Ridge Regression: Biased Estimation for Nonorthogonal Problems

ARTHUR E. HOERL AND ROBERT W. KENNARD

University of Delaware and E. I. du Pont de Nemours & Co.

In multiple regression it is shown that parameter estimates based on minimum residual sum of squares have a high probability of being unsatisfactory, if not incorrect, if the prediction vectors are not orthogonal. Proposed is an estimation procedure based on adding small positive quantities to the diagonal of $\mathbf{X}'\mathbf{X}$. Introduced is the ridge trace, a method for showing in two dimensions the effects of nonorthogonality. It is then shown how to augment $\mathbf{X}'\mathbf{X}$ to obtain biased estimates with smaller mean square error.

0. INTRODUCTION

Consider the standard model for multiple linear regression, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where it is assumed that $\mathbf{X}$ is $(n \times p)$ and of rank p, β is $(p \times 1)$ and unknown, $E[\varepsilon] = \mathbf{0}$, and $E[\varepsilon\varepsilon'] = \sigma^2\mathbf{I}_n$. If an observation on the factors is denoted by $\mathbf{x}_v = \{x_{1v}, x_{2v}, \ldots, x_{pv}\}$, the general form $\mathbf{X}\beta$ is $\{\sum_{i=1}^{p} \beta_i\theta_i(\mathbf{x}_v)\}$ where the θ_i are functions free of unknown parameters.

The usual estimation procedure for the unknown β is Gauss-Markov—linear functions of $\mathbf{Y} = \{y_v\}$ that are unbiased and have minimum variance. This estimation procedure is a good one if $\mathbf{X}'\mathbf{X}$, when in the form of a correlation matrix, is nearly a unit matrix. However, if $\mathbf{X}'\mathbf{X}$ is not nearly a unit matrix, the least squares estimates are sensitive to a number of "errors." The results of these errors are critical when the specification is that $\mathbf{X}\beta$ is a true model. Then the least squares estimates often do not make sense when put into the context of the physics, chemistry, and engineering of the process which is generating the data. In such cases, one is forced to treat the estimated predicting function as a black box or to drop factors to destroy the correlation bonds among the $\mathbf{X}_i$ used to form $\mathbf{X}'\mathbf{X}$. Both these alternatives are unsatisfactory if the original intent was to use the estimated predictor for control and optimization. If one treats the result as a black box, he must caution the user of the model not to take partial derivatives (a useless caution in practice), and in the other case, he is left with a set of dangling controllables or observables.

Estimation based on the matrix $[\mathbf{X}'\mathbf{X} + k\mathbf{I}_p]$, $k \geq 0$ rather than on $\mathbf{X}'\mathbf{X}$ has been found to be a procedure that can be used to help circumvent many of the difficulties associated with the usual least squares estimates. In particular, the procedure can be used to portray the sensitivity of the estimates to the particular set of data being used, and it can be used to obtain a point estimate with a smaller mean square error.

Received Aug. 1968; revised June 1969.

1. Properties of Best Linear Unbiased Estimation

Using unbiased linear estimation with minimum variance or maximum likelihood estimation when the random vector, ε, is normal gives

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{1.1}$$

as an estimate of β and this gives the minimum sum of squares of the residuals:

$$\phi(\hat{\beta}) = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}). \tag{1.2}$$

The properties of $\hat{\beta}$ are well known [18]. Here the concern is primarily with cases for which $\mathbf{X}'\mathbf{X}$ is not nearly a unit matrix (unless specified otherwise, the model is formulated to give an $\mathbf{X}'\mathbf{X}$ in correlation form). To demonstrate the effects of this condition on the estimation of β, consider two properties of $\hat{\beta}$ — its variance-covariance matrix and its distance from its expected value.

(*i*) $\mathbf{VAR}\,(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ (1.3)

(*ii*) $L_1 \equiv$ Distance from $\hat{\beta}$ to β.

$$L_1^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta) \tag{1.4}$$

$$E[L_1^2] = \sigma^2 \text{ Trace } (\mathbf{X}'\mathbf{X})^{-1} \tag{1.5}$$

or equivalently

$$E[\hat{\beta}'\hat{\beta}] = \beta'\beta + \sigma^2 \text{ Trace } (\mathbf{X}'\mathbf{X})^{-1} \tag{1.5a}$$

When the error ε is normally distributed, then

$$\text{VAR } [L_1^2] = 2\sigma^4 \text{ Trace } (\mathbf{X}'\mathbf{X})^{-2}\,. \tag{1.6}$$

These related properties show the uncertainty in $\hat{\beta}$ when $\mathbf{X}'\mathbf{X}$ moves from a unit matrix to an ill-conditioned one. If the eigenvalues of $\mathbf{X}'\mathbf{X}$ are denoted by

$$\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p = \lambda_{\min} > 0, \tag{1.7}$$

then the average value of the squared distance from $\hat{\beta}$ to β is given by

$$E[L_1^2] = \sigma^2 \sum_{i=1}^{p} (1/\lambda_i) \tag{1.8}$$

and the variance when the error is normal is given by

$$\text{VAR } [L_1^2] = 2\sigma^4 \sum (1/\lambda_i)^2\,. \tag{1.9}$$

Lower bounds for the average and variance are $\sigma^2/\lambda_{\min}$ and $2\sigma^4/\lambda_{\min}^2$, respectively. Hence, if the shape of the factor space is such that reasonable data collection results in an $\mathbf{X}'\mathbf{X}$ with one or more small eigenvalues, the distance from $\hat{\beta}$ to β will tend to be large. Estimated coefficients, $\hat{\beta}_i$, that are large in absolute value have been observed by all who have tackled live nonorthogonal data problems.

The least squares estimate (1.1) suffers from the deficiency of mathematical optimization techniques that give point estimates; the estimation procedure does not have built into it a method for portraying the sensitivity of the solution (1.1) to the optimization criterion (1.2). The procedures to be discussed in the sections to follow portray the sensitivity of the solutions and utilize nonsensitivity as an aid to analysis.

2. Ridge Regression

A. E. Hoerl first suggested in 1962 [9] [11] that to control the inflation and general instability associated with the least squares estimates, one can use

$$\hat{\beta}^* = [\mathbf{X}'\mathbf{X} + k\mathbf{I}]^{-1}\mathbf{X}'\mathbf{Y}; k \geq 0 \tag{2.1}$$

$$= \mathbf{W}\mathbf{X}'\mathbf{Y} . \tag{2.2}$$

The family of estimates given by $k \geq 0$ has many mathematical similarities with the portrayal of quadratic response functions [10]. For this reason, estimation and analysis built around (2.1) has been labeled "ridge regression." The relationship of a ridge estimate to an ordinary estimate is given by the alternative form

$$\hat{\beta}^* = [\mathbf{I}_p + k(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\hat{\beta} \tag{2.3}$$

$$= \mathbf{Z}\hat{\beta} . \tag{2.4}$$

This relationship will be explored further in subsequent sections. Some properties of $\hat{\beta}^*$, $\mathbf{W}$, and $\mathbf{Z}$ that will be used are:

(*i*) Let $\xi_i(\mathbf{W})$ and $\xi_i(\mathbf{Z})$ be the eigenvalues of $\mathbf{W}$ and $\mathbf{Z}$, respectively. Then

$$\xi_i(\mathbf{W}) = 1/(\lambda_i + k) \tag{2.5}$$

$$\xi_i(\mathbf{Z}) = \lambda_i/(\lambda_i + k) \tag{2.6}$$

where λ_i are the eigenvalues of $\mathbf{X}'\mathbf{X}$. These results follow directly from the definitions of $\mathbf{W}$ and $\mathbf{Z}$ in (2.2) and (2.4) and the solution of the characteristic equations $|\mathbf{W} - \xi\mathbf{I}| = 0$ and $|\mathbf{Z} - \xi\mathbf{I}| = 0$.

(*ii*) $$\mathbf{Z} = \mathbf{I} - k(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} = \mathbf{I} - k\mathbf{W} \tag{2.7}$$

The relationship is readily verified by writing $\mathbf{Z}$ in the alternative form $\mathbf{Z} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{W}\mathbf{X}'\mathbf{X}$ and multiplying both sides of (2.7) on the left by $\mathbf{W}^{-1}$.

(*iii*) $\hat{\beta}^*$ for $k \neq 0$ is shorter than $\hat{\beta}$, i.e.

$$(\hat{\beta}^*)'(\hat{\beta}^*) < \hat{\beta}'\hat{\beta}. \tag{2.8}$$

By definition $\hat{\beta}^* = \mathbf{Z}\hat{\beta}$. From its definition and the assumptions on $\mathbf{X}'\mathbf{X}$, $\mathbf{Z}$ is clearly symmetric positive definite. Then the following relation holds [17]:

$$(\hat{\beta}^*)'\,(\hat{\beta}^*) \leq \xi^2_{\max}(\mathbf{Z})\;\hat{\beta}'\hat{\beta}. \tag{2.9}$$

But $\xi_{\max}(\mathbf{Z}) = \lambda_1/(\lambda_1 + k)$ where λ_1 is the largest eigenvalue of $\mathbf{X}'\mathbf{X}$ and (2.8) is established. From (2.6) and (2.7) it is seen that $\mathbf{Z}(0) = \mathbf{I}$ and that $\mathbf{Z}$ approaches $\mathbf{0}$ as $k \to \infty$.

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

$$\phi^*(k) = (\mathbf{Y} - \mathbf{X}\hat{\beta}^*)'\,(\mathbf{Y} - \mathbf{X}\hat{\beta}^*) \tag{2.10}$$

which can be written in the form

$$\phi^*(k) = \mathbf{Y}'\mathbf{Y} - (\hat{\beta}^*)'\,\mathbf{X}'\mathbf{Y} - k(\hat{\beta}^*)'\,(\hat{\beta}^*). \tag{2.11}$$

The expression shows that $\phi^*(k)$ is the total sum of squares less the "regression" sum of squares for $\hat{\beta}^*$ with a modification depending upon the squared length of $\hat{\beta}^*$.

3. The Ridge Trace

a. Definition of the Ridge Trace

When $\mathbf{X'X}$ deviates considerably from a unit matrix, that is, when it has small eigenvalues, (1.5) and (1.6) show that the probability can be small that $\hat{\beta}$ will be close to β. In any except the smallest problems, it is difficult to untangle the relationships among the factors if one is confined to an inspection of the simple correlations that are the elements of $\mathbf{X'X}$. That such untangling is a problem is reflected in the "automatic" procedures that have been put forward to reduce the dimensionality of the factor space or to select some "best" subset of the predictors. These automatic procedures include regression using the factors obtained from a coordinate transformation using the principal components of $\mathbf{X'X}$, stepwise regression, computation of all 2^p regressions, and some subset of all regressions using fractional factorials or a branch and bound technique [3][5][6][7][8][14][19]. However, with the occasional exception of principal components, these methods don't really give an insight into the structure of the factor space and the sensitivity of the results to the particular set of data at hand. But by computing $\hat{\beta}^*(k)$ and $\phi^*(k)$ for a set of values of k, such insight can be obtained. A detailed study of two nonorthogonal problems and the conclusions that can be drawn from their ridge traces is given in [12].

b. Characterization of the Ridge Trace

Let $\mathbf{B}$ be any estimate of the vector β. Then the residual sums of squares can be written as

$$\begin{aligned}\phi &= (\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB}) \\ &= (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) + (\mathbf{B} - \hat{\beta})'\mathbf{X'X}(\mathbf{B} - \hat{\beta}) \\ &= \phi_{\min} + \phi(\mathbf{B})\end{aligned} \tag{3.1}$$

Contours of constant ϕ are the surfaces of hyperellipsoids centered at $\hat{\beta}$, the ordinary least squares estimate of β. The value of ϕ is the minimum value, $\phi_{\min}$, plus the value of the quadratic form in $(\mathbf{B} - \hat{\beta})$. There is a continuum of values of $\mathbf{B}_0$ that will satisfy the relationship $\phi = \phi_{\min} + \phi_0$ where $\phi_0 > 0$ is a fixed increment. However, the relationships in Section 2 show that on the average the distance from $\hat{\beta}$ to β will tend to be large if there is a small eigenvalue of $\mathbf{X'X}$. In particular, the worse the conditioning of $\mathbf{X'X}$, the more $\hat{\beta}$ can be expected to be too long. On the other hand, the worse the conditioning, the further one can move from $\hat{\beta}$ without an appreciable increase in the residual sums of squares. In view of (1.5a) it seems reasonable that if one moves away from the minimum sum of squares point, the movement should be in a direction which will shorten the length of the regression vector.

The ridge trace can be shown to be following a path through the sums of

squares surface so that for a fixed ϕ a single value of $\mathbf{B}$ is chosen and that is the one with minimum length. This can be stated precisely as follows:

Minimize $\mathbf{B}'\mathbf{B}$

$$\text{subject to } (\mathbf{B} - \hat{\beta})'\mathbf{X}'\mathbf{X}(\mathbf{B} - \hat{\beta}) = \phi_0 . \tag{3.2}$$

As a Lagrangian problem this is

$$\text{Minimize } F = \mathbf{B}'\mathbf{B} + (1/k)[(\mathbf{B} - \hat{\beta})'\mathbf{X}'\mathbf{X}(\mathbf{B} - \hat{\beta})' - \phi_0] \tag{3.3}$$

where $(1/k)$ is the multiplier. Then

$$\frac{\partial F}{\partial \mathbf{B}} = 2\mathbf{B} + (1/k)[2(\mathbf{X}'\mathbf{X})\mathbf{B} - 2(\mathbf{X}'\mathbf{X})\hat{\beta}] = 0 \tag{3.4}$$

This reduces to

$$\mathbf{B} = \hat{\beta}^* = [\mathbf{X}'\mathbf{X} + k\mathbf{I}]^{-1}\,\mathbf{X}'\mathbf{Y} \tag{3.5}$$

where k is chosen to satisfy the restraint (3.2). This is the ridge estimator. Of course, in practice it is easier to choose a $k \geq 0$ and then compute ϕ_0 . In terms of $\hat{\beta}^*$ the residual sum of squares becomes

$$\phi^*(k) = (\mathbf{Y} - \mathbf{X}\hat{\beta}^*)'(\mathbf{Y} - \mathbf{X}\hat{\beta}^*) = \phi_{\min} + k^2\hat{\beta}^{*\prime}(\mathbf{X}'\mathbf{X})^{-1}\,\hat{\beta}^*. \tag{3.6}$$

A completely equivalent statement of the path is this: If the squared length of the regression vector $\mathbf{B}$ is fixed at R^2, then $\hat{\beta}^*$ is the value of $\mathbf{B}$ that gives a minimum sum of squares. That is, $\hat{\beta}^*$ is the value of $\mathbf{B}$ that minimizes the function

$$F_1 = (\mathbf{Y} - \mathbf{XB})'\,(\mathbf{Y} - \mathbf{XB}) + (1/\mathrm{k})\,(\mathbf{B}'\mathbf{B} - R^2). \tag{3.7}$$

c. Likelihood Characterization of the Ridge Trace.

Using the assumption that the error vector is Normal $(\mathbf{0}, \sigma^2\mathbf{I}_n)$ the likelihood function is

$$(2\pi\sigma^2)^{-n/2} \exp\{ - (1/2\sigma^2)(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)\}. \tag{3.8}$$

The kernel of this function is the quadratic form in the exponential which can be written in the form

$$(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) + (\beta - \hat{\beta})'\mathbf{X}'\mathbf{X}(\beta - \hat{\beta}). \tag{3.9}$$

With (3.1) in 3*b*, this shows that an increase in the residual sum of squares is equivalent to a decrease in the value of the likelihood function. So the contours of equal likelihood also lie on the surface of hyperellipsoids centered at $\hat{\beta}$.

The ridge trace can thereby be interpreted as a path through the likelihood space, and the question arises as why this particular path can be of special interest. The reasoning is the same as for the sum of squares. Although long vectors give the same likelihood values as shorter vectors, they will not always have equal physical meaning. Implied is a restraint on the possible values of $\hat{\beta}$ that is not made explicit in the formulation of the general linear model given in the Introduction. This implication is discussed further in the sections that follow.

4. Mean Square Error Properties of Ridge Regression

a. Variance and Bias of a Ridge Estimator

To look at $\hat{\beta}^*$ from the point of view of mean square error it is necessary to obtain an expression for $E[L_1^2(k)]$. Straightforward application of the expectation operator and (2.3) gives the following:

$$
\begin{aligned}
E[L_1^2(k)] &= E[(\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta)] \\
&= E[(\hat{\beta} - \beta)'\mathbf{Z}'\mathbf{Z}(\hat{\beta} - \beta)] + (\mathbf{Z}\beta - \beta)'(\mathbf{Z}\beta - \beta) && (4.2) \\
&= \sigma^2 \text{ Trace } (\mathbf{X}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Z} + \beta'(\mathbf{Z} - \mathbf{I})'(\mathbf{Z} - \mathbf{I})\beta && (4.3) \\
&= \sigma^2[\text{Trace } (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} - k \text{ Trace } (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-2}] \\
&\qquad + k^2\beta'(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-2}\beta && (4.4) \\
&= \sigma^2 \sum_1^p \lambda_i/(\lambda_i + k)^2 + k^2\beta'(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-2}\beta && (4.5) \\
&= \gamma_1(k) + \gamma_2(k) && (4.6)
\end{aligned}
$$

The meanings of the two elements of the decomposition, $\gamma_1(k)$ and $\gamma_2(k)$, are readily established. The second element, $\gamma_2(k)$, is the squared distance from $\mathbf{Z}\beta$ to β. It will be zero when $k = 0$, since $\mathbf{Z}$ is then equal to $\mathbf{I}$. Thus, $\gamma_2(k)$ can be considered the square of a bias introduced when $\hat{\beta}^*$ is used rather than $\hat{\beta}$. The first term, $\gamma_1(k)$, can be shown to be the sum of the variances (total variance) of the parameter estimates. In terms of the random variable $\mathbf{Y}$,

$$\hat{\beta}^* = \mathbf{Z}\hat{\beta} = \mathbf{Z}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{4.7}$$

Then

$$
\begin{aligned}
\mathbf{VAR}\ [\hat{\beta}^*] &= \mathbf{Z}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{VAR}\ [\mathbf{Y}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{Z}' \\
&= \sigma^2\mathbf{Z}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{Z}' . && (4.8)
\end{aligned}
$$

The sum of the variances of all the $\hat{\beta}_i^*$ is the sum of the diagonal elements of (4.8).

Figure 1 shows in qualitative form the relationship between the variances, the squared bias, and the parameter k. The total variance decreases as k increases, while the squared bias increases with k. As is indicated by the dotted line, which is the sum of $\gamma_1(k)$ and $\gamma_2(k)$ and thus is $E[L_1^2(k)]$, the possibility exists that there are values of k (admissible values) for which the mean square error is less for $\hat{\beta}^*$ than it is for the usual solution $\hat{\beta}$. This possibility is supported by the mathematical properties of $\gamma_1(k)$ and $\gamma_2(k)$. [See Section 4*b*.] The function $\gamma_1(k)$ is a monotonic decreasing function of k, while $\gamma_2(k)$ is monotonic increasing. However, the most significant feature is the value of the derivative of each function in the neighborhood of the origin. These derivatives are:

$$\lim_{k \to 0^+} (d\gamma_1/dk) = -2\sigma^2\Sigma(1/\lambda_i^2) \tag{4.9}$$

$$\lim_{k \to 0^+} (d\gamma_2/dk) = 0 . \tag{4.10}$$

Thus, $\gamma_1(k)$ has a negative derivative which approaches $-2p\sigma^2$ as $k \to 0^+$ for an

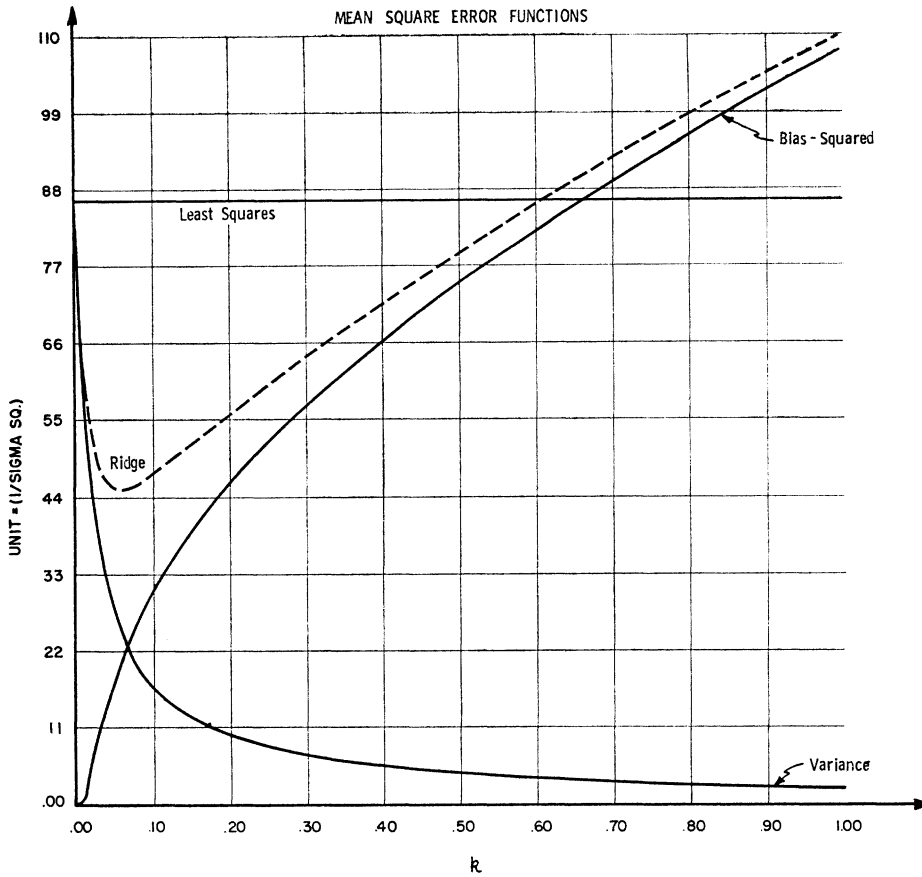

FIGURE 1

orthogonal $\mathbf{X}'\mathbf{X}$ and approaches $-\infty$ as $\mathbf{X}'\mathbf{X}$ becomes ill-conditioned and $\lambda_p \to 0$. On the other hand, as $k \to 0^+$, (4.10) shows that $\gamma_2(k)$ is flat and zero at the origin. These properties lead to the conclusion that it is possible to move to $k > 0$, take a little bias, and substantially reduce the variance, thereby improving the mean square error of estimation and prediction. An existence theorem to validate this conclusion is given in Section 4*b*.

b. Theorems on the Mean Square Function

Theorem 4.1. The total variance $\gamma_1(k)$ is a continuous, monotonically decreasing function of k.

Corollary 4.1.1. The first derivative with respect to k of the total variance $\gamma_1'(k)$, approaches $-\infty$ as $k \to 0^+$ and $\lambda_p \to 0$.

Both the theorem and the corollary are readily proved by use of $\gamma_1(k)$ and its derivative expressed in terms of λ_i.

Theorem 4.2. The squared bias $\gamma_2(k)$ is a continuous, monotonically increasing function of k.

Proof: From (4.5) $\gamma_2(k) = k^2\beta'(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-2}\beta$.

Corollary 4.1.1. The first derivative of the total variance, $\gamma_1'(k)$, approaches $-\infty$ as $k \to 0^+$ and the matrix $\mathbf{X}'\mathbf{X}$ becomes singular.

Both the theorem and the corollary are readily proved by use of $\gamma_1(k)$ and its derivative expressed in terms of λ_i.

Theorem 4.2. The squared bias $\gamma_2(k)$ is a continuous, monotonically increasing function of k.

Proof: From (4.5) $\gamma_2(k) = k^2\beta'(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-2}\beta$. If $\mathbf{\Lambda}$ is the matrix of eigenvalues of $\mathbf{X}'\mathbf{X}$ and $\mathbf{P}$ the orthogonal transformation such that $\mathbf{X}'\mathbf{X} = \mathbf{P}'\mathbf{\Lambda}\mathbf{P}$, then

$$\gamma_2(k) = k^2 \sum_1^p \alpha_i^2/(\lambda_i + k)^2 \tag{4.11}$$

$$\text{where } \alpha = \mathbf{P}\beta. \tag{4.12}$$

Since $\lambda_i > 0$ for all i and $k \geq 0$, each element $(\lambda_i + k)$ is positive and there are no singularities in the sum. Clearly, $\gamma_2(0) = 0$. Then $\gamma_2(k)$ is a continuous function for $k \geq 0$. For $k > 0$ (4.11) can be written as

$$\gamma_2(k) = \sum_1^p \alpha_i^2/[1 + (\lambda_i/k)]^2 . \tag{4.13}$$

Since $\lambda_i > 0$ for all i, the functions λ_i/k are clearly monotone decreasing for increasing k and each term of $\gamma_2(k)$ is monotone increasing. So $\gamma_2(k)$ is monotone increasing. q.e.d.

Corollary 4.2.1. The squared bias $\gamma_2(k)$ approaches $\beta'\beta$ as an upper limit.

Proof: From (4.13) $\lim_{k\to\infty} \gamma_2(k) = \sum_1^p \alpha_i^2 = \alpha'\alpha =$

$$\beta'\mathbf{P}'\mathbf{P}\beta = \beta'\beta \quad \text{q.e.d.}$$

Corollary 4.2.2. The derivative $\gamma_2'(k)$ approaches zero as $k \to 0^+$.

Proof: From (4.11) it is readily established that

$$d\gamma_2(k)/dk = 2k \sum_1^p \lambda_i\alpha_i^2/(\lambda_i + k)^3 . \tag{4.14}$$

Each term in the sum $2k\lambda_i\alpha_i^2/(\lambda_i + k)^3$ is a continuous function. And the limit of each term as $k \to 0^+$ is zero. q.e.d.

Theorem 4.3. (Existence Theorem) There always exists a $k > 0$ such that $E[L_1^2(k)] < E[L_1^2(0)] = \sigma^2 \sum_1^p (1/\lambda_i)$.

Proof: From (4.5), (4.11), and (4.14)

$$dE[L_1^2(k)]/dk = d\gamma_1(k)/dk + d\gamma_2(k)/dk$$

$$= -2\sigma^2 \sum_1^p \lambda_i/(\lambda_i + k)^3 + 2k \sum_1^p \lambda_i\alpha_i^2/(\lambda_i + k)^3 . \tag{4.15}$$

First note that $\gamma_1(0) = \sigma^2 \sum_1^p (1/\lambda_i)$ and $\gamma_2(0) = 0$. In Theorems 4.1 and 4.2

it was established that $\gamma_1(k)$ and $\gamma_2(k)$ are monotonically decreasing and increasing, respectively. Their first derivatives are always non-positive and non-negative, respectively. Thus, to prove the theorem, it is only necessary to show that there always exists a $k > 0$ such that $dE[L_1^2(k)]/dk < 0$. The condition for this is shown by (4.15) to be:

$$k < \sigma^2/\alpha_{\max}^2 \quad \text{q.e.d.} \tag{4.16}$$

c. *Some Comments On The Mean Square Error Function*

The properties of $E[L_1^2(k)] = \gamma_1(k) + \gamma_2(k)$ show that it will go through a minimum. And since $\gamma_2(k)$ approaches $\beta'\beta$ as a limit as $k \to \infty$, this minimum will move toward $k = 0$ as the magnitude of $\beta'\beta$ increases. Since $\beta'\beta$ is the squared length of the unknown regression vector, it would appear to be impossible to choose a value of $k \neq 0$ and thus achieve a smaller mean square error without being able to assign an upper bound to $\beta'\beta$. On the other hand, it is clear that $\beta'\beta$ does not become infinite in practice, and one should be able to find a value or values for k that will put $\hat{\beta}^*$ closer to β than is $\hat{\beta}$. In other words, unboundedness, in the strict mathematical sense, and practical unboundedness are two different things. In Section 7 some recommendations for choosing a $k > 0$ are given, and the implicit assumptions of boundedness are explored further.

5. A General Form of Ridge Regression

It is always possible to reduce the general linear regression problem as defined in the Introduction to a canonical form in which the $\mathbf{X}'\mathbf{X}$ matrix is diagonal. In particular there exists an orthogonal transformation $\mathbf{P}$ such that $\mathbf{X}'\mathbf{X} = \mathbf{P}'\mathbf{\Lambda}\mathbf{P}$ where $\mathbf{\Lambda} = (\delta_{ij}\lambda_i)$ is the matrix of eigenvalues of $\mathbf{X}'\mathbf{X}$. Let

$$\mathbf{X} = \mathbf{X}^*\mathbf{P} \tag{5.1}$$

and

$$\mathbf{Y} = \mathbf{X}^*\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{5.2}$$

where

$$\boldsymbol{\alpha} = \mathbf{P}\beta, \ (\mathbf{X}^*)'(\mathbf{X}^*) = \mathbf{\Lambda}, \text{ and } \boldsymbol{\alpha}'\boldsymbol{\alpha} = \beta'\beta \ . \tag{5.3}$$

Then the general ridge estimation procedure is defined from

$$\boldsymbol{\alpha}^* = [(\mathbf{X}^*)'(\mathbf{X}^*) + \mathbf{K}]^{-1}(\mathbf{X}^*)'\mathbf{Y} \tag{5.4}$$

where

$$\mathbf{K} = (\delta_{ij}k_i), \ k_i \geq 0 \ .$$

All the basic results given in Section 4 can be shown to hold for this more general formulation. Most important is that there is an equivalent to the existence theorem, Theorem 4.3. In the general form; one seeks a k_i for each canonical variate defined by $\mathbf{X}^*$. By defining $(L_1^*)^2 = (\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha})'(\hat{\boldsymbol{\alpha}}^* - \boldsymbol{\alpha})$ it can be shown that the optimal values for the k_i will be $k_i = \sigma^2/\alpha_i^2$. There is no graphical equivalent to the RIDGE TRACE but an iterative procedure initiated at $\hat{k}_i = \hat{\sigma}^2/\hat{\alpha}_i^2$ can be used. (See Section 7)

6. Relation to Other Work in Regression

Ridge regression has points of contact with other approaches to regression analysis and to work with the same objective. Three should be mentioned.

- In a series of papers, Stein [20][21] and James and Stein [13] investigated the improvement in mean square error by a transformation on $\hat{\beta}$ of the form $C\hat{\beta}$, $0 \leq C < 1$, which is a shortening of the vector $\hat{\beta}$. They show that such a $C > 0$ can always be found and indicate how it might be computed.
- A Bayesian approach to regression can be found in Jeffreys [15] and Raiffa and Schlaifer [16]. Viewed in this context, each ridge estimate can be considered as the posterior mean based on giving the regression coefficients, β, a prior normal distribution with mean zero and variance-covariance matrix $\mathbf{\Sigma} = (\delta_{ij}\delta^2/k)$. For those that do not like the philosophical implications of assuming β to be a random variable, all this is equivalent to constrained estimation by a nonuniform weighting on the values of β.
- Constrained estimation in a context related to regression can be found in [1]. For the model in the present paper, let β be constrained to be in a closed, bounded convex set C, and, in particular, let C be a hypersphere of radius R. Let the estimation criterion be minimum residual sum of squares $\phi = (\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB})$ where $\mathbf{B}$ is the value giving the minimum. Under the constraint, if $\hat{\beta}'\hat{\beta} \leq R^2$, than B is chosen to be $\hat{\beta}$; otherwise $\mathbf{B}$ is chosen to be $\hat{\beta}^*$ where k is chosen so that $(\hat{\beta}^*)'(\hat{\beta}^*) = R^2$.

7. Selecting a Better Estimate of β

In Section 2 and in the example of Section 3, it has been demonstrated that the ordinary least squares estimate of the regression vector β suffers from a number of deficiencies when $\mathbf{X}'\mathbf{X}$ does not have a uniform eigenvalue spectrum. A class of biased estimators $\hat{\beta}^*$, obtained by augmenting the diagonal of $\mathbf{X}'\mathbf{X}$ with small positive quantities, has been introduced both to portray the sensitivity of the solution to $\mathbf{X}'\mathbf{X}$ and to form the basis for obtaining an estimate of β with a smaller mean square error. In examining the properties of $\hat{\beta}^*$, it can be shown that its use is equivalent to making certain boundedness assumptions regarding either the individual coordinates of β or its squared length, $\beta'\beta$. As Barnard [12] has recently pointed out, an alternative to unbiasedness in the logic of the least squares estimator $\hat{\beta}$ is the prior assurance of bounded mean square error with no boundedness assumption on β. If it is possible to make specific mathematical assumptions about β, then it is possible to constrain the estimation procedure to reflect these assumptions.

The inherent boundedness assumptions in using $\hat{\beta}^*$ make it clear that it will not be possible to construct a clear-cut, automatic estimation procedure to produce a point estimate (a single value of k or a specific value for each k_i) as can be constructed to produce $\hat{\beta}$. However, this is no drawback to its use because with any given set of data it is not difficult to select a $\hat{\beta}^*$ that is better than $\hat{\beta}$. In fact, put in context, any set of data which is a candidate for analysis using linear regression has implicit in it restrictions on the possible values of the estimates that can be consistent with known properties of the data generator. Yet it is difficult to be explicit about these restrictions; it is especially difficult to

be mathematically explicit. In a recent paper [4] it has been shown that for the problem of estimating the mean μ of a distribution, a set of data has in it implicit restrictions on the values of σ that can be logical contenders as generators. Of course, in linear regression the problem is much more difficult; the number of possibilities is so large. First, there is the number of parameters involved. To have ten to twenty regression coefficients is not uncommon. And their signs have to be considered. Then there is $\mathbf{X}'\mathbf{X}$ and the $\binom{p}{2}$ different factor correlations and the ways in which they can be related. Yet in the final analysis these many different influences can be integrated to make an assessment as to whether the estimated values are consistent with the data and the properties of the data generator. Guiding one along the way, of course, is the objective of the study. In [12] it is shown for two problems how such an assessment can be made.

Based on experience, the best method for achieving a better estimate $\hat{\beta}^*$ is to use $k_i = k$ for all i and use the Ridge Trace to select a single value of k and a unique $\hat{\beta}^*$. These kinds of things can be used to guide one to a choice.

- At a certain value of k the system will stabilize and have the general characteristics of an orthogonal system.
- Coefficients will not have unreasonable absolute values with respect to the factors for which they represent rates of change.
- Coefficients with apparently incorrect signs at $k = 0$ will have changed to have the proper sign.
- The residual sum of squares will not have been inflated to an unreasonable value. It will not be large relative to the minimum residual sum of squares or large relative to what would be a reasonable variance for the process generating the data.

Another approach is to use estimates of the optimum values of k_i developed in Section 5. A typical approach here would be as follows:

- Reduce the system to canonical by the transformations $\mathbf{X} = \mathbf{X}^*\mathbf{P}$ and $\boldsymbol{\alpha} = \mathbf{P}\beta$.
- Determine estimates of the optimum k_i's using $\hat{k}_{i0} = \hat{\sigma}^2/\hat{\alpha}_i^2$. Use the $\hat{k}_{i0}$ to obtain $\hat{\beta}^*$.
- The $\hat{k}_{i0}$ will tend to be too small because of the tendency to overestimate $\boldsymbol{\alpha}'\boldsymbol{\alpha}$. Since use of the $\hat{k}_{i0}$ will shorten the length of the estimated regression vector, k_{i0} can be re-estimated using the $\hat{\alpha}_i^*$. This re-estimation can be continued until there is a stability achieved in $(\hat{\boldsymbol{\alpha}}^*)'(\boldsymbol{\alpha}^*)$ and $\hat{k}_{i0} = \hat{\sigma}^2/(\hat{\alpha}_i^*)^2$.

8. Conclusions

It has been shown that when $\mathbf{X}'\mathbf{X}$ is such that it has a nonuniform eigenvalue spectrum, the estimates of β in $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, based on the criterion of minimum residual sum of squares, can have a high probability of being far removed from β. This unsatisfactory condition manifests itself in estimates that are too large in absolute value and some may even have the wrong sign. By adding a small positive quanity to each diagonal element the system $[\mathbf{X}'\mathbf{X} + \mathbf{K}]\hat{\beta}^* = \mathbf{X}'\mathbf{Y}$ acts more like an orthogonal system. When $\mathbf{K} = k\mathbf{I}$ and all solutions in the interval $0 \leq k \leq 1$ are obtained, it is possible to obtain a two-dimensional characteriza-

tion of the system and a portrayal of the kinds of difficulties caused by the intercorrelations among the predictors. A study of the properties of the estimator $\hat{\beta}^*$ shows that it can be used to improve the mean square error of estimation, and the magnitude of this improvement increases with an increase in spread of the eigenvalue spectrum. An estimate based on $\hat{\beta}^*$ is biased and the use of a biased estimator implies some prior bound on the regression vector β. However, the data in any particular problem has information in it that can show the class of generators β that are reasonable. The purpose of the ridge trace is to portray this information explicitly and, hence, guide the user to a better estimate $\hat{\beta}^*$.

Nomenclature

$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
$\hat{\beta}^* = \hat{\beta}^*(k) = [\mathbf{X}'\mathbf{X} + k\mathbf{I}]^{-1}\mathbf{X}'\mathbf{Y}; k > 0$
$\mathbf{W} = \mathbf{W}(k) = [\mathbf{X}'\mathbf{X} + k\mathbf{I}]^{-1}$
$\mathbf{Z} = \mathbf{Z}(k) = [\mathbf{I} + k\,(\mathbf{X}'\mathbf{X})^{-1}]^{-1} = \mathbf{I} - k\mathbf{W}$
λ_i = Eigenvalue of $\mathbf{X}'\mathbf{X}$; $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$
$\mathbf{\Lambda} = (\delta_{ij}\lambda_i)$ = the matrix of eigenvalues
$\mathbf{P}$ = An orthogonal matrix such that $\mathbf{P}'\mathbf{\Lambda}\mathbf{P} = \mathbf{X}'\mathbf{X}$
$L_1^2(k) = E[\,(\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta)\,] = \gamma_1(k) + \gamma_2(k)$
$\gamma_1(k)$ = Variance of the estimate $\hat{\beta}^*$
$\gamma_2(k)$ = Squared bias of the estimate $\hat{\beta}^*$
$\mathbf{K} = (\delta_{ij}k_i)$; $k_i \geq 0$ A diagonal matrix of non-negative constants.
$\alpha = \mathbf{P}\beta$
$\mathbf{X}^* = \mathbf{X}\mathbf{P}'$
$\hat{\alpha}^* = [\,(\mathbf{X}^*)'(\mathbf{X}^*) + \mathbf{K}]^{-1}\,(\mathbf{X}^*)'\mathbf{Y}$
$\mathbf{W}^* = [\,(\mathbf{X}^*)'(\mathbf{X}^*) + \mathbf{K}]^{-1}$
$\mathbf{Z}^* = \{\mathbf{I} + [\,(\mathbf{X}^*)'(\mathbf{X}^*)\,]^{-1}\,\mathbf{K}]\} = \mathbf{I} - \mathbf{K}\mathbf{W}$

References

[1] Balakrishnan, A. V. (1963). An operator theoretic formulation of a class of control problems and a steepest descent method of solution. *Journal on Control 1*, 109–127.

[2] Barnard, G. A. (1963). The logic of least squares. *Journal of the Royal Statistical Society, Series B 25*, 124–127.

[3] Beale, E. M. L., Kendall, M. G., and Mann, D. W. (1967). The discarding of variables in multivariate analysis. *Biometrika 54*, 356–366.

[4] Clutton-Brock, M. (1965). Using the observations to estimate prior distribution. *Journal of the Royal Statistical Society, Series B 27*, 17–27.

[5] Efroymson, M. A. (1960). Multiple regression analysis. Chapter 17 in *Mathematical Methods for Digital Computers*. Edited by A. Ralston and H. S. Wilf, John Wiley & Sons, Inc., New York.

[6] Garside, M. J. (1965). The best subset in multiple regression analysis. *Applied Statistics 14*.

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

[8] Hocking, R. R. and Leslie, R. N. (1967). Selection of the best subset in regression analysis. *Technometrics 9*, 531–540.

[9] Hoerl, A. E. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress 58*, 54–59.

[10] HOERL, A. E. (1964). Ridge analysis. *Chemical Engineering Progress Symposium Series 60*, 67–77.

[11] HOERL, A. E. and KENNARD, R. W. (1968). On regression analysis and biased estimation. *Technometrics 10*, 422–423. Abstract.

[12] HOERL, A. E. and KENNARD, R. W. (1970). Ridge Regression. Applications to non-orthogonal problems. *Technometrics 12*.

[13] JAMES, W. and STEIN, C. M. (1961). Estimation with quadratic loss. *Proc. 4th Berkeley Symposium 1*, 361–379.

[14] JEFFERS, J. N. R. (1967). Two case studies in the application of principal component analysis. *Applied Statistics 16*, 225–236.

[15] JEFFREYS, H. (1961). *Theory of Probability*. Third Edition, Oxford University Press, London, Chapter III.

[16] RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory*, Harvard University, Boston, Chapters 11 and 13.

[17] RILEY, J. D. (1955). Solving systems of linear equations with a positive definite, Symmetric, but possibly ill-conditioned matrix. *Mathematics of Computation 9*, 96–101.

[18] SCHEFFÉ, H. (1960). *The Analysis of Variance*. John Wiley & Sons, Inc., New York, Chapters 1 and 2.

[19] SCOTT, J. T., Jr. (1966). Factor analysis and regression. *Econometrica 34*, 552–562.

[20] STEIN, C. M. (1960). Multiple regression. Chapter 37 in *Essays in Honor of Harold Hotelling*, Stanford University Press.

[21] STEIN, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution. *Journal of the Royal Statistical Society, Series B, 24*, 265–296.