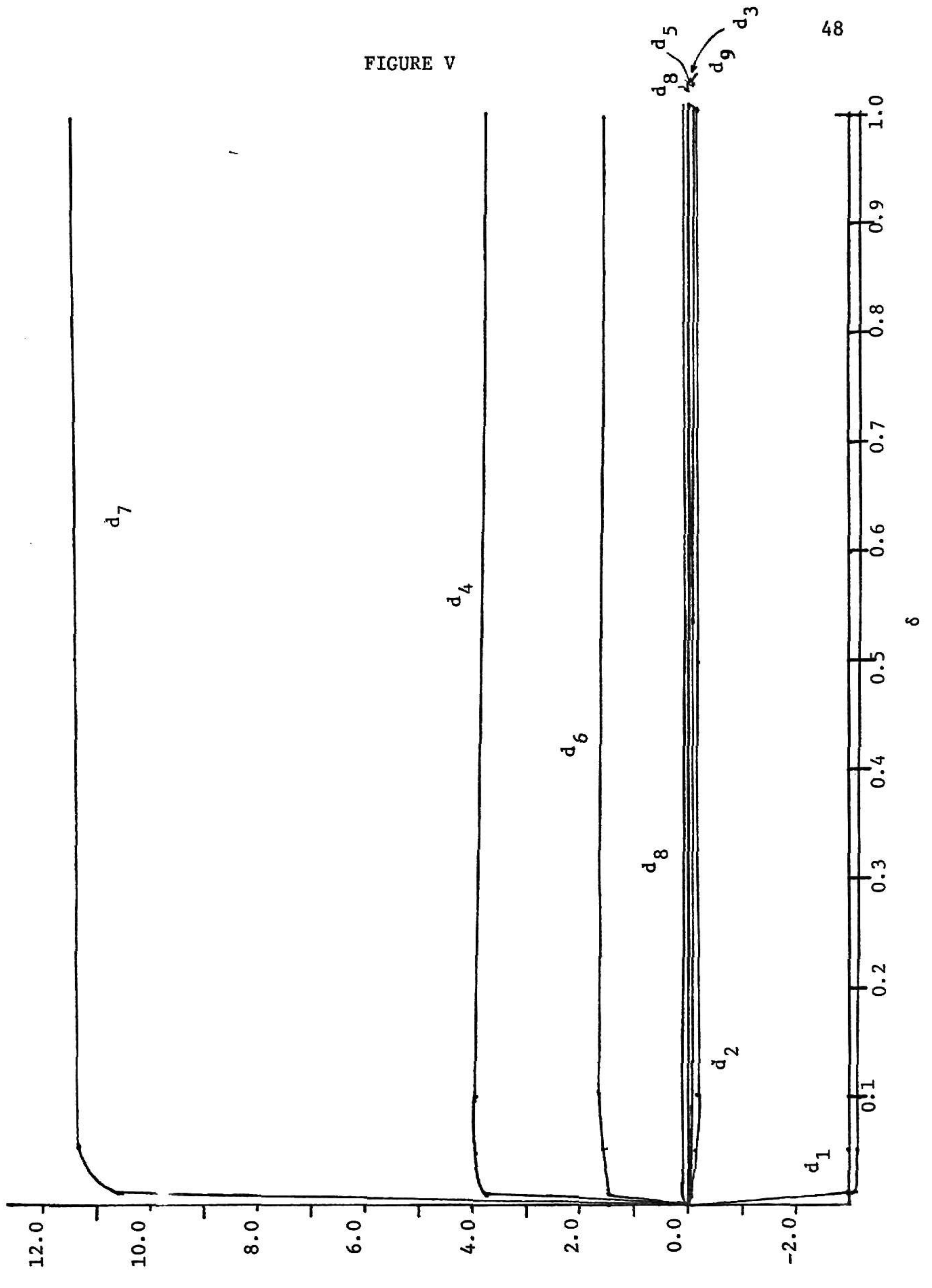
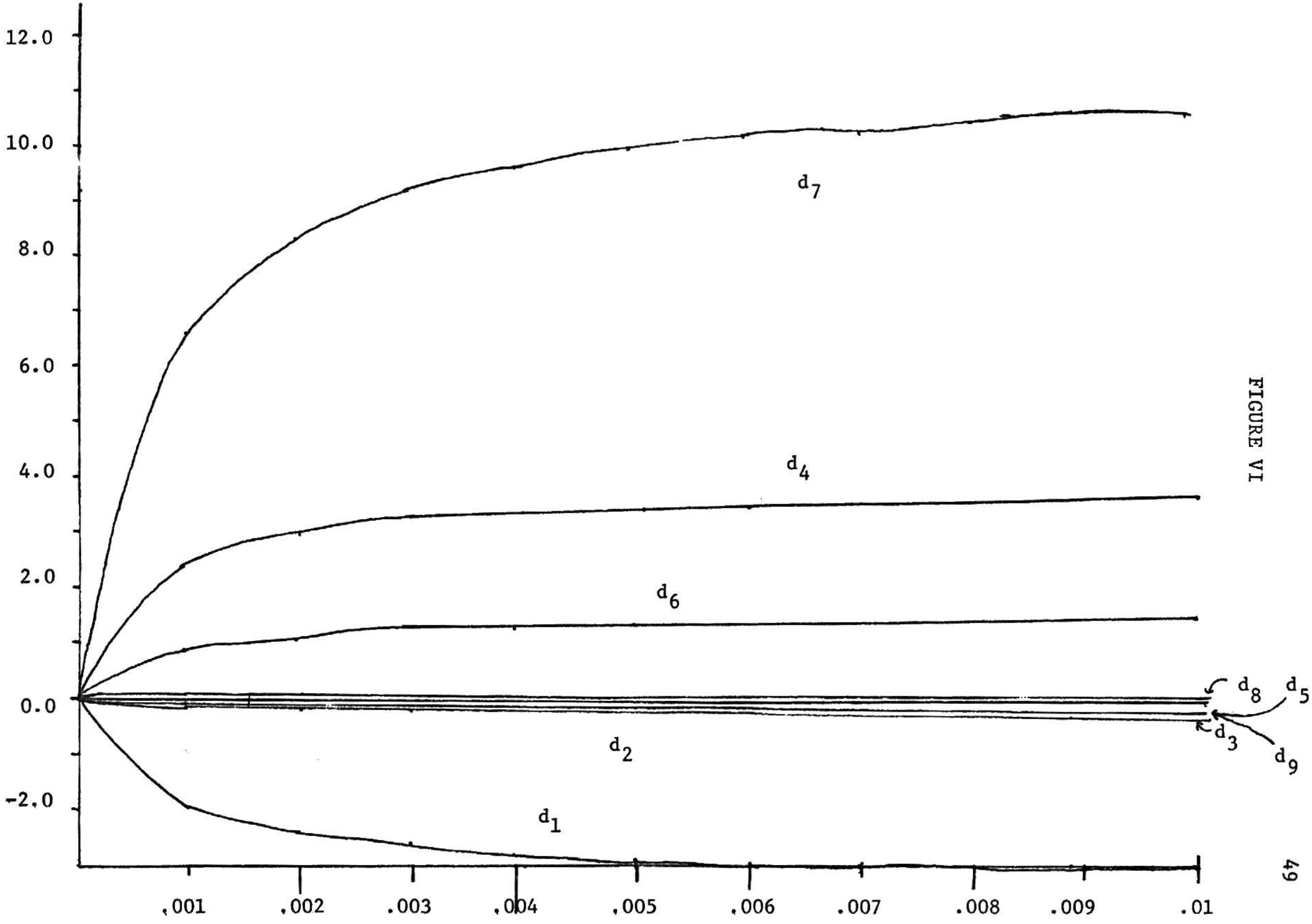

12.0
10.0
8.0
6.0
4.0
2.0
0.0
-2.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
δ
d_7
d_4
d_6
d_8
d_2
d_1
d_8
d_5
d_9
d_3

FIGURE VI


12.0
10.0
8.0
6.0
4.0
2.0
0.0
-2.0
.001
.002
.003
.004
.005
.006
.007
.008
.009
.01
d_7
d_4
d_6
d_8
d_5
d_3
d_9
d_2
d_1

C. $\xi_o = 7/8 = .875$ $s^2 = 20.5821$

$$\hat{\beta}'\hat{\beta} = 159.444$$

Hence $[1 - \xi_o s^2(\hat{\beta}'\hat{\beta})^{-1}] = .887049$

This amounts to an "optimum" value for α in $\underline{c}_\alpha$, the deterministically shrunken estimator. The result is

$$[1 ,- \xi_o s^2(\hat{\beta}'\hat{\beta})^{-1}] = \underline{e}_{\xi_o} = \begin{bmatrix} -2.9397 \\ -0.1732 \\ -0.1269 \\ 3.4878 \\ 0.0839 \\ 1.3873 \\ 10.1322 \\ 0.1128 \\ -0.0035 \end{bmatrix}$$

$(\underline{e}_{\xi_o}) = 31.7437$

(54.2% increase relative to OLS)

$||\underline{e}_{\xi_o}||^2 = 125.459$

(21.3% decrease relative to OLS)

7.5 - Latent Root Solution

The λ_j and γ_{oj} for A'A are in Table V. The $\gamma_{oj}\lambda_j^{-1}$, the coefficients each latent vector $\underline{\gamma}_j^o$ is multiplied by to obtain $\hat{\beta}$ are also shown. It was determined to eliminate $\underline{\gamma}_2^o$ and $\underline{\gamma}_4^o$ from the linear combinations used to calculate b^*,i.e.,since λ_2 and λ_4 are<.05; and both$|\gamma_{o2}|$ and$|\gamma_{o4}|$ are < .10. Observe how the extremely small magnitude of λ_4 causes $\underline{\gamma}_4^o$ to be grossly overweighted in the computation of $\hat{\beta}$.

Elimimating $\underline{\gamma}_2^o$ and $\underline{\gamma}_4^o$ from the analysis;

$$b^* = \begin{bmatrix} -0.1910 \\ -0.0284 \\ 0.2303 \\ -0.1176 \\ -0.3894 \\ 3.0071 \\ 0.3816 \\ 0.0214 \\ -0.0127 \end{bmatrix}$$

$\Phi(b^*) = 27.1880$

(32.1% increase over OLS)

$||b^*||^2 = 9.444$

(94.1% decrease relative to OLS)

TABLE V

j	λ_j	γ_{oj}	$\gamma_{oj}\lambda_j^{-1}$
0	0.2355	0.7447	3.1620
1	5.4531	-0.2965	-0.0544
2	0.0310	-0.0074	-0.2403
3	0.0588	-0.2741	-4.6659
4	0.0038	-0.0737	-19.6013
5	2.0918	-0.4121	-0.1970
6	0.1241	-0.2333	-1.8801
7	1.2384	-0.0329	-0.0266
8	0.1379	-0.1428	-1.0360
9	0.6257	-0.1765	-0.2821

7.6 Table VI presents all the estimates, their residual sum of squares, and squared coefficient vector lengths together for ease of comparison. The example dat appears in Table VII.

TABLE VI

$\hat{\beta}$	$\hat{\beta}^*$	exp. $\hat{\beta}^*$	$\hat{\beta}^+$	$\underline{c}_\alpha$	$\underline{d}_\delta$	$\underline{e}_{\xi_o}$	b^*
-3.314	-0.147	0.100	-0.280	-2.817	-2.848	-2.940	-0.191
-0.195	-0.048	-0.043	-0.022	-0.166	-0.168	-0.173	-0.028
-0.143	-0.001	-0.071	0.183	-0.122	-0.123	-0.127	0.230
3.932	0.147	0.025	0.056	3.342	3.379	3.488	-0.117
0.095	-0.468	-0.885	-0.341	0.080	0.081	0.084	-0.389
1.564	1.633	0.839	2.857	1.329	1.344	1.387	3.007
11.422	5.023	6.584	1.047	9.709	9.817	10.132	0.382
0.127	0.039	0.035	0.028	0.108	0.109	0.113	0.021
-0.004	-0.009	-0.005	-0.012	-0.003	-0.003	-0.004	-0.013
[a] 20.582	35.491	44.828	27.208	35.405	34.475	31.744	27.188
[b] 159.444	28.165	44.854	9.489	115.198	117.763	125.459	9.444

[a] $\Phi(B)$

[b] $||B||^2$

FIGURE VII

Data for Example

Point	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1	2.0	91.5	85.7	61.5	87.0	8.21	1.23	.25	24.9	80.0
2	4.5	96.0	88.0	65.5	91.0	7.08	1.87	.26	25.9	80.0
3	1.0	95.0	88.5	68.9	90.9	6.38	1.19	.19	22.6	110.0
4	1.0	90.6	82.0	58.0	86.0	7.97	1.75	.16	26.1	140.0
5	5.5	92.8	84.5	60.2	88.0	6.56	2.76	.37	31.6	80.0
6	8.0	92.3	84.2	60.0	88.2	6.65	2.89	.25	35.2	65.0
7	9.0	92.0	84.0	57.5	87.2	5.64	3.33	.25	47.1	80.0
8	9.0	90.5	82.0	55.0	85.0	5.86	3.74	.23	84.9	80.0
9	7.5	92.5	84.5	61.0	87.8	5.78	2.43	.26	44.5	65.0
10	10.0	91.2	81.2	51.8	85.7	4.82	3.49	.25	77.0	80.0
11	4.0	93.3	84.5	59.0	88.1	5.45	2.56	.25	43.6	80.0
12	3.0	93.0	85.2	59.0	88.1	5.96	2.32	.22	42.7	60.0
13	4.0	91.6	83.2	58.6	86.5	6.55	2.74	.26	45.6	60.0
14	5.0	89.0	81.0	55.2	84.0	6.02	2.77	.24	45.6	280.0
15	4.5	91.9	83.5	56.9	86.8	6.18	3.47	.36	36.2	260.0

APPENDIX I

THE CORRELATION MATRIX

Represent the original model by

$$Y = \psi_o + \psi_1 Z_1 + \psi_2 Z_2 + \ldots + \psi_p Z_p + \varepsilon \tag{1}$$

or

$$Y = Z\psi + \varepsilon \qquad \text{where } \varepsilon \sim N(0, \sigma^2 I) \tag{2}$$

Express the dependent observations Y and the independent observations Z_i as deviations from the respective means;

$$(I_n - \tfrac{1}{n} J_n)Y = (I_n - \tfrac{1}{n} J_n)Z\psi + \varepsilon^* \tag{3}$$

where $\varepsilon^* \sim N(0, \sigma^2(I_n - \frac{1}{n} J_n))$

or

$$Y^* = Z^*\psi + \varepsilon^* \tag{4}$$

Apply the transformation $P^{-1}(Z^{*\prime}Z^*)P^{-1}$ where P^2 is the matrix of diagonal elements of $Z^{*\prime}Z^*$. Then $P^{-1}(Z^{*\prime}Z^*)P^{-1}$ is in the form of a correlation matrix;

i.e.

$$P^{-1}(Z^{*\prime}Z^*)P^{-1} = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{12} & 1 & r_{23} & \cdots & r_{2p} \\ r_{13} & r_{23} & 1 & \cdots & r_{3p} \\ & & \vdots & & \\ r_{1p} & r_{2p} & r_{3p} & \cdots & 1 \end{bmatrix}. \tag{5}$$

Inserting $P^{-1}P = I$ in model (4),

$$Y^* = (Z^*P^{-1})(P\psi) + \varepsilon^* = X\beta + \varepsilon^* \tag{6}$$

where $\beta = P\psi$ and $X = Z^*P^{-1}$ which implies $\psi = P^{-1}\beta$ (7)

The least squares solution for β is now

$$\hat{\beta} = (X'X)^{-1} X'Y^{*} = (P^{-1}Z^{*}{}'Z^{*}P^{-1})^{-1} P^{-1}Z^{*}{}'Y^{*} \qquad (8)$$

After the estimate B of β has been obtained by whatever method, the estimates of the original parameters are;

$$\hat{\psi} = P^{-1} B \qquad (9)$$

APPENDIX II

ITERATIVE SOLUTION FOR RIDGE REGRESSION

Reduce X'X to a diagonal matrix by applying the orthogonal transformation S.

$$S(X'X)S' = \Lambda \qquad (10)$$

where S is pxp such that $S'S = I$ and Λ is a diagonal matrix whose iagonal elements $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the characteristic roots of X'X. We write $X^* = XS'$ and $\mu = S\beta$. Then the model (1.1) may be written

$$Y = X^*\mu + \varepsilon \qquad (11)$$

where $(X^*)'(X^*) = \Lambda$ (12)

Therefore $\hat{\mu} = [(X^*)'(X^*)]^{-1} X^{*}{}'Y = \Lambda^{-1}SX'Y$ (13)

The iterative procedure is described by the formula

$$k_{i(j)} = \hat{\sigma}^2 \quad (\hat{\mu}^*_{i(j)})^2 \qquad (14)$$

where the j subscript denotes the jth iterate and $\hat{\sigma}^2$ is the residual sum of squares for the model (1.1) divided by (n-p-1), the unbiased estimator for σ^2.

As initial values, use

$$\hat{\mu}^*_{i(o)} = \hat{\mu}_i \qquad ; \quad i = 1, 2, \ldots, p \qquad (15)$$

where $\hat{\mu}_i$ is the OLS estimate of μ_i. The $k_{i(j)}$ values are used in equation (2.1) to obtain the next $\hat{\mu}^*_{i(j+1)}$ values for use in (14).

Although the authors ([5] and [6]) do not recommend a criterion for terminating the iteration, it seems reasonable to terminate when $k_{i(j+1)} - k_{i(j)} < \Delta$ where Δ is some predetermined small value.

To obtain the estimates of the original β_i's,

$\mu = S\beta$ implies that $S'\mu = S'S\beta = \beta$ (16)

Hence

$$S' \hat{\mu}^* = \hat{\beta} \qquad (17)$$

APPENDIX III

EXPLICIT SOLUTION FOR RIDGE REGRESSION

Begin by performing the orthogonal transformation on X'X described in Appendix II. The explicit solution depends on certain convergence/ divergence conditions related to the iterative solution.

(i) Compute $q_{i(o)} = \dfrac{\hat{\sigma}^2}{\lambda_i \hat{\mu}_i^2}$; $i = 1, 2, . . ., p$ (18)

where $\hat{\sigma}^2$ and $\hat{\mu}_i$ are the OLS estimates of σ^2 and μ_i, repectively and λ_i is the ith characteristic root of X'X.

(ii) Hemmerle [5] shows that the iterative process defined by (14) converges whenever $0 < q_o \leq 1/4$ and diverges for $q_o > 1/4$. Thus, we let

$$\hat{\mu}_i^* = 0 \text{ for } q_{i(o)} > 1/4 \qquad (19)$$

$$\text{and } \hat{\mu}_i^* = \frac{\hat{\mu}_i}{(1 + q_i^*)} \quad \text{for } 0 < q_{i(o)} \leq 1/4 \qquad (20)$$

$$\text{where } q_i^* = \frac{(1 - 2q_{i(o)}) - \sqrt{(1 - 4q_{i(o)})}}{2q_{i(o)}} \qquad (21)$$

(iii) This procedure may produce an undesired large increase in the residual sum of squares. One may desire to follow some additional steps to prevent this. These may be found in [5]. Compute the estimates of the original β_i as in equation (17).

APPENDIX IV

APL ROUTINES

The following programs written in the APL programming language were used for the computations of Chapter 7. XPX, XPY, I9, DI, LS and T are user-supplied global variables.

```
I.      ∇RIDGE K

  [1] BH←(⌹(XPX + K x I9)) +.x XPY
  [2] SSE←YPY - ((⍉BH) +.x XPY)
  [3] 'SSE= ';SSE
  [4] BH←DI +.x BH
  [5] 'BH= ';BH
  [6] 'LENGTH= ':(⍉BH) +.x BH
  [7] ∇
```

K is the user-input value for k in equation (2.1). Output is BH, the vector of estimates of the original parameter vector β; SSE, the residual sum of squares; and the squared length of BH.

```
II.      ∇GINV

  [1] BH←T +.x XPY
  [2] SSE←YPY - ((⍉BH) +.x XPY)
  [3] 'SSE= ';SSE
  [4] BH←DI +.x BH
  [5] 'BH= ';BH
  [6] 'LENGTH= ';(⍉BH) +.x BH
  [7] 'T= ';+/TxI9
  [8] ∇
```

T is computed from (3.2) prior to program use. Ouput is the same as ∇RIDGE. The diagonal elements of T are also output for selection of the maximum variance inflation factor.

III. ∇SHRINK C

```
[1] BH←C x LS
[2] SSE←YPY - ((⍉BH) +.x XPY)
[3] 'SSE= ';SSE
[4] BH←DI+.x BH
[5] 'BH= ';BH
[6] 'LENGTH= ';(⍉BH) +.x BH
[7] ∇
```

Input is a value of C in the interval [0,1]. Output is the same as for ∇RIDGE.

IV. ∇SHRINK DELTA

```
[1] BH←DELTA x (I9 + (DELTA x (⌹(LS +.x(⍉LS))))) +.x LS
[2] SSE←YPY - ((⍉BH) +.x XPY)
[3] 'SSE= ';SSE
[4] BH←DI +.x BH
[5] 'BH= ';BH
[6] 'LENGTH= ';(⍉BH) +.x BH
[7] ∇
```

Input is a value for DELTA in the interval [0,1]. Output is the same as for ∇RIDGE.

V. Characteristic roots and characteristic vectors were computed using EIGEN, which can be copied from Library 5.

REFERENCES

[1] Allen,David M. (1974).
"The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction." Technometrics,16,125-128.

[2] Draper,N.R. and Smith,H. (1966).
Applied Regression Analysis. John Wiley and Sons,Inc., New York, Chapter 5.

[3] Gunst,R.F., Webster,J.T. and Mason,R.L. (1976).
"A Comparison of Least Squares and Latent Root Regression Analysis". Technometrics, 18, 75-84.

[4] Hawkins,Douglas M. (1975).
"Relations Between Rikge Regression and Eigenanalysis of the Augmentation Correlation Matrix". Technometrics, 17, 477-480.

[5] Hemmerle,William J. (1975).
"An Explicit Solution for Generalized Ridge Regression". Technometrics, 17, 309-314.

[6] Hoerl,Arthur E. and Kennard,Robert W. (1970).
"Ridge Regression: Biased Estimation for Nonorthogonal Problems". Technometrics, 12, 69-82.

[7] Hoerl,Arthur E. and Kennard,Robert W. (1970).
"Ridge Regression:Applications to Nonorthogonal Problems".Technometrics, 12, 69-82.

[8] Marquardt,D.W. (1970).
"Generalized Inverses, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation". Technometrics, 12, 591-612.

[9] Mayer,Lawrence S. and Willke,Thomas A. (1973).
"On Biased Estimation in Linear Models". Technometrics, 15, 497-508.

[10] Sclove,Stanley L. (1968).
"Improved Estimators for Coefficients in Linear Regression". Journal of the American Statistical Association, 63, 597-606.

[11] Swindel,Benee F. (1974).
"Instability of Regression Coefficients Illustrated". The American Statistician, 28, 63-65.

[12] Webster,Jack T., Gunst,R.F. and Mason,R.L. (1974).
"Latent Root Regression Analysis". Technometrics, 16, 513-522.

BIASED ESTIMATION TECHNIQUES FOR MULTIPLE LINEAR REGRESSION

by

PHILLIP DEAN WITTMER

B. S., Kansas State University, 1975

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

Kansas State University
Manhattan, Kansas

1976

Multiple linear regression analysis involves the attempt to explain the variability in a dependent variable by a linear combination of certain independent variables. The serious problems resulting from estimating the parameters of the model by least squares when a great deal of multicollinearity exists among the independent variables are demonstrated.

Some alternative techniques for estimating the parameters under these circumstances are then presented. The techniques discussed are Ridge Regression, Generalized Inverses, Shrunken Estimators and Latent Root Regression Analysis. All four techniques provide biased estimators of the model parameters.

The methods for computing parameter estimates are given in step-by-step fashion for each technique as well as a summary of the justification and reasoning behind the technique. Basically, the theory supporting any biased technique is to "break up" some of the interrelationships among the independent variables in order to arrive at estimators with smaller mean square error than the least squares estimator.

The four techniques are compared from a theoretical standpoint, and it is seen that many similarities exist. A data set contining high correlations among the explanatory variables is analyzed utilizing the four techniques to illustrate the computations involved.