1976-03

An investigation of the probability distribution of the ridge regression estimator for linear models.

Lewis, Edgar Barry

Monterey, California. Naval Postgraduate School

AN INVESTIGATION OF THE PROBABILITY DISTRIBUTION OF THE RIDGE REGRESSION ESTIMATOR FOR LINEAR MODELS

Edgar Barry Lewis

NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

AN INVESTIGATION OF THE PROBABILITY DISTRIBUTION OF THE RIDGE REGRESSION ESTIMATOR FOR LINEAR MODELS

by

Edgar Barry Lewis

March 1976

Thesis Advisor: H. J. Larson

Approved for public release; distribution unlimited.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) An Investigation of the Probability Distribution of the Ridge Regression Estimator for Linear Models		5. TYPE OF REPORT & PERIOD COVERED Master's Thesis; March 1976 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Edgar Barry Lewis		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940		12. REPORT DATE March 1976 13. NUMBER OF PAGES 37
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Naval Postgraduate School Monterey, California 93940		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Probability Distribution
Ridge Regression
Estimator

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The estimation of the parameters of a linear statistical model is generally accomplished by the method of least squares. However, when the method of least squares is applied to non-orthogonal problems the resulting estimates may be significantly different from the true parameters. The method of ridge regression may provide better estimates in these cases; however, a probability distribution of the ridge estimator is

presently not known. The form of such a distribution is dependent upon how the ridge parameter, k, is selected. Two possible objective methods of choosing k are examined to determine if either one leads to a useful probability distribution.

An Investigation of the Probability Distribution of the Ridge Regression Estimator for Linear Models

by

Edgar Barry Lewis
Lieutenant, United States Navy
B.S.E.E., University of New Mexico, 1967

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
March 1976

ABSTRACT

The estimation of the parameters of a linear statistical model is generally accomplished by the method of least squares. However, when the method of least squares is applied to nonorthogonal problems the resulting estimates may be significantly different from the true parameters. The method of ridge regression may provide better estimates in these cases; however, a probability distribution of the ridge estimator is presently not known. The form of such a distribution is dependent upon how the ridge parameter, k, is selected. Two possible objective methods of choosing k are examined to determine if either one leads to a useful probability distribution.

TABLE OF CONTENTS

LIST OF FIGURES

I. BACKGROUND

The following conventions will be used throughout. Unless otherwise noted, capital letters and Greek letters will refer to matrices and vectors while lower case letters will refer to scalars.

A. INTRODUCTION

The use of linear statistical models is widespread in scientific fields of all kinds. Generally, the linear statistical model is postulated as

$$Y = X\beta + \varepsilon \qquad (1)$$

where Y is an n x 1 vector of n observed values of a dependent variable, X is an n x p matrix containing n values for each of p predictor (independent) variables, β is a p x 1 vector of p unknown parameters (or coefficients) to be estimated from data, and ε is an n x 1 vector representing experimental errors. Usually, the experimental error is assumed to have a multivariate normal distribution with mean equal to zero and variance covariance matrix equal to $\sigma^2 I$ where σ^2 is the scalar value of the common variance of the experimental errors. This assumption will be made throughout this paper.

In practice, the modeling problem is to estimate the parameters β from data Y and X. The most common method of

doing this is called least squares estimation or sometimes ordinary least squares (OLS). The latter designation will be used in this paper.

Under certain fairly general and common conditions OLS is an adequate method of estimating β. However, when the data is "ill-conditioned" or nonorthogonal OLS may yield poor estimates of the true parameters.

Ridge regression (RR) has been proposed [Ref. 1] as an alternative estimation method that might yield better estimates under conditions where OLS does poorly.

B. ORDINARY LEAST SQUARES

For convenience, it is assumed that the elements of X are scaled such that X'X has the form of a correlation matrix. This is done by forming from each element x_{ij} a new element x'_{ij} such that

$$x'_{ij} = (x_{ij} - \bar{x}_j)/s_{x_j} \tag{2}$$

where $\bar{x}_j$ is the mean value of the elements of the $j^{\underline{th}}$ independent variable and s_{x_j} is its standard deviation times an appropriate constant such that the diagonal elements of X'X are equal to one. The OLS estimator of β is then

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{3}$$

so long as $(X'X)^{-1}$ exists.[1] The estimator $\hat{\beta}$ is unique, unbiased and is the best linear unbiased estimator (BLUE) of β (it has the minimum variance among all linear unbiased estimators of β) so long as $E(Y) = X\beta$ and $E(Y - X\beta)(Y - X\beta)' = \sigma^2 I$ where σ^2 is a scalar, as assumed previously.

The OLS estimator $\hat{\beta}$ is commonly used and is particularly useful when it can be assumed that Y is a multivariate normal vector with mean vector $X\beta$ and covariance matrix $\sigma^2 I$. In this case, it can be shown[2] that the maximum likelihood estimator of β is the same as the OLS estimator and furthermore, since $\hat{\beta}$ is a linear function of the elements of Y, $\hat{\beta}$ has a multivariate normal distribution with mean vector equal to β and covariance matrix $\sigma^2(X'X)^{-1}$. This latter characteristic of $\hat{\beta}$ allows the use of hypothesis tests and the computation of confidence bounds.

Unfortunately, in some cases X'X is "ill-conditioned" and OLS yields poor estimates. This typically occurs when an experiment is poorly designed or there are economic or physical restraints causing strong correlations among the predictor variables. In this case X'X, in its correlation matrix form, will not be orthogonal.

---

[1]For a derivation and details of properties of the OLS estimator, see, for example, Ref. 2.

[2]For example, see Ref. 2, page 182.

Hoerl and Kennard [Ref. 3] address the eigenvalues of X'X (denoted by λ_j, j = 1, 2, ... p) and point out that nonorthogonal data are characterized by the smallest eigenvalue (λ_{min}) being much less than unity and that, since σ^2/λ_{min} is a lower bound for the mean squared distance between $\hat{\beta}$ and β, then for X'X nonorthogonal, the difference between $\hat{\beta}$ and β has a high probability of being large. When X'X is nonorthogonal $\hat{\beta}$ is characterized by one or more of the following difficulties, for example:

(1) large variance,

(2) large magnitude of residual errors,

(3) incorrect signs of parameter estimates.

C. RIDGE REGRESSION

A. E. Hoerl suggested [Refs. 1 and 4] that the large variance of $\hat{\beta}$ for nonorthogonal data could be reduced by the addition of a constant $k \geq 0$ to the diagonal elements of X'X, thus yielding

$$\hat{\beta}^* = (X'X + kI)^{-1} X'Y \qquad (4)$$

as as estimator. Equation (4) is derived in Appendix A. Note that for k equal to zero the estimator $\hat{\beta}^*$ is equal to the OLS estimator $\hat{\beta}$. Therefore, OLS can be thought of as a special case of ridge regression.[3] Hoerl suggested

[3]See Appendix B for a discussion of an even more general estimator.

the name "ridge regression" for this procedure because of its mathematical similarity to some of his earlier work [Ref. 5] on quadratic response functions. Appendix A contains a derivation of the ridge regression estimator.

1. Mean Squared Error

The rationale behind using the ridge estimator is to minimize the mean squared error (MSE) associated with the estimate instead of minimizing the sum of squares of residuals as is done in OLS.[4] Hoerl and Kennard show that the mean squared error is given by

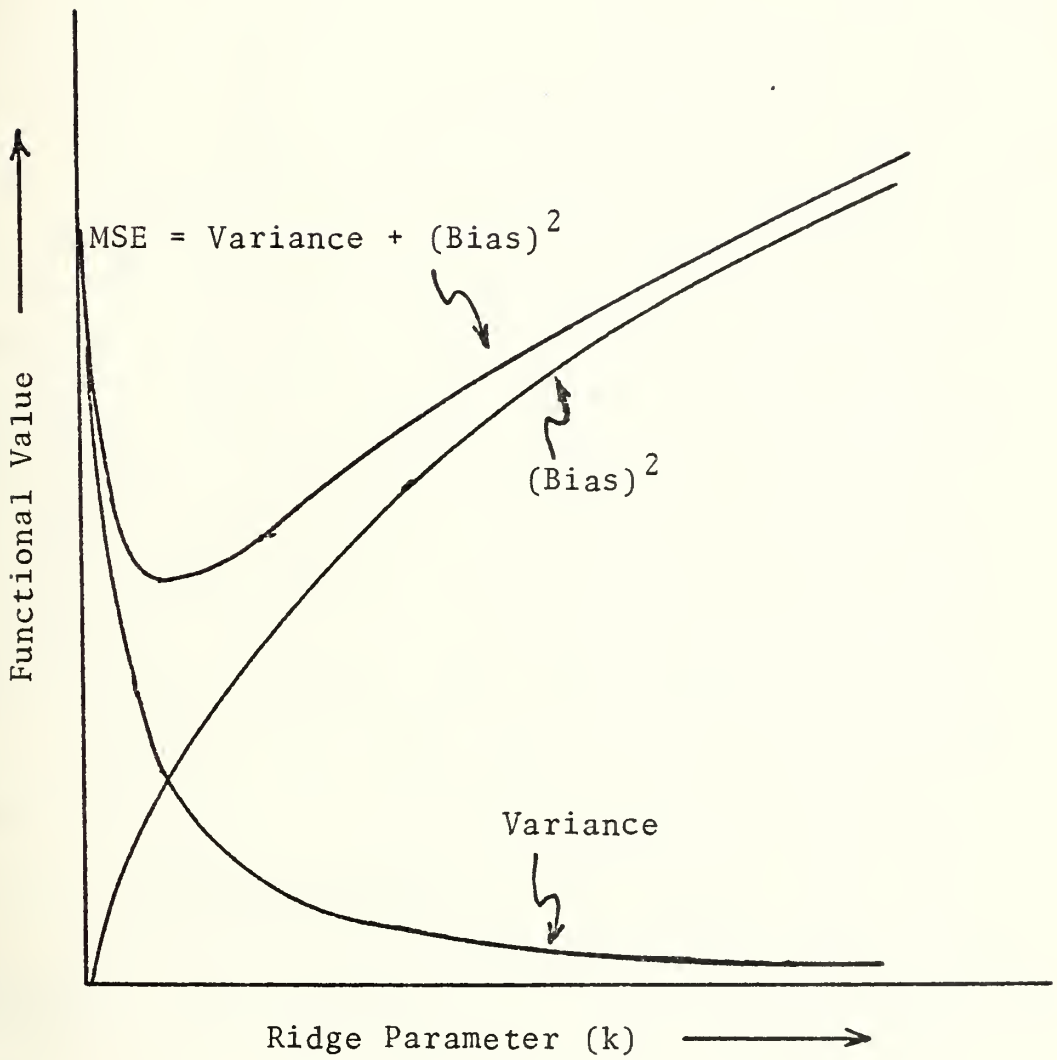
$$\text{MSE} = \text{Variance} + (\text{Bias})^2 \tag{5}$$

Furthermore, they show that variance is a monotonically decreasing function of k, that the squared bias is a monotonically increasing function of k and that the rate of change of variance, for nonorthogonal data and small k, is considerably larger than the rate of change of the squared bias. Figure 1 is a graphical illustration of these relationships. Hoerl and Kennard argue that it is possible to find some $k \geq 0$ such that the variance is greatly reduced while only a small amount of bias is introduced, thus yielding a smaller MSE than if OLS ($k = 0$)

[4] In the case of unbiased estimation, which OLS is, these are equivalent criteria.

FIGURE 1

MEAN SQUARED ERROR FUNCTIONS



were used. Indeed they show that if $\beta'\beta$ is bounded, then such a k always exists. Thus, proper use of ridge regression on nonorthogonal data insures a reduced MSE of estimation.

The problem remains to select an appropriate value of k. Hoerl and Kennard [Ref. 6] suggest the use of two graphical devices as aids to determining an appropriate value of k. The first is the ridge trace, a two-dimensional plot of the elements of $\hat{\beta}^*$ as functions of k and the second is an estimate of the squared length of the coefficient vector $\hat{\beta}^{*\prime}\hat{\beta}^*$. The ridge trace is used to gain an understanding of the underlying correlations between the various predictor variables while the plot of $\hat{\beta}^{*\prime}\hat{\beta}^*$ is used to subjectively determine a suitable range of values of k. A typical ridge trace is illustrated in Figure 2 and a typical plot of $\hat{\beta}^{*\prime}\hat{\beta}^*$ is depicted in Figure 3. Notice that $\hat{\beta}^{*\prime}\hat{\beta}^*$, in Figure 3, decreases steeply for small k ($k < 0.2$) but in the range about 0.3 to 0.4 has become much less sensitive to further increases in k.

2. Alternative Methods of Choosing k

The previously described method of subjectively choosing a suitable value of k is the current method in use and appears to be useful. A major problem arises, however, because the method denies to the analyst knowledge of the probability distribution of $\hat{\beta}^*$ and, therefore, any probabilistic inferences concerning the resulting

FIGURE 2

TYPICAL RIDGE TRACE

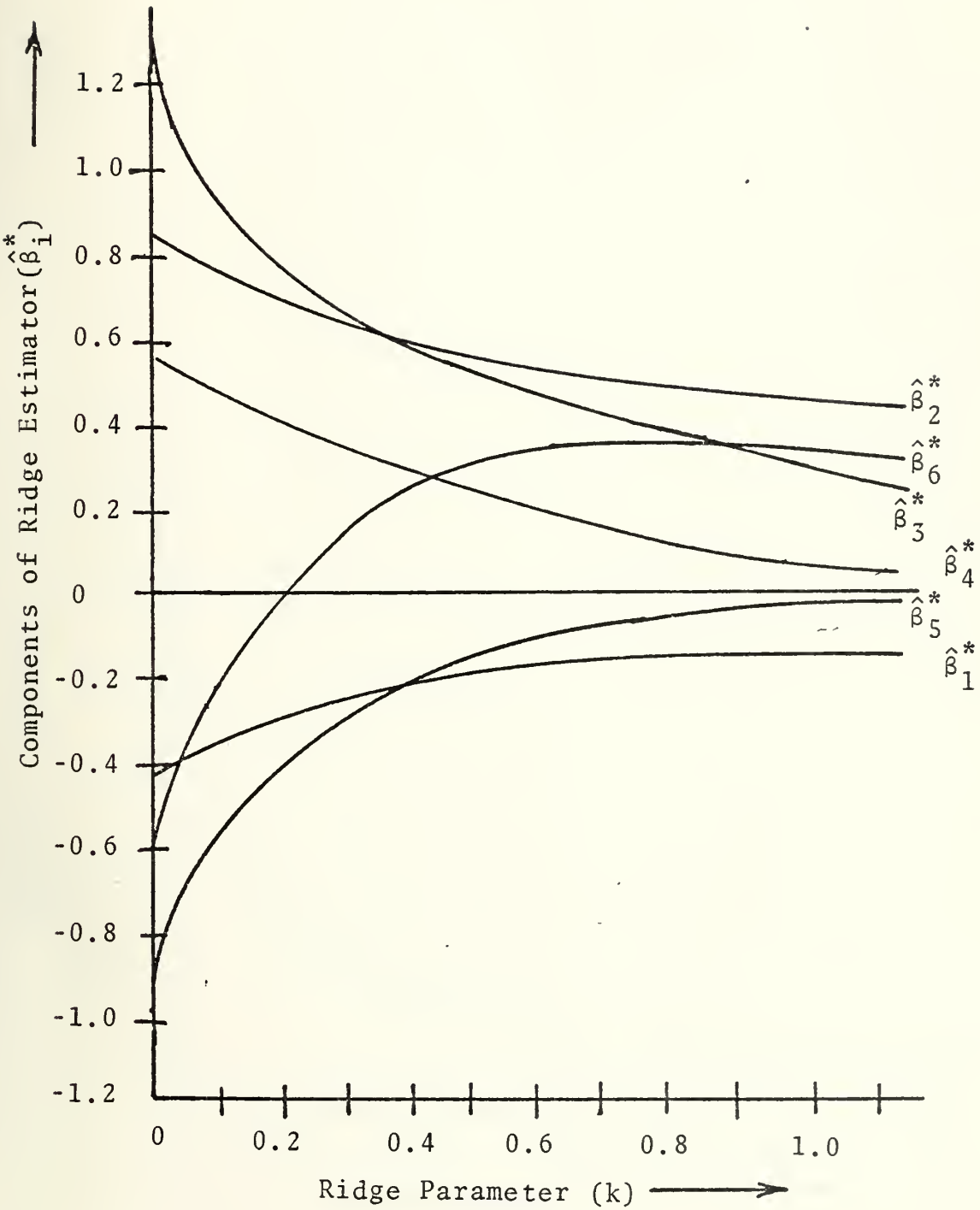
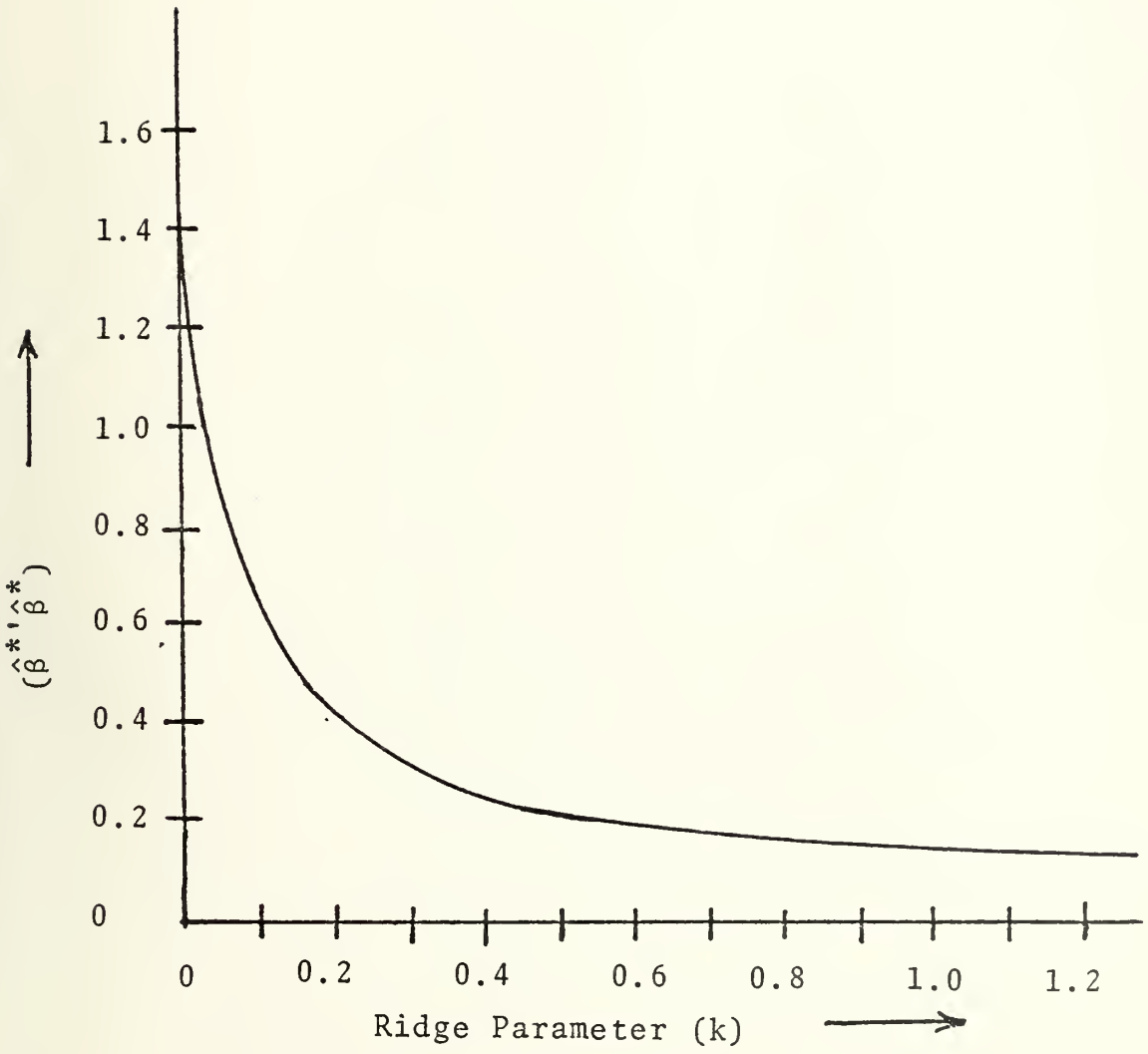


FIGURE 3

TYPICAL PLOT OF THE

SQUARED LENGTH OF THE RIDGE ESTIMATOR

$(\hat{\beta}^{*\prime}\hat{\beta}^{*})$



estimator. Hoerl and Kennard have suggested a general form of ridge regression [Ref. 3] and an iterative method of determining k. In addition, Hemmerle [Ref. 7] has derived a closed form solution based on this method. Another possibility is to use the ridge trace or the plot of $\hat{\beta}^{*'}\hat{\beta}^{*}$ quantitatively to calculate a point value for k in such a way that the marginal probability distribution, $f_{\hat{\beta}^{*}}$, may be determined. Two such methods using the ridge trace are examined in the next section.

II. PROPOSED OBJECTIVE RULES FOR CHOOSING k

The slope (rate of change) of the ridge trace curves or the absolute change of the ridge trace curves over a specified interval may be used to determine a value of the ridge parameter, k, objectively. These criteria are discussed here.

Either of these criteria may be sensitive to the behavior of each coefficient $\hat{\beta}_i^*$. In general, $\hat{\beta}_i^*$ is not monotonic in k, although they all approach zero as k is increased without bound. It has been noted by Marquardt and Snee [Ref. 8] that it is not uncommon for one or more $\hat{\beta}_i^*$ to increase in absolute value as k is increased. (See, for example, $\hat{\beta}_6^*$ in Figure 2.) Therefore, the ridge trace should be examined by the analyst to detect any behavior of $\hat{\beta}_i^*$ that might adversely affect the proper selection of k even though the ridge trace is not to be used directly to select a specific value of k.

It is clear that $\hat{\beta}^*$ is distributed multivariate normal if Y is distributed multivariate normal and a specific value of k is selected a priori. However, whenever the value of k is dependent on a data sample its value will not generally be the same for each data sample. Therefore, k is a random variable. Let K denote this (scalar) random variable.

The marginal probability distribution of $\hat{\beta}^*$ may be derived from the joint probability distribution of K and $\hat{\beta}^*$ which can be determined by

$$f_{\hat{\beta}^*,K} = f_{\hat{\beta}^*/K} \cdot f_K \tag{6}$$

if the conditional distribution of $\hat{\beta}^*$ given K, $f_{\hat{\beta}^*/K}$, and the marginal distribution of K, f_K, are known. As stated above, when K is given, the distribution of $\hat{\beta}^*$ is known. It remains to determine the marginal of K, f_K. Clearly, this distribution depends on how K is related to Y. The procedure will be to find a mapping from the range of Y into the range of K which gives the marginal distribution of K. With this distribution and the known conditional distribution of $\hat{\beta}^*$ given K, the joint distribution of $\hat{\beta}^*$ and K may be determined. It is convenient to consider the cumulative distribution function, $F_K(k)$, since, if the functional relationship of K to Y, K = h(Y), is known then

$$F_K(k) = P[K \leq k] = P[h(Y) \leq k] = P[Y \varepsilon R_k] \tag{7}$$

where R_k is a region in the space of Y corresponding to $h(Y) \leq k$. Thus if R_k can be determined then, since the marginal distribution of Y is known, $F_K(k) = P[Y \varepsilon R_k]$ can be determined and f_K may be determined from F_K by differentiation. It remains to determine R_k corresponding

to a specified region in the space of K and an objective rule for mapping from Y to K.

A. ABSOLUTE VALUE CRITERION

The practical range of the ridge parameter is taken to be $0 < k \leq 1$ in the literature. It seems reasonable then to choose the smallest value of k such that all $\hat{\beta}_i^*(k)$ are close to their respective values at k = 1. In other words,

$$|\hat{\beta}_i^*(k) - \hat{\beta}_i^*(1)| \leq \delta_i; \; i = 1, 2, \ldots, p \qquad (8)$$

where δ_i is a constant selected by the analyst. The criterion expressed by (8) means that the ridge trace curves, $\hat{\beta}_i^*$, at k are within δ_i of their value at k = 1 beyond which there is no interest. Here δ_i refers to the $i^{\underline{th}}$ scalar component of a p x 1 vector, δ. Suppose that at some $k = k_0$ the $m^{\underline{th}}$ component of the left hand size of (1) is the one whose absolute magnitude is largest. Define a p x 1 vector τ such that $\tau_m = \pm\delta_m$, as appropriate, and the other components of τ are equal to the corresponding values of $|\hat{\beta}_i^*(k_0) - \hat{\beta}_i^*(1)|$. Then equation (8) can be rewritten in vector form

$$\hat{\beta}^*(k_0) - \hat{\beta}^*(1) = \tau \qquad (9)$$

B. DERIVATIVE CRITERION

Another potential criterion to use for selecting k is to require that the slopes of all $\hat{\beta}_i^*$ be "flat enough" in the sense that

$$\frac{\partial \hat{\beta}_i^*(k)}{\partial k} = \delta_i; \; i = 1, 2, \ldots, p \qquad (10)$$

where δ_i is as previously defined. Define m such that the $m^{\underline{th}}$ component of the left hand side of (10) is the one whose absolute magnitude is largest and define a p x 1 vector π such that $\pi_m = \pm\delta_m$, as appropriate, and the other components of π are equal to the corresponding values of $|\frac{\partial \hat{\beta}^*}{\partial k}|$.

Then equation (1) can be written, in vector form

$$\frac{\partial \hat{\beta}^*(k)}{\partial k} = \pi \qquad (11)$$

III. PROBLEM

The problem is to determine the probability distribution of K given Y. It is proposed to determine this by attempting to derive and examine the functional relationship of Y and K.

A. ABSOLUTE VALUE CRITERION

The criterion expressed by equation (9) may be stated, by substituting from equation (4)

$$(X'X + kI)^{-1}X'Y - (X'X + I)^{-1}X'Y = \tau \qquad (12)$$

and by factoring

$$[(X'X + kI)^{-1} - (X'X + I)^{-1}]X'Y = \tau \qquad (13)$$

but, as shown in Appendix C, equation (C-4), the expression in brackets may be expanded to

$$(X'X + kI)^{-1}[(X'X + I) - (X'X + kI)]\ (X'X + I)^{-1} \qquad (14)$$

Therefore, by canceling terms and simplifying, equation (13) becomes

$$(1 - k)(X'X + kI)^{-1}(X'X + I)^{-1}X'Y = \tau \qquad (15)$$

If $k \neq 1$ and if $(X'X + kI)^{-1}$ and $(X'X + I)^{-1}$ exist, then

$$\boxed{X'Y = \left(\frac{1}{1 - k}\right)(X'X + kI)(X'X + I)\tau} \qquad (16)$$

The task then is to solve the linear equations in (16) for Y in order to determine R_k. Unfortunately, equation (18) represents p linear restraints (hyperplanes) on n unknown variables where, in general, $n > p$. Furthermore, τ is a function of Y. Thus, R_k is not easily determined under this criterion.

B. DERIVATIVE CRITERION

The criterion given by equation (11) may be stated by substituting from equation (4)

$$\frac{\partial}{\partial k}[(X'X + kI)^{-1}X'Y] = \pi \qquad (17)$$

or since $\frac{\partial}{\partial k}(X'X + kI) = I$

$$-(X'X + kI)^{-2}X'Y = \frac{\partial \pi}{\partial k} \qquad (18)$$

Now, if $(X'X + kI)$ is not singular then

$$\boxed{X'Y = (X'X + kI)^2 \frac{\partial \pi}{\partial k}} \qquad (19)$$

where the negative sign has been dropped since the criterion actually specifies the absolute value of the components of the derivative and the notation of π accounts for proper signs.

Equation (19) is similar to equation (16), as it should be since the criteria are similar, and the same difficulties are encountered in determining R_k as for the previous criterion. In addition, the derivative of π will be difficult to determine. Therefore, the derivative criterion does not lead to a useful result either.

IV. NOTES ON THE FULL BAYESIAN RIDGE ESTIMATOR

The full Bayesian ridge estimator (FBRE) is suggested by Eskew [Ref. 9] and is given as

$$\underline{\hat{\beta}}^{*} = (X'X + kI)^{-1}(X'Y + k\beta_0) \tag{20}$$

where β_0 is a prior estimate of β. There are two interesting properties of $\underline{\hat{\beta}}^{*}$ not noted by Eskew.

First suppose that the prior β_0 is chosen to be the OLS estimate $\hat{\beta}$. Then

$$\underline{\hat{\beta}}^{*} = (X'X + kI)^{-1}[X'Y + k(X'X)^{-1}X'Y] \tag{21}$$

and hence

$$\underline{\hat{\beta}}^{*} = (X'X + kI)^{-1}[I + k(X'X)^{-1}]X'Y \tag{22}$$

But

$$[I + k(X'X)^{-1}] = (X'X + kI)(X'X)^{-1} \tag{23}$$

Substituting (23) into (22)

$$\underline{\hat{\beta}}^{*} = (X'X)^{-1}X'Y = \hat{\beta} \tag{24}$$

Thus if the OLS estimator is used as a prior estimate for the FBRE, equation (21), then the resulting estimate is equal to the OLS estimate.

Now, suppose that any prior estimate β_0 is used in equation (21) but the resulting estimate is then used as a prior in (21) to compute another estimate. If this procedure is repeated indefinitely, in the limit the result will again be the OLS estimator regardless of what prior, β_0, was initially used. The proof of this is shown in Appendix B.

V. CONCLUSIONS AND RECOMMENDATIONS

A. CONCLUSIONS

The determination of a probability distribution of the ridge estimator, $\hat{\beta}^*$, is desirable in order to facilitate the use of hypothesis tests and the computation of confidence bounds concerning $\hat{\beta}^*$. The probability distribution of $\hat{\beta}^*$ depends on the objective rule used to select the ridge parameter, k. Neither of the two objective rules examined here appears to lead to a simply determined probability distribution.

B. RECOMMENDATIONS

The search for a useful probability distribution of $\hat{\beta}^*$ should be pursued further. In particular, the closed form solution for k presented by Hemmerle [Ref. 7] may prove fruitful. Other possibilities include investigating other criteria based on the ridge trace such as minimizing the sum of squares, over all $i = 1, 2, \ldots, p$, of the difference between $\hat{\beta}_i^*(k)$ and $\hat{\beta}_i^*(1)$. Also, the same criteria applied to the ridge trace could be considered for the squared length of $\hat{\beta}^*$.

APPENDIX A

DERIVATION OF THE RIDGE REGRESSION ESTIMATOR

The residual sum of squares for any estimator can be written

$$\Phi(\beta) = (Y - X\beta)' (Y - X\beta) = \varepsilon'\varepsilon \qquad \text{(A-1)}$$

In ridge regression it is desirable to minimize the residual sum of squares subject to an acceptable length, c, of the regression vector $\hat{\beta}^*$. Expressed as a Lagrangian restraint problem this is

$$\min \Phi'(\hat{\beta}^*) = (Y - X\hat{\beta}^*)' (Y - X\hat{\beta}^*) + k(\hat{\beta}^{*\prime}\hat{\beta}^* - c^2) \qquad \text{(A-2)}$$

where k is the inverse of the Lagrangian multiplier.

Taking partial derivatives of Φ' with respect to $\hat{\beta}^*$ and setting them equal to zero

$$\frac{\partial \Phi'}{\partial \hat{\beta}^*} = 0$$

$$= \frac{\partial}{\partial \hat{\beta}^*} [Y'Y - Y'X\hat{\beta}^* - \hat{\beta}^{*\prime}X'Y + \hat{\beta}^{*\prime}X'X\hat{\beta}^* + k\hat{\beta}^{*\prime}\hat{\beta}^*] \qquad \text{(A-3)}$$

Hence

$$0 = -(Y'X)' - X'Y + 2X'X\hat{\beta}^* + 2k\hat{\beta}^* \qquad \text{(A-4)}$$

or

$$2X'Y = 2X'X\hat{\beta}^* + 2kI\hat{\beta}^* \qquad \text{(A-5)}$$

Therefore,

$$X'Y = (X'X + kI)\hat{\beta}^* \qquad \text{(A-6)}$$

Now, if $(X'X + kI)$ is non-singular (which k is selected to ensure), then

$$\hat{\beta}^* = (X'X + kI)^{-1}X'Y \qquad \text{(A-7)}$$

APPENDIX B

FULL BAYESIAN RIDGE ESTIMATION

A. BACKGROUND

Eskew [Ref. 9] points out that ridge estimation is equivalent to minimizing the squared differences between the regression estimates and a prior estimate of zero subject to a constraint on the sum of squares and suggests that a non-zero prior might be more reasonable. Following this line of reasoning he derives the full Bayesian ridge estimator (FBRE)

$$\underline{\hat{\beta}}^{*} = (X'X + kI)^{-1}(X'Y + k\beta_0) \qquad \text{(B-1)}$$

where β_0 is a prior estimate of the true parameters β. Note that the ridge estimator is a special case of FBRE where the prior is taken to be zero.

Eskew shows that the variance of the FBRE is the same as the variance of the ridge regression estimator (RRE) while the squared bias of the FBRE is less than that for the RRE, thereby resulting in a reduction of mean squared error.

B. ITERATIVE USE OF THE FULL BAYESIAN RIDGE ESTIMATOR

Suppose that the FBRE is calculated using any prior, β_0, and then the result, $\underline{\hat{\beta}}_1^{*}$, is used as a prior to calculate another FBRE, $\underline{\hat{\beta}}_2^{*}$. If this procedure is repeated

m times the result may be written

$$\underline{\beta}_m^* = (1/k) \sum_{i=1}^{m} (kA)^i X'Y + (kA)^m \beta_0 \qquad \text{(B-2)}$$

where $A = (X'X + kI)^{-1}$. It is interesting to determine the form of $\hat{\underline{\beta}}_m^*$ in the limit as m approaches infinity. Since A and X'X are positive definite matrices their eigenvalues are positive. Let $\lambda_i > 0$ be an eigenvalue of A and $\rho_i > 0$ be an eigenvalue of X'X. Hoerl and Kennard show the relationship between λ_i and ρ_i to be

$$\lambda_i = 1/(\rho_i + k) \qquad \text{(B-3)}$$

Now there exists an orthogonal p x p matrix P with P'P = I such that

$$P'AP = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p) \qquad \text{(B-4)}$$

or since the eigenvalues of kA are $k\lambda_i$ and the eigenvalues of A^m are $(\lambda_i)^m$

$$P'(kA)^m P = \text{diag}(k_1^m \lambda_1^m, k_2^m \lambda_2^m, \ldots, k_p^m \lambda_p^m) \qquad \text{(B-5)}$$

Now

$$P'[\lim_{m\to\infty} (kA)^m]P = \lim_{m\to\infty} P'(kA)^m P \qquad \text{(B-6)}$$

The right hand side of (B-6) is the limit of the right hand side of (B-5). By substituting from equation (B-3) a typical diagonal element is $0 \leq [k/(\rho_i + k)]^m < 1$, since $\rho_i > 0$ for all $i = 1, 2, \ldots, P$. Therefore, each of the elements of the right hand side of (B-5) approaches zero as m approaches infinity. Hence

$$P' \lim_{m\to\infty} (kA)^m P = 0 \tag{B-7}$$

This can only occur if

$$\lim_{m\to\infty} (kA)^m = 0 \tag{B-8}$$

Therefore, the last term of equation (B-2) is zero in the limit. Now define a matrix function $S = S(kA)$ where

$$S = \sum_{i=1}^{\infty} (kA)^i \tag{B-9}$$

DeRusso, Roy, and Close [Ref. 10] show that S(kA) converges if and only if $S(k\lambda_i)$ converges for all $k\lambda_i$, the eigenvalues of kA. Clearly this will occur if and only if

$$|k\lambda_{max}| < 1 \tag{B-10}$$

Substituting equation (B-3)

$$|k/(\rho_{min} + k)| < 1 \qquad \text{(B-11)}$$

or, after some algebra

$$\rho_{min} > -2k \quad \text{and} \quad \rho_{min} > 0 \qquad \text{(B-12)}$$

Since ρ_{min} is an eigenvalue of a positive definite matrix, X'X, then $\rho_{min} > 0$ and both conditions of (B-13) are met. Therefore S(kA) does converge. To see what it converges to, define S' = S + I and multiply S' on the left by (I - kA)

$$(I - kA)S' = (I - kA)(I + kA + (kA)^2 + \ldots) \qquad \text{(B-13)}$$

and multiplying the right hand side out

$$(I - kA)S' = [I + kA + (kA)^2 + \ldots] - [kA + (kA)^2 + \ldots]$$

$$= I \qquad \text{(B-14)}$$

Then

$$S' = (I - kA)^{-1} \qquad \text{(B-15)}$$

Then

$$S = [I[(kA)^{-1} - I]kA]^{-1} - I$$

$$= (1/k)A^{-1}[(1/k)A^{-1} - I]^{-1} - I \qquad \text{(B-16)}$$

Substituting $A = (X'X + kI)^{-1}$

$$S = [(1/k)X'X + I][(1/k)X'X]^{-1} - I$$

$$= k(X'X)^{-1} \qquad \text{(B-17)}$$

Substituting S into equation (B-2)

$$\lim_{m\to\infty} \hat{\beta}_m^* = (1/k)k(X'X)^{-1}X'Y \qquad \text{(B-18)}$$

Therefore

$$\lim_{m\to\infty} \hat{\beta}_m^* = (X'X)^{-1}X'Y = \hat{\beta} \qquad \text{(B-19)}$$

Thus the iterative procedure, starting with any prior β_0, converges to the OLS estimator, $\hat{\beta}$.

APPENDIX C

MISCELLANEOUS MATRIX ALGEBRA AND CALCULUS

Let A, B, and C denote m x n matrices. Denote their inverses by A^{-1}, B^{-1}, and C^{-1}, respectively.

A. MATRIX ALGEBRA

First, note that

$$C(A + B)^{-1} = (AC^{-1} + BC^{-1})^{-1} \tag{C-1}$$

since

$$C(A + B)^{-1} = [(A + B)C^{-1}]^{-1} \tag{C-2}$$

$$= (AC^{-1} + BC^{-1})^{-1} \tag{C-3}$$

Also

$$A^{-1} \pm B^{-1} = A^{-1}(B \pm A)B^{-1} = B^{-1}(B \pm A)A^{-1} \tag{C-4}$$

since

$$A^{-1}(B \pm A)B^{-1} = (A^{-1}B \pm I)B^{-1}$$

$$= (A^{-1} \pm B^{-1}) \tag{C-5}$$

and

$$B^{-1}(B \pm A)A^{-1} = (I \pm B^{-1}A)A^{-1}$$

$$= (A^{-1} \pm B^{-1}) \qquad \text{(C-6)}$$

B. MATRIX CALCULUS

Let A(t), B(t), and C(t) denote m x n matrices whose elements may be functions of the scalar variable t. Let $\dot{A}(t)$ and $\dot{B}(t)$ denote the derivatives of A(t) and B(t), respectively, with respect to t.

The following are shown to be true by DeRusso, Roy, and Close [Ref. 10].

$$\frac{d}{dt} A(t)B(t) = \dot{A}(t)B(t) + A(t)\dot{B}(t) \qquad \text{(C-7)}$$

and

$$\frac{d}{dt} A^{-1}(t) = -A^{-1}(t)\dot{A}(t)A^{-1}(t) \qquad \text{(C-8)}$$

LIST OF REFERENCES

1. Hoerl, A. E., "Application of ridge analysis to regression problems," Chemical Engineering Progress, v. 58, p. 54-59, 1962.

2. Larson, H. J., Introduction to the Theory of Statistics, Ch. 7, Wiley, 1973.

3. Hoerl, A. E. and Kennard, R. W., "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, v. 12, no. 1, p. 55-67, February 1970.

4. Hoerl, A. E., "Ridge Analysis," Chemical Engineering Progress Symposium Series 60, p. 67-77, 1964.

5. Hoerl, A. E., "Optimum Solution of Many Variable Equations," Chemical Engineering Progress, v. 55, p. 69ff., 1959.

6. Hoerl, A. E and Kennard, R. W., "Ridge Regression: Applications to Nonorthogonal Problems," Technometrics, v. 12, no. 1, p. 69-83, February 1970.

7. Hemmerle, W. J., "An Explicit Solution for Generalized Ridge Regression," Technometrics, v. 17, no. 3, p. 309-314, August 1975.

8. Marquardt, D. W. and Snee, R. D., "Ridge Regression in Practice," The American Statistician, v. 29, no. 1, p. 3-20, February 1975.

9. Eskew, H. L., "Ridge Regression and Bayesian Inference in Costing," Proceedings, 31st Military Operations Research Symposium, 1974.

10. DeRusso, P. M., Roy, R. J., and Close, C. M., State Variables for Engineers, p. 201, 213, Wiley, 1965.

11. Marquardt, D. W., "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation," Technometrics, v. 12, no. 3, p. 591-612, August 1970.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3.	Department Chairman, Code 55 Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	1
4.	Professor Harold J. Larson, Code 55 La Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	1
5.	Professor Donald R. Barr, Code 55 Br Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	1
6.	LT Edgar B. Lewis, USN Naval Postgraduate School SMC 1277 Monterey, California 93940	1