The Annals of Statistics
1978, Vol. 6, No. 5, 1111-1121

GOOD AND OPTIMAL RIDGE ESTIMATORS

BY R. L. OBENCHAIN

Bell Telephone Laboratories

In generalized ridge estimation, the components of the ordinary least squares (OLS) regression coefficient vector which lie along the principal axes of the given regressor data are rescaled using known ridge factors. Generalizing a result of Swindel and Chapman, it is shown that, if each ridge factor is nonstochastic, nonnegative, and less than one, then there is at most one unknown direction in regression coefficient space along which ridge coefficients have larger mean squared error than do OLS coefficients. Then, by decomposing the mean squared error of a ridge estimator into components parallel to and orthogonal to the unknown true regression coefficient vector, new insight is gained about definitions for optimal factors. Estimators of certain unknown quantities are displayed which are maximum likelihood or unbiased under normal theory or which have correct range.

1. Introduction. Generalized ridge estimators for the unknown coefficient vector, $\boldsymbol{\beta}$, in a multiple linear regression model utilize ridge factors, $\delta_1, \cdots, \delta_p$, which rescale the components of the least squares vector, $\mathbf{b}^0$, along the principal axes of the given regressors. Specifically, with $\mathbf{c}$ denoting these components and with $\boldsymbol{\gamma}$ denoting the corresponding true components of $\boldsymbol{\beta}$, the estimator of γ_i is $\delta_i c_i$ for $i = 1, \cdots, p$. The mean squared error computations which will be made here assume that each δ_i is nonstochastic; the ridge estimator is then a linear estimator of $\boldsymbol{\beta}$.

Attention is focuced upon characterizations of two types of nonstochastic ridge factors which are unknown because they are functions of $(\boldsymbol{\beta}, \sigma^2)$. First, a set of nonstochastic ridge factors is said to be "good" for a fixed $(\boldsymbol{\beta}, \sigma^2)$ if the corresponding linear ridge estimator dominates least squares in every mean squared error sense. Secondly, a set of nonstochastic ridge factors is said to be "optimal" for a fixed $(\boldsymbol{\beta}, \sigma^2)$ if the corresponding linear ridge estimator achieves minimum possible mean squared error in one specific (univariate) sense.

Normal theory inferences are also considered in which stochastic ridge factors result from inserting estimates for $(\boldsymbol{\beta}, \sigma^2)$ into the formulas which characterize good or optimal nonstochastic factors. No attempt is made to estimate the exact mean squared error matrices of the resulting nonlinear ridge estimators.

Section 2 presents all necessary notation and definitions. The observations made in Section 2 illustrate the basic concepts which are used to establish the ridge function theorem of Section 3 and the optimal factor theorems of Section 4.

Received November, 1975; revised November 1976.
AMS 1970 *subject classifications*. Primary 62J05; Secondary 62F10.
Key words and phrases. Ridge regression, mean squared error optimality.

2. Notation, definitions, and known results. The notation of Obenchain (1975) will be utilized. Thus, given an $(n \times p)$ matrix of regressors $\mathbf{X}$ and an $(n \times 1)$ vector of the corresponding responses $\mathbf{y}$, assume that sample means have been removed from the data (so that $\mathbf{1}'\mathbf{X} = \mathbf{0}'$ and $\mathbf{1}'\mathbf{y} = 0$), and write the standard multiple linear regression model as $E(\mathbf{y}\,|\,\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ and $D(\mathbf{y}\,|\,\mathbf{X}) = \sigma^2(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)$, where $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of unknown regression coefficients ($\boldsymbol{\beta}'\boldsymbol{\beta} < \infty$) and $\sigma^2 > 0$ is the unknown error variance. The singular value decomposition of $\mathbf{X}$, Rao (1973), page 42, will be denoted by

$$\mathbf{X} = \mathbf{H}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{G}' \tag{2.1}$$

where $\boldsymbol{\Lambda}^{\frac{1}{2}}$ is the diagonal matrix of ordered singular values of $\mathbf{X}$, $\lambda_1^{\frac{1}{2}} \geqq \cdots \geqq \lambda_p^{\frac{1}{2}}$. Assume that $\lambda_p > 0$ so that $\boldsymbol{\beta}$ is estimable, and let $\mathbf{b}^0$ denote the ordinary least square estimator of $\boldsymbol{\beta}$; $\mathbf{b}^0 = \mathbf{X}^+\mathbf{y}$, where the superscript $+$ denotes the Moore-Penrose inverse. Then, as in Obenchain (1975), $\mathbf{b}^0 = \mathbf{G}\mathbf{c}$ where the vector $\mathbf{c} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{H}'\mathbf{y}$ contains the uncorrelated components of $\mathbf{b}^0$, and $E(\mathbf{c}\,|\,\mathbf{X}) = \mathbf{G}'\boldsymbol{\beta} \equiv \boldsymbol{\gamma}$, the unknown true components of $\boldsymbol{\beta}$.

The ith element of $\mathbf{c}$ has the form

$$c_i = (\mathbf{y}'\mathbf{y}/\lambda_i)^{\frac{1}{2}} r_{yi} \tag{2.2}$$

where r_{yi} is the correlation coefficient between $\mathbf{y}$ and the coordinates of the regressors along their ith principal axis, and the normal theory F-statistic for $\gamma_i = 0$ is

$$F_i = r_{yi}^2(n - p - 1)/(1 - R^2)\,, \tag{2.3}$$

where $R^2 = r_{y1}^2 + \cdots + r_{yp}^2$. In the conditional distribution theory of interest, $\mathbf{X}$ (and thus $\mathbf{H}$) is given, so the principal correlations, $r_{y1}, \cdots, r_{yp}$, are not then distributed as correlation coefficients. The unknown noncentrality of F_i is $\phi_i^2 = \gamma_i^2\lambda_i/\sigma^2$, but it is clear from (2.3) that F_i does not depend in any known way upon λ_i.

Generalized ridge estimators result from utilizing $\delta_i c_i$ as the estimator of γ_i for $i = 1, \cdots, p$, where δ_i is a realizable (fixed or stochastic) ridge factor. The estimator of $\boldsymbol{\beta}$ is thus

$$\mathbf{b}^* = \mathbf{G}\boldsymbol{\Delta}\mathbf{c} \tag{2.4}$$

where $\boldsymbol{\Delta}$ is the diagonal matrix with elements $\delta_1, \cdots, \delta_p$. If $\delta_1, \cdots, \delta_p$ are non-stochastic given $\mathbf{X}$, then the mean squared error matrix of $\mathbf{b}^*$ as an estimator of $\boldsymbol{\beta}$ is $\mathbf{G}[\mathrm{MSE}\,(\boldsymbol{\Delta}\mathbf{c})]\mathbf{G}'$ where

$$\mathrm{MSE}\,(\boldsymbol{\Delta}\mathbf{c}) = \sigma^2\boldsymbol{\Delta}^2\boldsymbol{\Lambda}^{-1} + (\mathbf{I} - \boldsymbol{\Delta})\boldsymbol{\gamma}\boldsymbol{\gamma}'(\mathbf{I} - \boldsymbol{\Delta}) \tag{2.5}$$

is the mean squared error matrix of $\boldsymbol{\Delta}\mathbf{c}$ as an estimator of $\boldsymbol{\gamma}$.

A weighted, univariate measure of the mean squared error of $\mathbf{b}^*$ as an estimator of $\boldsymbol{\beta}$ is given by

$$\mathrm{wmse}\,(\mathbf{b}^*, \mathbf{W}) = E[(\mathbf{b}^* - \beta)'\mathbf{W}(\mathbf{b}^* - \boldsymbol{\beta})]\,, \tag{2.6}$$

where $\mathbf{W}$ is any p by p, nonnegative definite, nonstochastic weight matrix. An often considered special case of (2.6) is that in which $\mathbf{W} = \mathbf{I}$, Hoerl and Kennard (1970); the resulting wmse $(\mathbf{b}^*, \mathbf{I})$ = trace [MSE $(\mathbf{b}^*)$] is called the summed mean squared error of $\mathbf{b}^*$. As another example, if $\boldsymbol{\alpha}$ is a unit vector ($\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$), then wmse $(\mathbf{b}^*, \boldsymbol{\alpha}\boldsymbol{\alpha}') = \boldsymbol{\alpha}'[\text{MSE}\,(\mathbf{b}^*)]\boldsymbol{\alpha} = \text{MSE}\,(\boldsymbol{\alpha}'\mathbf{b}^*) = \text{MSE}\,(-\boldsymbol{\alpha}'\mathbf{b}^*)$ is the mean squared error of $\mathbf{b}^*$ parallel to $\pm\boldsymbol{\alpha}$ in p-dimensional Euclidean space, i.e., the mean squared error of $\boldsymbol{\alpha}'\mathbf{b}^*$ as an estimator of $\boldsymbol{\alpha}'\boldsymbol{\beta}$.

A matrix measure of the mean squared error of $\mathbf{b}^*$ relative to $\mathbf{b}^0$ is given by EMSE $(\mathbf{b}^*)$ = MSE $(\mathbf{b}^0)$ − MSE $(\mathbf{b}^*)$, which will be called the Excess MSE matrix. Note that EMSE $(\mathbf{b}^*)$ is not MSE $(\mathbf{b}^0 - \mathbf{b}^*)$.

DEFFINITION. $\mathbf{b}^*$ of (2.4) is said to be a "good" generalized ridge estimator of $\boldsymbol{\beta}$ iff the following three equivalent properties hold:

$$\text{EMSE}\,(\mathbf{b}^*) = \text{MSE}\,(\mathbf{b}^0) - \text{MSE}\,(\mathbf{b}^*) \tag{2.7.i}$$

is a positive definite matrix, or

$$\text{MSE}\,(\boldsymbol{\alpha}'\mathbf{b}^0) > \text{MSE}\,(\boldsymbol{\alpha}'\mathbf{b}^*) \tag{2.7.ii}$$

for every unit vector $\boldsymbol{\alpha}$, or

$$\text{wmse}\,(\mathbf{b}^\circ, \mathbf{W}) > \text{wmse}\,(\mathbf{b}^*, \mathbf{W}) \tag{2.7.iii}$$

for every positive definite weight matrix $\mathbf{W}$. (Theobald (1974), Theorem 1, proved the equivalence of (2.7.i) and (2.7.iii).)

Necessary and sufficient conditions are given in Section 3 for a $\mathbf{b}^*$ with nonstochastic ridge factors on the range $0 \leqq \delta_i < 1$ for $i = 1, \cdots, p$ to be good in the above sense for fixed $(\boldsymbol{\beta}, \sigma^2)$. These conditions are generalizations of the results of Swindel and Chapman (1973), who considered only the one "parameter" family of factors $\delta_i = \lambda_i/(\lambda_i + k)$ for $i = 1, \cdots, p$ and $k > 0$. Furthermore, if a $\mathbf{b}^*$with nonstochastic $0 \leqq \delta_i < 1$ is not good, it is shown that there exists an unknown direction $\boldsymbol{\alpha}^*$ such that MSE $(\boldsymbol{\alpha}'\mathbf{b}^0) >$ MSE $(\boldsymbol{\alpha}'\mathbf{b}^*)$ for every $\boldsymbol{\alpha}$ orthogonal to $\boldsymbol{\alpha}^*$ but such that MSE $(\boldsymbol{\alpha}'\mathbf{b}^0)$ can be smaller than MSE $(\boldsymbol{\alpha}'\mathbf{b}^*)$ when $\boldsymbol{\alpha}'\boldsymbol{\alpha}^* \neq 0$.

No realizable (stochastic or nonstochastic) $\delta_1, \cdots, \delta_p$ are good for every $\boldsymbol{\beta}$ and σ^2; this point is clarified by Theorem 1 for the case of nonstochastic (δ)'s, and the case of stochastic (δ)'s is treated by Brown (1975) and Bunke (1975a, b). For example, the Stein-type $\mathbf{b}^*$ estimators which dominate $\mathbf{b}^0$ in wmse $(\mathbf{b}, \mathbf{X}'\mathbf{X})$ for every $(\boldsymbol{\beta}, \sigma^2)$ when $p \geqq 3$, Efron and Morris (1976), equation (4.1), are not good since, as suggested by Stein (1962), page 267, MSE $(\boldsymbol{\alpha}'\mathbf{b}^0) <$ MSE $(\boldsymbol{\alpha}'\mathbf{b}^*)$ is possible when $\boldsymbol{\alpha}$ is parallel to the unknown vector $(\mathbf{X}'\mathbf{X})^{\frac{1}{2}}\boldsymbol{\beta} = \mathbf{G}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\gamma}$. As a second example, the choice $\delta_i = \gamma_i/c_i$, considered by Vinod (1976) and Kennard (1976) when $c_i \neq 0$, yields zero error but is clearly not realizable.

Linear $(\mathbf{b}^*)$'s with $0 \leqq \delta_i < 1$ are admissible estimators of $\boldsymbol{\beta}$ because they are Bayes with respect to some proper prior on $\boldsymbol{\beta}$ (e.g., Lindley and Smith (1972)); $\mathbf{b}^0$ is admissible in the matrix mean squared error sense even though it is Bayes with respect to an improper prior on $\boldsymbol{\beta}$. However, Thisted (1976) shows that

heuristic rules for choosing among linear ridge estimators for a particular problem can yield inadmissible (nonlinear) estimators.

Section 4 considers "optimal," unknown nonstochastic choices of $\delta_1, \cdots, \delta_p$ which mimimize MSE $(\boldsymbol{\alpha}'\mathbf{b}^*)$ for given $\boldsymbol{\alpha}$ or which minimize wmse $(\mathbf{b}^*, \mathbf{W})$ for given $\mathbf{W}$. For example, the unknown δ_i which minimizes the mean squared error of $\delta_i c_i$ as a linear estimator of γ_i, Hoerl and Kennard (1970), is

$$\delta_i^{\text{MSE}} = \phi_i^2/(1 + \phi_i^2)\,, \tag{2.8}$$

where $\phi_i^2 = \gamma_i^2\lambda_i/\sigma^2$ is again the unknown noncentrality of F_i, (2.3). Then, as observed in Obenchain (1975), $\delta_i = \delta_i^{\text{MSE}}$ for $i = 1, \cdots, p$ is the unknown, optimal set of ridge factors with respect to any criterion which depends only upon the diagonal elements of MSE $(\boldsymbol{\Delta}\mathbf{c})$. Two such examples are the wmse $(\mathbf{b}^*, \mathbf{I})$ of Hoerl and Kennard (1970) and the scaled, predictive mean squared error $= 1 + \sigma^{-2}$ wmse $(\mathbf{b}^*, \mathbf{X}'\mathbf{X})$ of Mallows (1973). Restricted and unrestricted maximum likelihood estimation of $\delta_1^{\text{MSE}}, \cdots, \delta_p^{\text{MSE}}$ is considered in Obenchain (1975). The most interesting new results of Section 4 correspond to minimizing MSE$(\boldsymbol{\alpha}'\mathbf{b}^*)$ by choice of $\delta_1, \cdots, \delta_p$ when $\boldsymbol{\alpha}$ is either parallel to or orthogonal to the unknown true $\boldsymbol{\beta}$.

The generalized ridge estimator for shrinkage toward an arbitrary known point $\boldsymbol{\beta}_0$ is

$$\mathbf{b}^* = \mathbf{G}\boldsymbol{\Delta}(\mathbf{c} - \boldsymbol{\gamma}_0) + \boldsymbol{\beta}_0\,, \tag{2.4$'$}$$

where $\boldsymbol{\gamma}_0 = \mathbf{G}'\boldsymbol{\beta}_0$ is also known. Then MSE $(\mathbf{b}^*) = \mathbf{G}[\text{MSE}(\boldsymbol{\Delta}\mathbf{c} + (\mathbf{I} - \boldsymbol{\Delta})\boldsymbol{\gamma}_0)]\mathbf{G}'$, so it is only necessary to replace $\boldsymbol{\gamma}$ by $(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)$ in (2.5) and in the following to consider this generalization. Another generalization occurs when the original linear model was $E(\mathbf{z}\,|\,\mathbf{W}) = \mu\mathbf{1} + \mathbf{W}\boldsymbol{\beta}$ and $D(\mathbf{z}\,|\,\mathbf{W}) = \sigma^2\mathbf{V}$ for a known positive definite covariance structure $\mathbf{V}$. Writing $\boldsymbol{\omega} = \mathbf{Q}\mathbf{V}^{-\frac{1}{2}}\mathbf{1}$, $\mathbf{y} = (\mathbf{I} - \boldsymbol{\omega}\boldsymbol{\omega}^+)\mathbf{Q}\mathbf{V}^{-\frac{1}{2}}\mathbf{z}$, and $\mathbf{X} = (\mathbf{I} - \boldsymbol{\omega}\boldsymbol{\omega}^+)\mathbf{Q}\mathbf{V}^{-\frac{1}{2}}\mathbf{W}$ for any n by n orthogonal matrix $\mathbf{Q}$, a corresponding linear model is $E(\mathbf{y}\,|\,\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ and $D(\mathbf{y}\,|\,\mathbf{X}) = \sigma^2(\mathbf{I} - \boldsymbol{\omega}\boldsymbol{\omega}^+)$ where $\boldsymbol{\omega}'\mathbf{y} = 0$ and $\boldsymbol{\omega}'\mathbf{X} = \mathbf{0}'$. The principal components analysis terminology of Obenchain (1975) is not strictly appropriate unless $\boldsymbol{\omega}$ is proportional to $\mathbf{1}$, but this can be made to be the case by choice of $\mathbf{Q}$. (For example, take $\mathbf{Q} = \mathbf{Q}_1\mathbf{Q}_2'$ where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are any orthogonal matrices whose first columns are proportional to $\mathbf{1}$ and $\mathbf{V}^{-\frac{1}{2}}\mathbf{1}$, respectively, so that $\boldsymbol{\omega} = \mathbf{1}(\mathbf{1}'\mathbf{V}^{-1}\mathbf{1}/n)^{\frac{1}{2}}$.) Whatever be $\boldsymbol{\omega}$, the normal distribution theory and (2.3) are unchanged.

3. Good ridge factors. Let $\zeta_1 \geqq \cdots \geqq \zeta_p$ denote the ordered eigenvalues of EMSE $(\mathbf{b}^*)$, (2.7.i). Then $\mathbf{b}^*$ is good if and only if $\zeta_p > 0$. Define a scalar valued function, $RF(\boldsymbol{\Delta})$, of $\delta_1, \cdots, \delta_p$ to be

$$\begin{aligned} RF(\boldsymbol{\Delta}) &= \textstyle\sum_{i=1}^p [\delta_i^{\text{MSE}}(1 - \delta_i)]/[(1 - \delta_i^{\text{MSE}})(1 + \delta_i)] \\ &= \textstyle\sum_{i=1}^p \phi_i^2(1 - \delta_i)/(1 + \delta_i)\,, \end{aligned} \tag{3.1}$$

where δ_i^{MSE} is the unknown shrinkage factor of (2.8). The following theorem

gives necessary and sufficient conditions for certain (**b***)'s to be good and displays the eigenvector $\boldsymbol{\alpha}^*$ of EMSE (**b***) corresponding to ζ_p when $\zeta_p < 0$.

THEOREM 1 (ridge function theorem). *If* $(\boldsymbol{\beta}, \sigma^2)$ *are fixed parameters such that* $\sigma^2 > 0$ *and* $\boldsymbol{\beta}'\boldsymbol{\beta} < \infty$, *the given* $\mathbf{X}$ *is of rank* $p(\lambda_p > 0)$, *and* δ_i *is nonstochastic on the range* $0 \leqq \delta_i < 1$ *for* $i = 1, \cdots, p$, *then*

(i) $\zeta_{p-1} > 0$ *if* $p > 1$,
(ii) $\zeta_p > 0$ iff $\delta_1, \cdots, \delta_p$ *are all sufficiently close to one that* $RF(\boldsymbol{\Delta}) < 1$,
(iii) $\zeta_p = 0$ iff $RF(\boldsymbol{\Delta}) = 1$,
(iv) $\zeta_p < 0$ iff $RF(\boldsymbol{\Delta}) > 1$, *and*
(v) *the eigenvector of* EMSE (**b***) *corresponding to a* $\zeta_p < 0$ *has* jth *element*

$$\alpha_j^* \propto \sum_{i=1}^{p} \frac{g_{ji}(1 - \delta_i)\gamma_i}{[\sigma^2\lambda_i^{-1}(1 - \delta_i^2) + |\zeta_p|]}, \tag{3.2}$$

for $\mathbf{G} = ((g_{ji}))$ *and* λ_i *of* (2.1) *and* $\boldsymbol{\gamma} = \mathbf{G}'\boldsymbol{\beta}$.

PROOF. It will only be necessary to show that EMSE (**b***) can be written in the notation of the lemma of the Appendix under the assumptions of the theorem. EMSE (**b***) $= \mathbf{GAG}'$ where $\mathbf{A} = \boldsymbol{\Lambda}^{-1}\sigma^2 - \text{MSE}(\boldsymbol{\Delta}\mathbf{c})$ can be rewritten as $\mathbf{A} = \mathbf{D}(\mathbf{I} - \mathbf{zz}')\mathbf{D}$ where $\mathbf{D}$ is a positive definite diagonal matrix of the form $\mathbf{D} = (\mathbf{I} - \boldsymbol{\Delta}^2)^{\frac{1}{2}}\boldsymbol{\Lambda}^{-\frac{1}{2}}\sigma$ and $\mathbf{z} = \sigma^{-1}(\mathbf{I} - \boldsymbol{\Delta}^2)^{-\frac{1}{2}}\boldsymbol{\Lambda}^{\frac{1}{2}}(\mathbf{I} - \boldsymbol{\Delta})\boldsymbol{\gamma}$ is a column vector. Note then that

$$\mathbf{z}'\mathbf{z} = \sum_{i=1}^{p} \frac{\phi_i^2(1 - \delta_i)^2}{(1 - \delta_i^2)} = RF(\boldsymbol{\Delta}),$$

that EMSE (**b***) and $\mathbf{A}$ have the same eigenvalues, and that the eigenvector of EMSE(**b***) corresponding to a $\zeta_p < 0$ is proportional to $\mathbf{G}[\mathbf{D}^2 + |\zeta_p|\mathbf{I}]^{-1}\mathbf{Dz}$.

REMARKS. (a) In the one-parameter Hoerl–Kennard (1970) family, $\delta_i = \lambda_i/(\lambda_i + k)$, the ridge function is $RF(\boldsymbol{\Delta}) = \sum \phi_i^2/(1 + 2\lambda_i k^{-1})$ and the results of Swindel and Chapman (1973) are a special case of part (ii) of the above theorem. Namely, every $k > 0$ yields a good ridge estimator if $\sum \phi_i^2 < 1$; otherwise the good range is $0 < k < 2/|\eta_p|$ where η_p is the negative eigenvalue of $(\mathbf{X}'\mathbf{X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}'/\sigma^2$. The sufficient condition of Theobald (1974), Theorem 2, that $0 < k < 2\sigma^2/\boldsymbol{\beta}'\boldsymbol{\beta}$ is thus usually too stringent to be necessary.

(b) When $p = 1$, $\delta_1 c_1$ is good for $\delta_1 > \delta_1^{\text{MIN}} = \max(0; 2\delta_1^{\text{MSE}} - 1)$, and $RF(\boldsymbol{\Delta}^{\text{MSE}}) < 1$ when $p = 2$. $RF(\boldsymbol{\Delta}^{\text{MSE}})$ can exceed one when $p > 2$. It is easily shown that the following "(2/p)th's rule," $\delta_i \geqq 1 - 2(1 - \delta_i^{\text{MSE}})/p$ for $i = 1, \cdots, p$, is sufficient for **b*** to be good because each term of (3.1) will then be less than $1/p$.

(c) Since $f(\delta) = (1 - \delta)/(1 + \delta)$ is convex on $0 \leqq \delta < 1$, the good ridge factors form a convex, open, nonempty set within $0 \leqq \delta_i < 1$ for $i = 1, \cdots, p$ for each fixed $(\boldsymbol{\beta}, \sigma^2)$ with $\boldsymbol{\beta}'\boldsymbol{\beta} < \infty$ and $\sigma^2 > 0$.

(d) If $RF_1(\boldsymbol{\Delta})$ is the value of the ridge function at $(\boldsymbol{\beta}_1, \sigma_1^2)$ and f is any nonzero constant factor, then $RF_2(\boldsymbol{\Delta}) = f^2 RF_1(\boldsymbol{\Delta})$ is its value at $(\boldsymbol{\beta}_2, \sigma_2^2)$ of the form

$(f\boldsymbol{\beta}_1, \sigma_1^2)$ or $(\boldsymbol{\beta}_1, \sigma_1^2/f^2)$. Since $RF_1(\boldsymbol{\Delta})$ must be greater than zero for some $(\boldsymbol{\beta}_1, \sigma_1^2)$ when δ_i is nonstochastic and $0 \leqq \delta_i < 1$ for $i = 1, \cdots, p$, no such $\boldsymbol{\Delta}$ can be good for all $(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ with $\boldsymbol{\beta}'\boldsymbol{\beta} < \infty$ and $\sigma^2 > 0$.

Normal theory estimation. The maximum likelihood estimator of $\phi_i^2 = \gamma_i^2\lambda_i/\sigma^2$ under normal distribution theory is $nF_i/(n - p - 1)$ for F_i of (2.3). The maximum likelihood estimators of $\zeta_1, \cdots, \zeta_p$ are the p solutions of

$$\sum_{i=1}^{p} c_i^2(1 - \delta_i)^2/[s^2\lambda_i^{-1}(1 - \delta_i^2) - \zeta] = 1 \tag{3.3}$$

if $c_i = (\mathbf{y}'\mathbf{y}/\lambda_i)^{\frac{1}{2}}r_{yi}$ is not zero for $i = 1, \cdots, p$, and $s^2 = (\mathbf{y}'\mathbf{y})(1 - R^2)/n$. If the estimate of ζ_p is negative, substitution into (3.2) yields the corresponding maximum likelihood estimator of $\boldsymbol{\alpha}^*$. For numerical examples, see Obenchain (1976a, b).

The estimator of ϕ_i^2 which is unbiased under normal distribution theory is $\hat{\phi}_i^2 = -1 + (n - p - 3)F_i/(n - p - 1)$, and the corresponding "correct range" estimator is $\phi_i^{*2} = \max(0, \hat{\phi}_i^2)$. All of the above estimators can be used to estimate $RF(\boldsymbol{\Delta}) = \sum \phi_i^2(1 - \delta_i)/(1 + \delta_i)$, and it is noted that the maximum likelihood estimate $\geqq$ the correct range estimate $\geqq$ the unbiased estimate.

The scaled mean squared error matrix, MSE $(\boldsymbol{\Delta}\mathbf{c})/\sigma^2$, is unbiasedly estimated for nonstochastic $\boldsymbol{\Delta}$ by

$$\begin{aligned}\hat{\mathbf{T}} &= \frac{(n - p - 3)}{(n - p - 1)}(\mathbf{I} - \boldsymbol{\Delta})\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{t}\mathbf{t}'\boldsymbol{\Lambda}^{-\frac{1}{2}}(\mathbf{I} - \boldsymbol{\Delta}) + \boldsymbol{\Lambda}^{-1}(2\boldsymbol{\Delta} - \mathbf{I}) \\ &= ((\hat{\tau}_{ij}))\,,\end{aligned} \tag{3.4}$$

where $\mathbf{t}$ is the column vector of t-statistics, $t_i = \pm F_i^{\frac{1}{2}}$ of (2.3). The ith diagonal element of MSE $(\boldsymbol{\Delta}\mathbf{c})/\sigma^2$ is $\tau_{ii} = \text{MSE}(\delta_i c_i)/\sigma^2 = [(1 - \delta_i)^2\phi_i^2 + \delta_i^2]/\lambda_i \geqq \delta_i^2/\lambda_i$, this lower limit being the known scaled variance of $\delta_i c_i$. A correct range estimator of τ_{ii} is provided by $\tau_{ii}^* = \max(\hat{\tau}_{ii}, \delta_i^2/\lambda_i)$. The estimators of Mallows' C-statistic, $C(\mathbf{b}^*) = 1 + \sum \lambda_i\tau_{ii}$, corresponding to $\hat{\tau}_{ii}$ and τ_{ii}^*, Obenchain (1975), equation (5.1′), are minimized term-by-term when δ_i equals

$$\delta_i^{M\neq} = \max\{0, 1 - (n - p - 1)/[F_i(n - p - 3)]\}\,. \tag{3.5}$$

The corresponding estimators of MSE $(\boldsymbol{\alpha}'\mathbf{b}^*)/\sigma^2$, which is the scaled mean squared parallel to $\pm\boldsymbol{\alpha}$, are $\boldsymbol{\alpha}'\mathbf{G}\hat{\mathbf{T}}\mathbf{G}'\boldsymbol{\alpha}$ for $\mathbf{G}$ of (2.1), where the off-diagonal elements of $\hat{\mathbf{T}}$ of (3.4) are used and the diagonal elements are either the $\hat{\tau}_{ii}$ or the τ_{ii}^*. If the $\hat{\tau}_{ii}$ are used, a correct range estimator of MSE $(\boldsymbol{\alpha}'\mathbf{b}^*)/\sigma^2$ is $\max(\boldsymbol{\alpha}'\mathbf{G}\hat{\mathbf{T}}\mathbf{G}'\boldsymbol{\alpha}, \boldsymbol{\alpha}'\mathbf{G}\boldsymbol{\Delta}^2\boldsymbol{\Lambda}^{-1}\mathbf{G}'\boldsymbol{\alpha})$, and this estimator can be smaller than the correct range estimator which utilizes the τ_{ii}^*. Special cases of interest in the above occur when $\boldsymbol{\alpha}$ is a column of the identity matrix or when $\boldsymbol{\alpha}$ indicates a contrast between a pair of coefficients which are known to be highly negatively correlated in the least squares estimator. Just as the ridge coefficients are plotted to form a "trace," Hoerl and Kennard (1970), the above estimates of scaled mean squared error or of the eigenvalues of EMSE $(\mathbf{b}^*)$ can be plotted to aid in solution selection, Obenchain (1976a, b).

Under the restriction $\mathbf{\Delta} = \delta\mathbf{I}$, $\mathbf{b}^* = \delta\mathbf{b}^0$ is good for $\delta > 1 - 2/(1 + \sum \phi_i^2)$, where $\sum \phi_i^2$ is the noncentrality of the F-ratio, $F_0 = (F_1 + \cdots + F_p)/p = R^2(n - p - 1)/[p(1 - R^2)]$, for the hypothesis that $\boldsymbol{\beta} = \mathbf{0}$. Utilizing the maximum likelihood estimator, $npF_0/(n - p - 1)$, of $\sum \phi_i^2$, the resulting estimated lower limit for good δ exceeds $1 - 2(n - p - 1)/(npF_0)$. Now $\delta\mathbf{b}^0$ is known to dominate $\mathbf{b}^0$ in summed mean squared error under normal theory, Baranchik (1970, 1973), for $\delta = \max(0, 1 - c/F_0)$ when $p > 2$, $n > p + 1$, and the constant c is $0 < c < 2(p - 2)(n - p - 1)/[p(n - p + 1)]$. Note that this known upper limit on c is bigger than $2(n - p - 1)/np$ by a factor of $(p - 2)n/(n - p + 1)$. Thus sample evidence usually indicates that some Stein-type estimators may not be good.

Normal theory confidence statements are relatively simple to make when $\mathbf{\Delta} = \delta\mathbf{I}$ using arguments similar to those of Johnson and Welch (1939) for producing confidence sets for a coefficient of variation. Given a confidence probability π, $\Pr[L < \sum \phi_i^2] = \pi$ when L is the minimum noncentrality needed to make the 100π percent point of noncentral $F(p, n - p - 1)$ equal or exceed the observed F_0 for $\boldsymbol{\beta} = \mathbf{0}$. The resulting interval with (δ)'s which are good with specified confidence is $\Pr[RF(\delta\mathbf{I}) < 1 \text{ for } M < \delta < 1] = 1 - \pi$ where $M = \max[0, (L - 1)/(L + 1)]$. For example, $F_0 = 10$, $p = 4$, $n - p - 1 = 25$, $\pi = 0.1$, and standard approximations, Johnson and Kotz (1970), yield $L = 60.3$ and $M = 0.967$, while (δ)'s as small as $1 - 25/270 = 0.908$ imply that $\delta\mathbf{b}^0$ dominates $\mathbf{b}^0$ in the sense of Baranchik (1970, 1973).

The "strong" MSE criterion of Toro-Vizcarrondo and Wallace (1968) is related to the ridge function for the case where each δ_i is zero or one, so that $RF(\mathbf{\Delta}) = \sum_{i=1}^{p}(1 - \delta_i)\phi_i^2$. (In their notation, this noncentrality parameter is divided by two.) The "weak" MSE criterion of Wallace (1972) corresponds to observing that $\text{wmse}(\mathbf{b}^0, \mathbf{X}'\mathbf{X}) - \text{wmse}(\mathbf{b}^*, \mathbf{X}'\mathbf{X}) = \sigma^2 \sum_{i=1}^{p}(1 - \delta_i)(1 - \phi_i^2)$, again when each δ_i is zero or one in $\mathbf{b}^*$. There criteria suggest forming the F-ratio $\sum_{i=1}^{p}(1 - \delta_i)F_i/m$ with degrees-of-freedom $m = p - \delta_1 - \cdots - \delta_p$ and $(n - p - 1)$ then testing whether its unknown noncentrality is either $\leqq 1$ for the "strong" criterion or $\leqq m$ for the "weak" criterion. However, generalizations of this procedure to cases where the δ_i are neither zeros and ones nor all equal do not seem to be straightforward.

4. Optimal ridge factor theorems. The following theorems characterize the ridge factors which achieve either minimum mean squared error parallel to an arbitrary direction, $\boldsymbol{\alpha}$, in coefficient space or minimum wmse $(\mathbf{b}^*, \mathbf{W})$ for arbitrary positive definite weight matrix, $\mathbf{W}$. The directions parallel to and orthogonal to the unknown $\boldsymbol{\beta}$ will be shown to play pivotal roles in this theory.

THEOREM 2. *Given a unit vector* $\boldsymbol{\alpha}$, $\sigma^2 > 0$, *and* $\boldsymbol{\beta}'\boldsymbol{\beta} < \infty$,

(i) MSE $(\boldsymbol{\alpha}'\mathbf{b}^*)$ *does not depend upon* δ_i *when* $\boldsymbol{\alpha}'\mathbf{g}_i = 0$, *and*

(ii) MSE $(\boldsymbol{\alpha}'\mathbf{b}^*)$ *is minimized by choice of nonstochastic* δ_i *at*

$$\delta_i(\boldsymbol{\alpha}) = (\boldsymbol{\alpha}'\boldsymbol{\beta})\lambda_i\gamma_i/[(\boldsymbol{\alpha}'\mathbf{g}_i)(\sigma^2 + \sum{}^{*}\gamma_j^2\lambda_j)] \tag{4.1}$$

when $\boldsymbol{\alpha}'\mathbf{g}_i \neq 0$, *where* $\sum^*$ *denotes summation only over subscripts* j *such that* $\boldsymbol{\alpha}'\mathbf{g}_j \neq 0$. *Equivalently,* $\delta_i\gamma_i = k\phi_i^2/(\boldsymbol{\alpha}'\mathbf{g}_i)$ *where* $k = (\boldsymbol{\alpha}'\boldsymbol{\beta})/(1 + \sum^* \phi_j^2)$.

PROOF. Let $\boldsymbol{\xi}' = \boldsymbol{\alpha}'\mathbf{G}$ be the unit vector giving the coordinates of $\boldsymbol{\alpha}$ with respect to the principal axes of $\mathbf{X}$. Then $\boldsymbol{\alpha}'\mathbf{b}^* = \boldsymbol{\xi}'\boldsymbol{\Delta}\mathbf{c}$ is the component of $\mathbf{b}^*$ in the $\boldsymbol{\alpha}$ direction, and this does not depend upon δ_i when $\boldsymbol{\alpha}$ is orthogonal to the ith principal axis ($\xi_i = \boldsymbol{\alpha}'\mathbf{g}_i = 0$). Nonstochastic (δ)'s imply MSE $(\boldsymbol{\alpha}'\mathbf{b}^*) = \sigma^2\boldsymbol{\xi}'\boldsymbol{\Delta}^2\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi} + (\boldsymbol{\xi}'(\mathbf{I}-\boldsymbol{\Delta})\boldsymbol{\gamma})^2$ and ∂ MSE $(\boldsymbol{\alpha}'\mathbf{b}^*)/\partial\delta_i = 2\delta_i\xi_i^2\lambda_i^{-1}\sigma^2 - 2(\boldsymbol{\xi}'(\mathbf{I}-\boldsymbol{\Delta})\boldsymbol{\gamma})\xi_i\gamma_i$, which is identically zero when $\xi_i = 0$ or becomes zero when $\delta_i = k^{(\alpha)}\gamma_i\lambda_i/\xi_i$ and $k^{(\alpha)} = [\boldsymbol{\xi}'(\mathbf{I} - \boldsymbol{\Delta})\boldsymbol{\gamma}]/\sigma^2$ for $\xi_i \neq 0$. The minimization problem is convex because the second derivative's matrix is nonnegative definite; so (4.1) results by solving $\sigma^2k^{(\alpha)} = \boldsymbol{\xi}'\boldsymbol{\gamma} - k^{(\alpha)}\sum^* \gamma_j^2\lambda_j$ for $k^{(\alpha)}$ and noting that $\boldsymbol{\xi}'\boldsymbol{\gamma} = \boldsymbol{\alpha}'\boldsymbol{\beta}$.

REMARKS. (a) It is easily shown that $\delta_i(\boldsymbol{\alpha}) = \delta_i(-\boldsymbol{\alpha})$, but neither $0 \leqq \delta_i(\boldsymbol{\alpha})$ nor $\delta_i(\boldsymbol{\alpha}) \leqq 1$ is necessarily the case.

(b) $\delta_i(\mathbf{g}_i) = \gamma_i^2\lambda_i/(\sigma^2 + \gamma_i^2\lambda_i) = \delta_i^{\text{MSE}}$ of (2.8) and $\delta_j(\mathbf{g}_i)$ is undetermined for $j \neq i$.

(c) If $\boldsymbol{\beta}$ is zero, every $\boldsymbol{\alpha}$ is orthogonal to $\boldsymbol{\beta}$ in the sense that $\boldsymbol{\alpha}'\boldsymbol{\beta} = 0$, and $\delta_i(\boldsymbol{\alpha}) = 0$ is an optimal choice for $i = 1, \cdots, p$ and for every $\boldsymbol{\alpha}$ in this case. When $\boldsymbol{\beta} \neq \mathbf{0}$, there is a $(p - 1)$-dimensional space of $\boldsymbol{\alpha}$'s orthogonal to $\boldsymbol{\beta}$, and $\delta_i(\boldsymbol{\alpha}) = 0$ when $\boldsymbol{\alpha}'\mathbf{g}_i \neq 0$ is again the optimal choice for these orthogonal $\boldsymbol{\alpha}$'s. If one always shrinks $\mathbf{b}^0$ to 0 by taking $\delta_1 = \cdots = \delta_p = 0$, one cannot make an error orthogonal to the unknown, true $\boldsymbol{\beta}$.

(d) If $\boldsymbol{\beta} \neq \mathbf{0}$, $\boldsymbol{\alpha}$ parallel to $\boldsymbol{\beta}$ yields $\xi_i = \boldsymbol{\alpha}'\mathbf{g}_i = \gamma_i(\boldsymbol{\gamma}'\boldsymbol{\gamma})^{-\frac{1}{2}}$. Thus $\delta_i(\boldsymbol{\alpha})$ is not determined for this $\boldsymbol{\alpha}$ when $\gamma_i = 0$, and $\delta_i = k^{(=)}\lambda_i$ is otherwise optimal for

$$k^{(=)} = (\boldsymbol{\gamma}'\boldsymbol{\gamma})/(\sigma^2 + \boldsymbol{\gamma}'\boldsymbol{\Lambda}\boldsymbol{\gamma}) \geqq 0\,. \tag{4.2}$$

The corresponding ridge estimator is $\mathbf{b}^* = k^{(=)}\mathbf{G}\boldsymbol{\Lambda}\mathbf{c} = k^{(=)}\mathbf{X}'\mathbf{y}$, where $\mathbf{X}'\mathbf{y}$ is the vector of p inner products of the given regressor values with the response vector. Thus a ridge estimator which achieves minimum mean squared error parallel to the unknown, true $\boldsymbol{\beta}$ is known to be parallel to $\mathbf{X}'\mathbf{y}$, and this is the only case (except $\boldsymbol{\Delta} = \mathbf{0}$) when the relative magnitudes of $\delta_1(\boldsymbol{\alpha}), \cdots, \delta_p(\boldsymbol{\alpha})$ are known in (4.1).

THEOREM 3. *Given a* p.d. *weight matrix* $\mathbf{W}$, $\sigma^2 > 0$, *and* $\boldsymbol{\beta}'\boldsymbol{\beta} < \infty$, wmse $(\mathbf{b}^*, \mathbf{W})$ *of* (2.6) *is minimized by choice of nonstochastic* $\delta_1, \cdots, \delta_p$ *at*

$$\delta_i(\mathbf{W}) = \gamma_i\lambda_i\eta_i/(\sigma^2 m_{ii})\,, \tag{4.3}$$

where η_i *is the* ith *element of* $\boldsymbol{\eta} = (\mathbf{D} + \mathbf{M}^{-1})^{-1}\boldsymbol{\gamma}$, $\mathbf{M} = ((m_{ij})) = \mathbf{G}'\mathbf{W}\mathbf{G}$, *and* $\mathbf{D}$ *is the diagonal matrix with* ith *element* ϕ_i^2/m_{ii}. *Equivalently,* $\boldsymbol{\Delta}\boldsymbol{\gamma} = \mathbf{D}\boldsymbol{\eta} = \mathbf{D}(\mathbf{D} + \mathbf{M}^{-1})^{-1}\boldsymbol{\gamma}$.

PROOF. wmse $(\mathbf{b}^*, \mathbf{W}) = \sigma^2$ trace $(\mathbf{M}\boldsymbol{\Delta}^2\boldsymbol{\Lambda}^{-1}) + \boldsymbol{\gamma}'(\mathbf{I} - \boldsymbol{\Delta})\mathbf{M}(\mathbf{I} - \boldsymbol{\Delta})\boldsymbol{\gamma}$ is a convex function of $\delta_1, \cdots, \delta_p$, so the minimum occurs at ∂ wmse$/\partial\boldsymbol{\delta} = \mathbf{0}$, which are fixed point equations of the form (4.3) with $\boldsymbol{\eta} = \mathbf{M}(\mathbf{I} - \boldsymbol{\Delta})\boldsymbol{\gamma}$. Thus $\boldsymbol{\eta} = \mathbf{M}\boldsymbol{\gamma} - \mathbf{M}\mathbf{D}\boldsymbol{\eta}$ yields the desired result when solved for $\boldsymbol{\eta}$ because $(\mathbf{M}\mathbf{D} + \mathbf{I})^{-1} = (\mathbf{D} + \mathbf{M}^{-1})^{-1}\mathbf{M}^{-1}$ where $\mathbf{M}$ is positive definite.

REMARKS. (a) If $\mathbf{M}$ is a diagonal matrix or if at most one element of $\boldsymbol{\gamma}$ is nonzero, then $\delta_i(\mathbf{W}) = \delta_i^{\text{MSE}}$ of (2.8).

(b) The one "parameter" family of weight matrices $\mathbf{W} = \mathbf{I} + (\nu - 1)\boldsymbol{\beta}\boldsymbol{\beta}^+$ and corresponding optimal factors of (4.3) have some interesting properties. The "parameter" ν is the Lagrange multiplier when the problem is to minimize the summed mean squared error orthogonal to $\boldsymbol{\beta}$ under a restriction on the mean squared error parallel to $\boldsymbol{\beta}$. Then $\delta_i(\mathbf{W})$ approaches 0 as ν approaches 0, $\delta_i(\mathbf{W}) = \delta_i^{\text{MSE}}$ at $\nu = 1$, and $\delta_i(\mathbf{W})$ approaches $k^{(=)}\lambda_i$ of (4.2) as ν approaches infinity for $i = 1, \cdots, p$.

Normal theory estimation. The maximum likelihood estimator of $k^{(=)}$ of (4.2) under normal distribution theory is $\hat{k}^{(=)} = \sum nr_{yi}^2\lambda_i^{-1}/[1 + (n - 1)R^2]$. Replacing n by $(n - p - 1)$ in this formula would remove the bias in the maximum likelihood estimate of σ^2 but would not produce an unbiased estimate of $k^{(=)}$. Note that $\hat{k}^{(=)}$ is rather sensitive to the small eigenvalues, $\lambda_p, \lambda_{p-1}, \cdots$, which occur when $\mathbf{X}$ is ill-conditioned. Estimates of (4.1) and (4.3) can be constructed by using c_i as the maximum likelihood, unbiased estimate of γ_i and using the maximum likelihood or correct range or unbiased estimates of ϕ_i^2, as described in Section 3.

5. Conclusions. Certain unrealizable estimators have been characterized as being either good relative to least squares or optimal for given direction $\boldsymbol{\alpha}$ or weight matrix $\mathbf{W}$. These two classes appear to contain all known definitions for "optimal," nonstochastic factors. It is not claimed that the sample statistics displayed here allow one to infer that any realizable estimators are guaranteed to be good or optimal for a particular problem. The high variability associated with an ill-conditioned regression problem affects the estimators of (3.1) to (3.4) and of (4.1) to (4.3). However, the above procedures do provide valuable sample evidence to the ridge practitioner; the data are used to indicate which ridge estimators are likely to be good or in what direction they are worse than least squares and also to estimate the scaled mean square error of linear ridge estimators in all directions of p space.

APPENDIX

LEMMA. *If* $\mathbf{D} = \text{Diag}(d_1, \cdots, d_p)$ *is a positive definite diagonal matrix and if* $\mathbf{z} = (z_1, \cdots, z_p)'$ *is a vector, then the matrix* $\mathbf{A} = \mathbf{D}(\mathbf{I} - \mathbf{z}\mathbf{z}')\mathbf{D}$ *is*

(i) *positive definite* iff $\mathbf{z}'\mathbf{z} < 1$,
(ii) *nonnegative definite of rank* $(p - 1)$ iff $\mathbf{z}'\mathbf{z} = 1$, *and*
(iii) *has one negative eigenvalue and* $(p - 1)$ *positive eigenvalues* iff $\mathbf{z}'\mathbf{z} > 1$.

An eigenvector $\boldsymbol{\tau}$ *corresponding to an eigenvalue* λ *of* $\mathbf{A}$ *is*

(a) $\boldsymbol{\tau} = i$th *column of the identity matrix and* $\lambda = d_i^2$ *if* $z_i = 0$ *for some* i.
(b) $\boldsymbol{\tau} \propto (\mathbf{D}^2 - \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{z}$ *if* $\lambda \neq d_i^2$ *for any* $i = 1, \cdots, p$.

PROOF. Since $\boldsymbol{\tau} \neq \mathbf{0}$ is to be an eigenvector of $\mathbf{A}$ with eigenvalue λ, $\mathbf{A}\boldsymbol{\tau} = \lambda\boldsymbol{\tau} = \mathbf{D}^2\boldsymbol{\tau} - (\mathbf{z}'\mathbf{D}\boldsymbol{\tau})\mathbf{D}\mathbf{z}$. Thus $\boldsymbol{\tau}$ must be $k(\mathbf{D}^2 - \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{z}$ for some scalar $k \neq \mathbf{0}$ if $\lambda \neq d_i^2 > 0$ for $i = 1, \cdots, p$ and if $\mathbf{z} \neq \mathbf{0}$. But then $(\mathbf{D}^2 - \lambda\mathbf{I})\boldsymbol{\tau} = k\mathbf{D}\mathbf{z} = (\mathbf{z}'\mathbf{D}\boldsymbol{\tau})\mathbf{D}\mathbf{z} = (k\mathbf{z}'\mathbf{D}^2(\mathbf{D}^2 - \lambda\mathbf{I})^{-1}\mathbf{z})\mathbf{D}\mathbf{z}$. It follows that every eigenvalue of $\mathbf{A}$ must either be a solution of

$$g(\lambda) = \sum_{i=1}^{i=p} z_i^2 d_i^2 (d_i^2 - \lambda)^{-1} = 1$$

or must equal one of the $d_i^2 > 0$ for $i = 1, \cdots, p$. Letting m denote the smallest d_i^2 for which the corresponding $z_i^2 > 0$, it is clear that $g(\lambda)$ is strictly increasing on $-\infty < \lambda < m$ and that $g(\lambda) = 1$ has exactly one solution in this interval. If this solution is $\lambda_{\text{MIN}} \leqq 0$, it cannot be a multiple eigenvalue of $\mathbf{A}$ because the corresponding eigenspace of $(\mathbf{D}^2 + |\lambda_{\text{MIN}}|\mathbf{I})^{-1}\mathbf{D}\mathbf{z}$ has rank one. The value, $g(0) = \mathbf{z}'\mathbf{z}$, of $g(\lambda)$ at $\lambda = 0$ is thus critical in determining λ_{MIN}. Specifically, $g(0) < 1$ implies that the solution of $g(\lambda) = 1$ is $\lambda_{\text{MIN}} > 0$, $g(0) = 1$ implies $\lambda_{\text{MIN}} = 0$, and $g(0) > 1$ implies $\lambda_{\text{MIN}} < 0$.

Acknowledgment. The author wishes to thank the referees and also Ronald Thisted for comments which helped to clarify the presentation.

REFERENCES

[1] BARANCHIK, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Math. Statist.* **41** 642–645.

[2] BARANCHIK, A. J. (1973). Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables. *Ann. Statist.* **1** 312–321.

[3] BHATTACHARYA, P. K. (1966). Estimating the mean of a multivariate normal population with general quadratic loss function. *Ann. Math Statist.* **37** 1819–1824.

[4] BROWN, L. D. (1975). Estimation with incompletely specified loss functions (the case of several location parameters). *J. Amer. Statist. Assoc.* **70** 417–427.

[5] BUNKE, O. (1975a). Least squares estimators as robust and minimax estimators. *Math. Operationsforch. u. Statist.* **6** 687–688.

[6] BUNKE, O. (1975). Improved inference in linear models with additional information. *Math. Operationsforsch. u. Statist.* **6** 817–829.

[7] EFRON, B. and MORRIS, C. (1976). Families of minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.* **4** 11–21.

[8] HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

[9] JOHNSON, N. L. and KOTZ, S. (1970). *Continuous Univariate Distributions*—2. (Chapter 30, Noncentral *F*-Distribution.) Houghton Mifflin, Boston.

[10] JOHNSON, N. L. and WELCH, B. L. (1939). Applications of the noncentral *t*-distribution. *Biometrika* **31** 362–381.

[11] KENNARD, R. W. (1976). Letter to the Editor. *Technometrics* **18** 504–505.

[12] LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B* **34** 1–41.

[13] MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

[14] OBENCHAIN, R. L. (1975). Ridge analysis following a preliminary test of the shrunken hypothesis. *Technometrics* **17** 431–441.

[15] OBENCHAIN, R. L. (1976a). Methods of ridge regression. *Proceedings of the Ninth International Biometric Conference*, Volume One, 37–57, Boston.

[16] OBENCHAIN, R. L. (1976b). Example output of RELAXR, a FORTRAN program for generalized ridge regression. Unpublished.
[17] RAO, C. R. (1973). *Linear Statistical Inference and Its Application*, 2nd ed. Wiley, New York.
[18] STEIN, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 197–206. Univ. of California Press.
[19] STEIN, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution. *J. Roy. Statist. Soc. Ser. B* **24** 265–296.
[20] SWINDEL, B. F. and CHAPMAN, D. D. (1973). Good ridge estimators. Abstracts Booklet, 1973 Joint Statistical Meetings, Page 126. New York City.
[21] THEOBALD, C. M. (1974). Generalizations of mean squared error applied to ridge regression. *J. Roy. Statist. Soc. Ser. B* **36** 103–106.
[22] THISTED, R. (1976). Ridge regression, minimax estimation, and empirical bayes methods. Technical Report No. 28, Division of Biostatistics, Stanford Univ.
[23] TORO-VIZCARRONDO, C. and WALLACE, T. D. (1968). A test of the mean squared error criterion in linear regression. *J. Amer. Statist. Assoc.* **63** 558–572.
[24] VINOD, H. D. (1976). Letter to the Editor. *Technometrics* **18** 504.
[25] WALLACE, T. D. (1972). Weaker criteria and tests for linear restrictions in regression. *Econometrica* **40** 689–698.

Addendum. Farebrother (1976) treated the case where the rank of $\mathbf{X}$ is $r < p$ and good estimators are sought in the one-parameter ridge family. In this case $\lambda_{r+1} = \cdots = \lambda_p = 0$, and the last $(p - r)$ principal axes and components are not uniquely determined. Farebrother's main result, given in his equation (6), implies that a fixed k yields an estimator which dominates least squares for every estimable linear function iff the sum of the first r-terms of the ridge function of (3.1) is less than one—but the last $(p - r)$ terms are zero because $\phi_{r+1}^2 = \cdots = \phi_p{}^2 = 0$. Thus it is not necessary to assume that $r = p$ to prove a generalization of Theorem 1, and the Farebrother (1976) result is more general than that of Swindel and Chapman (1973). Farebrother (1978) gives a further generalization of part (ii) of Theorem 1 to a class of estimators which includes (2.4) and (2.4′) as special cases.

FAREBROTHER, R. W. (1976). Further results on the mean squared error of ridge regression. *J. Roy. Statist. Soc. B* **38** 248–250.
FAREBROTHER, R. W. (1978). A class of shrinkage estimators. *J. Roy. Statist. Soc. B* **40** (to appear).

BELL LABORATORIES
HOLMDEL, NEW JERSEY 07733