MATRIX CONDITIONING AND MINIMAX ESTIMATION[1]

George Casella
Biometrics Unit, Cornell University, Ithaca, N. Y.

BU-732-M[2] March 1981

Abstract

Most of the research concerning ridge regression methods has dealt with improving the mean square error of the regression coefficient estimates. However, ridge regression was originally formulated with two goals in mind; the other being the numerical stabilization of the coefficient estimates. We show that, if the eigenvalues of the design matrix satisfy certain conditions, then a minimax ridge estimator (an estimator whose risk uniformly dominates that of the least squares estimator) can also be more numerically stable than the least squares estimator.

---

[1] This research was supported by NSF Grant No. MCS79-05771.

[2] In the Biometrics Unit Mimeo Series, Cornell University, Ithaca, NY 14853.

MATRIX CONDITIONING AND MINIMAX ESTIMATION

George Casella
Biometrics Unit, Cornell University, Ithaca, N. Y.

1. Introduction

The goal of ridge regression, as described by Hoerl and Kennard (1970), was to improve upon the least squares technique in ill-conditioned problems. The basic idea was that, by allow controlled amounts of bias in the original design matrix, it was possible to produce a more stable estimate of the coefficient vector.

The term 'stable estimate' is not, as yet, a well defined term. In general, a stable estimate can be considered to be one which is insensitive to small fluctuations in the data. Of course, the estimate must be sensitive to some degree but, as Hoerl and Kennard point out, in many cases the least squares estimate can change drastically when the data are perturbed by only a small amount.

Even though the original goal of ridge regression was to improve stability, almost all research concerning ridge regression estimators has dealt with another criterion: mean square error. Almost totally ignoring the question of stability, researchers have used both Monte Carlo methods (see, e.g., McDonald and Galarneau (1975)) and analytic methods (see, e.g., Thisted (1976), Strawderman (1978), Casella (1980)) to find classes of ridge regression estimators whose risk dominates that of the least squares estimator for all values of β, i.e., minimax ridge estimators.

Key words and phrases: Minimax, ridge regression, condition number, mean square error.

While the question of minimaxity is important, the goal of stability should not be disregarded. In this paper we investigate a class of ridge estimators to see when minimaxity and stability can be simultaneously realized. We find that, under certain conditions on the structure of the eigenvalues of the design matrix, both minimaxity and stability can be simultaneously achieved. We also give conditions under which no minimax estimator (in the class considered) can be guaranteed to be more stable than the least squares estimator.

2. Notation

We begin with the linear regression model

$$Y = X\beta + \epsilon \tag{2.1}$$

where Y is an $n \times 1$ vector of observations, X is the known $n \times p$ design matrix of rank p, β is the $p \times 1$ vector of unknown regression coefficients, and ϵ is an $n \times 1$ vector of experimental errors. We assume ϵ has a multivariate normal distribution with mean zero and covariance matrix $\sigma^2 I$, i.e., $\epsilon \sim N(0, \sigma^2 I)$. The least squares estimator is

$$\hat{\beta} = (X'X)^{-1}X'Y \quad . \tag{2.2}$$

Let P be the matrix of normalized eigenvectors of $X'X$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ be its eigenvalues. Then

$$\begin{aligned} P'X'XP &= D_\lambda = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p) \ , \\ P'P &= I \ , \\ \hat{\beta} &= PD_\lambda^{-1}P'X'Y \quad . \end{aligned} \tag{2.3}$$

The instability of $\hat{\beta}$ is a function of the eigenvalues of $X'X$. From (2.3) it can be seen that a small element of D_λ could make $\hat{\beta}$ very sensitive to small changes in the Y vector. (More precisely, it is not really λ_p being

close to zero that increases the sensitivity of $\hat{\beta}$, but rather that the ratio λ_1/λ_p is large. If the X'X matrix is put into correlation form, so that $\Sigma\lambda_i = p$, then λ_1/λ_p "large" and λ_p "small" are equivalent.)

A generalized ridge regression estimator is defined

$$\hat{\beta}(K) = P(D_\lambda + K)^{-1}P'X'Y \quad , \tag{2.4}$$

where $K = \text{diag}(k_1, k_2, \cdots, k_p)$, $k_i \geq 0$. If the k_i's are chosen in an appropriate manner, the eigenvalues of the matrix $D_\lambda + K$ will be drawn closer together, which would result in a more numerically stable estimator.

The loss of an estimator B will be measured by

$$L(B,\beta) = (1/\sigma^2)(B - \beta)'(B - \beta) \quad , \tag{2.5}$$

and its risk (mean square error), by

$$R(B,\beta) = EL(B,\beta) \quad , \tag{2.6}$$

where the expectation is over the distribution of B, so $R(B,\beta)$ is a function of β . B is a minimax estimator of β if and only if

$$R(B,\beta) \leq R(\hat{\beta},\beta) \qquad \text{for all } \beta \quad , \tag{2.7}$$

so the risk of a minimax estimator is uniformly smaller than that of the least squares estimator.

The results of this paper can be easily generalized to include losses of the form

$$L(B,\beta) = (1/\sigma^2)(B - \beta)'Q(B - \beta) \quad , \tag{2.8}$$

where Q is a known positive definite matrix (see Casella (1980) for details). We will confine our attention to the loss (2.5), however. This is done not only for ease of notation, but also because $Q = I$ seems to be the most widely used loss function.

3. Stochastic vs. Nonstochastic Ridge Estimators

A generalized ridge regression estimator is called stochastic if the k_i's are a function of the data (the Y vector), otherwise it is nonstochastic. Most ridge regression literature has dealt with nonstochastic estimators in theory, but with stochastic estimators in practice. Even choosing k based on a ridge trace must be viewed as a stochastic choice, since the ridge trace depends on the data. A major drawback of stochastic estimators chosen by ad hoc schemes (such as a ridge trace) is that there is no way to analytically determine their properties. (There are, of course, Monte Carlo studies, but no Monte Carlo study can cover all cases.)

Any theory developed for nonstochastic ridge estimators, unfortunately, has no bearing on stochastic ridge estimators. For example, formulas for variance and mean square error, easily derived for nonstochastic estimators, are not applicable to stochastic estimators. If $\hat{\beta}(K)$ is a nonstochastic estimator, it is easy to show that

$$\lim_{\beta'\beta \to \infty} R(\hat{\beta}(K),\beta) = \infty \quad , \tag{3.1}$$

while there exist stochastic $\hat{\beta}(K)$ whose risk uniformly (in β) dominates that of $\hat{\beta}$.

In view of the poor risk properties of nonstochastic estimators, it can be argued that stochastic estimators should be used exclusively. Since that seems to be the case anyway, it can further be argued that only stochastic estimators with biasing factors chosen by some fixed rule be used. It is only for these estimators that any optimality properties can be established analytically.

Table 1 lists eight (stochastic) methods for choosing ridge biasing factors which have appeared in the literature. Only three of these methods (2, 4 and 7) result in estimators that can be evaluated analytically.

Table 1. Some Methods for Choosing Ridge Biasing Factors

Authors	Method	Type
1. Hoerl and Kennard (1970)	Ridge Trace	Stochastic
2. Hoerl, Kennard and Baldwin (1975)	$k = p\hat{\sigma}^2/\hat{\beta}'\hat{\beta}$	Stochastic
3. McDonald and Galarneau (1975)	Choose k such that $\hat{\beta}'(k)\hat{\beta}(k) = \hat{\beta}'\hat{\beta} - \sigma^2 tr(X'X)^{-1}$	Stochastic
4. Lawless and Wang (1976)	$k = p\hat{\sigma}^2/\hat{\beta}'X'X\hat{\beta}$	Stochastic
5. Vinod (1976)	Several Criteria	Both Stochastic and Nonstochastic
6. Obenchain (1977)	Several Criteria	Both Stochastic and Nonstochastic
7. Hemmerle and Brantle (1978)	Minimize Estimated Risk	Stochastic
8. Golub, Heath and Wahba (1979)	Generalized Cross Validation	Stochastic

The results of Brown (1971) and Berger (1976) imply that k_i of the form

$$k_i = a_i \hat{\sigma}^2 / \hat{\beta}' T \hat{\beta} \quad , \tag{3.2}$$

where a_i is a positive constant and T is a positive definite matrix, will result in estimators with good risk properties. (Their results are more general than stated here.) The strategy in this article is to consider ridge estimators belonging to a subclass of those defined by (3.2), (a subclass known to contain minimax estimators) and examine the relationship between the minimax conditions and the numerical stability of the estimators. The class of estimators considered here is defined by

$$k_i = a_i \hat{\sigma}^2 / \hat{\beta}' X' X \hat{\beta} \quad , \tag{3.3}$$

where the shift from T to X'X is, unfortunately, due only to analytic considerations. The minimax condition for k_i of the form (3.3) is the weakest known. While more general choices T can also yield minimax estimators, the conditions derived thus far are stronger than necessary.

For any vectors $a = (a_1\, a_2, \cdots, a_p)$ and $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_p)$, define

$$M(a,\lambda) = \frac{\sum_{i=1}^{p} (a_i/\lambda_i^2) - 2 \max_i (a_i/\lambda_i^2)}{\max_i (a_i/\lambda_i^2)} \tag{3.4}$$

The following theorem is a special case of Theorem 4.1 of Casella (1980).

Theorem 1: If k_i is of the form (3.3), and $m\hat{\sigma}^2/\sigma^2 \sim \chi_m^2$ independent of $\hat{\beta}$, then $\hat{\beta}(K)$ is minimax against the loss (2.5) if

$$\max(a_i/\lambda_i) \leq \frac{2m}{m+2} M(a,\lambda) \quad . \tag{3.5}$$

If σ^2 is known, then the condition

$$\max_i(a_i/\lambda_i) \leq 2M(a,\lambda)$$

is necessary and sufficient for $\hat{\beta}(K)$ to be a minimax estimator of β . In this light it can be seen that condition (3.5) is a minimal requirement for minimaxity.

In order for $\hat{\beta}(K)$ to be different from $\hat{\beta}$, $M(a,\lambda)$ must be positive. This places a restriction on the way the a_i's can be chosen, a restriction dependent on the eigenvalues of X'X . We now examine this restriction more closely, to better understand how the a_i's should be chosen to yield a minimax estimator.

If we set $a_i = a\lambda_i$, then $\hat{\beta}(K)$ can be written

$$\hat{\beta}(K) = \left(\frac{\hat{\beta}'X'X\hat{\beta}}{a\hat{\sigma}^2+\hat{\beta}'X'X\hat{\beta}}\right)\hat{\beta} \quad , \tag{3.6}$$

which is a spherically symmetric estimator. Bock (1975) showed that if

$$\sum_{i=1}^{p} \lambda_i^{-1} \leq 2\lambda_p^{-1} \quad , \tag{3.7}$$

then no spherically symmetric minimax estimator (other than $\hat{\beta}$) exists. We can see that if $a_i = a\lambda_i$,

$$M(a,\lambda) = \left(\sum_{i=1}^{p} \lambda_i^{-1} - 2\lambda_p^{-1}\right)/\lambda_p^{-1} \quad , \tag{3.8}$$

and, hence, choosing k_i proportional to λ_i cannot produce a minimax ridge estimator if (3.7) holds. Also, if $a_i = a$ (ordinary ridge regression), then

$$M(a,\lambda) = \left(\sum_{i=1}^{p} \lambda_i^{-2} - 2\lambda_p^{-2}\right)/\lambda_p^{-2} \quad . \tag{3.9}$$

Thus, if the λ_i's make either (3.8) or (3.9) negative, in order for the estimator to be minimax, the a_i's must be chosen, in general, to weight the larger eigenvalues more heavily than the smaller eigenvalues. Notice also that if (3.7) holds, then at least one λ_i is much smaller than the others, a sign of multicollinearity and ill-conditioning. Thus, in general, to produce a minimax version of $\hat{\beta}(K)$ in ill-conditioned problems, the coordinates corresponding to the larger eigenvalues of X'X must be shrunk more than those corresponding to the smaller eigenvalues.

In another sense, if $\sum_{i=1}^{p} \lambda_i^{-1} < 2\lambda_p^{-1}$, one must truly <u>force</u> an estimator to be minimax. If $\lambda_i = 1$ for all i, (3.7) reduces to the statement $p \leq 2$. If the dimension is less than or equal to 2, $\hat{\beta}$ is the unique minimax estimator. Therefore if (3.7) holds, we might say that the "essential dimensionality" of the problem is less than 2, and perhaps minimax ridge estimators should not be sought. If minimaxity is forced, it may be required to shrink in a counterintuitive way, shrinking the coordinates with small variances much more than those with large variances.

4. <u>Condition Numbers</u>

The condition number of a matrix A is a measure (used mainly in numerical analysis) of the accuracy attainable in the solution of the linear system $Az = b$. Or, in other words, the condition number measures the sensitivity of the solution to perturbations in the data. The following definition can be found in Stewart (1973).

<u>Definition</u>: The <u>condition number</u>, $\kappa(A)$, of a matrix A with respect to the Euclidean norm $\|\cdot\|$ is

$$\kappa(A) = \|A\| \, \|A^{-1}\| = (\lambda_{max}(A'A)/\lambda_{min}(A'A))^{\frac{1}{2}} \quad , \tag{4.1}$$

where $\lambda_{max}(\cdot)$ and $\lambda_{min}(\cdot)$ denote, respectively, the largest and smallest eigenvalues.

We will be concerned only with estimators of the form $B = H^{-1}X'Y$ and, for convenience, will refer to the condition number of the estimator B (by which we will mean the condition number of the matrix H). A particularly useful interpretation of condition numbers is also given in Stewart's book, and we paraphrase it here. If $\hat{\beta}_1$ and $\hat{\beta}_2$ are given by

$$\hat{\beta}_1 = (X'X)^{-1}X'Y_1, \qquad \hat{\beta}_2 = (X'X)^{-1}X'Y_2 \quad ,$$

then

$$\|\hat{\beta}_1 - \hat{\beta}_2\|^2/\|\hat{\beta}_1\|^2 \leq \kappa(X'X)\|PY_1 - PY_2\|^2/\|PY_1\|^2 \quad ,$$

where PY_i is the projection of Y_i onto the space spanned by the columns of X, and $\kappa(X'X) = \lambda_1/\lambda_p$. Thus, the condition number reflects the sensitivity of an estimator to perturbations in the data.

From these definitions we see that

$$\begin{aligned} \kappa(\hat{\beta}) &= \lambda_1/\lambda_p \\ \kappa(\hat{\beta}(K)) &= \max_{1\leq i\leq p} \lambda_i + k_i \Big/ \min_{1\leq i\leq p} \lambda_i + k_i \quad . \end{aligned} \tag{4.2}$$

Thus, the k_i's can be chosen so that $\kappa(\hat{\beta}(K)) < \kappa(\hat{\beta})$. We also note that the choice $k_i \equiv k$ (ordinary ridge regression) always results in a smaller condition number. In general, without further assumptions on the ordering of the k_i's, we cannot explicitly compute $\kappa(\hat{\beta}(K))$.

If $\hat{\beta}$ itself is already an oversensitive estimator, it seems unwise to replace it with another estimator even more sensitive. Thus, a reasonable criteria for any ridge estimator to satisfy is

$$\kappa(\hat{\beta}(K)) \leq \lambda_1/\lambda_p \quad . \tag{4.3}$$

For estimators of the form (3.3), the following theorem gives conditions that guarantee (4.3).

Theorem 2: Let $t = \hat{\beta}'X'X\hat{\beta}/\hat{\sigma}^2$. For $\hat{\beta}(K)$ of the form (3.3),

$$\kappa(\hat{\beta}(K)) \leq \kappa(\hat{\beta}) = \lambda_1/\lambda_p \qquad \text{for all } t > 0$$

if and only if $a_{max}/a_{min} < \lambda_1/\lambda_p$.

Proof: $\kappa(\hat{\beta}(K)) \leq \lambda_1/\lambda_p$ for all $t > 0$ if and only if

$$\frac{\lambda_i+(a_i/t)}{\lambda_j+(a_j/t)} \leq \lambda_1/\lambda_p \qquad i,j = 1, \cdots, p, \quad t>0 \ .$$

This is equivalent to

$$\frac{\lambda_i}{\lambda_j}\left(\frac{t+(a_i/\lambda_i)}{t+(a_j/\lambda_j)}\right) \leq \lambda_1/\lambda_p \qquad i,j = 1, \cdots, p, \quad t>0 \ . \tag{4.4}$$

If $a_i/\lambda_i > a_j/\lambda_j$ then the left-hand side of (4.4) is decreasing in t, with maximum a_i/a_j . If $a_i/\lambda_i \leq a_j/\lambda_j$, then it is nondecreasing in t, with maximum λ_i/λ_j . Thus

$$\frac{\lambda_i+(a_i/t)}{\lambda_j+(a_j/t)} \leq \max\left\{\frac{a_i}{a_j}, \frac{\lambda_1}{\lambda_p}\right\} \leq \frac{\lambda_1}{\lambda_p} \ , \tag{4.5}$$

and the "if" part of the theorem is proved. Since the maximum can be attained, the "only if" part is also proved.

Notice that the condition $a_{max}/a_{min} \leq \lambda_1/\lambda_p$ precludes setting any $a_i = 0$, for then

$$\kappa(\hat{\beta}(K)) \geq (\lambda_1 + a_i/t)/\lambda_i \to \infty \text{ as } t \to 0 \ . \tag{4.6}$$

This is a drawback of these estimators, since it might be desirable not to shrink some coordinates, especially those with small variances. However, one can try to shrink these coordinates as little as possible. We also point out, however, that if $a_{max}/a_{min} \leq \lambda_1/\lambda_p$, then $\kappa(\hat{\beta}(K))$ achieves its maximum at either $t = 0$ or $t = \infty$. Thus, in practical applications, the condition number of $\hat{\beta}(K)$ will be strictly less than that of $\hat{\beta}$.

5. Stability and Minimaxity

5.1 Favorable Eigenstructures

In this section we present a result which characterizes the relationship between minimaxity and improvement of conditioning. The discussion at the end of Section 3 showed, heuristically, that if the eigenvalues of X'X are spread out, minimaxity must be forced into the estimator. Moreover, the shrinkage must be performed in a counterintuitive way. We now quantify the eigenstructures that will allow both minimaxity and improved condition number.

Recall the definition of $M(a,\lambda)$ from (3.4). We have the following theorem.

Theorem 3: Let $a = (a_1, \cdots, a_p)$, $\lambda = (\lambda_1, \cdots, \lambda_p)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. If $a_{max}/a_{min} \leq \lambda_1/\lambda_p$ then

$$\underset{a}{\text{maximum}}\, M(a,\lambda) = (\lambda_p/\lambda_1) + (p-3) - \lambda_1\lambda_p \sum_{i=2}^{p-1} \left((\lambda_i^2/\lambda_1\lambda_p) - 1\right)^{+}\lambda_i^{-2} \quad , \qquad (5.1)$$

where "+" denotes the positive part.

Proof: From the definition of $M(a,\lambda)$, it can be seen that $M(a,\lambda)$ is scale invariant. (That is, for any scalars c and d, $M(ca,d\lambda) = M(a,\lambda)$.) Therefore, without loss of generality, set $a_{max} = 1$. It can be seen that M is maximized when as many as possible of the terms $a_i\lambda_i^{-2}$ are set equal to the maximum term. Indeed, the unrestricted maximum of $M(a,\lambda)$ is $p-2$, achieved when $a_i\lambda_i^{-2} = a_j\lambda_j^{-2}$ for all i,j .

Under the restriction $a_{max}/a_{min} \leq \lambda_1/\lambda_p$, with $a_{max} = 1$, we must have $a_{min} \geq \lambda_p/\lambda_1$. Now

$$a_{min}\lambda_p^{-2} \geq (\lambda_1\lambda_p)^{-1} \geq \lambda_1^{-2} = a_{max}\lambda_1^{-2} \quad .$$

Now it follows that $M(a,\lambda)$ will be maximized when $a_i\lambda_i^{-2}$ is as close as possible to $(\lambda_1\lambda_p)^{-1}$ for $i = 2, \cdots, p-1$. This is achieved by setting $a_i = \min(1, \lambda_i^2/\lambda_1\lambda_p) = (\lambda_i^2/\lambda_1\lambda_p) - ((\lambda_i^2/\lambda_1\lambda_p) - 1)^+$. (Recall that $a_i \le a_{max} = 1$.) Substituting these values into $M(a,\lambda)$ proves the theorem.

If $\lambda_i^2 \le \lambda_1\lambda_p$ for $i = 2, \cdots, p-1$ then Theorem 3 shows that, under the restriction $a_{max}/a_{min} < \lambda_1/\lambda_p$,

$$\text{maximum } M(a,\lambda) = (\lambda_p/\lambda_1) + (p-3) \quad , \tag{5.2}$$

the largest value that the restricted maximum can attain. Thus, eigenvalues which satisfy $\lambda_i^2 < \lambda_1\lambda_p$, $i = 2, \cdots, p-1$ are most favorable, and allow minimax estimators to improve conditioning. Eigenvalues which satisfy this condition will not be very spread out, a situation which is known to be favorable to minimax estimators.

If, on the other hand, $\lambda_i^2 > \lambda_1\lambda_p$ for $i = 2, \cdots, p-1$, then, under the restriction $a_{max}/a_{min} < \lambda_1/\lambda_p$,

$$\text{maximum } M(a,\lambda) = \lambda_1\lambda_p\Big(\sum_{i=1}^{p-1} \lambda_i^{-2} - (\lambda_1\lambda_p)^{-1}\Big) \quad . \tag{5.3}$$

Therefore, the condition $\sum_{i=1}^{p-1} \lambda_i^{-2} > (\lambda_1\lambda_p)^{-1}$ is necessary and sufficient for the existence of a minimax estimator (of the form (3.3)) which will improve conditioning. We summarize this in the following theorem.

Theorem 4: There exists a minimax estimator of the form (3.3) with the property that

$$\max_{i,j} \frac{\lambda_i + a_i/t}{\lambda_j + a_j/t} \le \lambda_1/\lambda_p \qquad \text{for all } t > 0$$

if and only if $\sum_{i=1}^{p-1} \lambda_i^{-2} > (\lambda_1\lambda_p)^{-1}$.

5.2 Choosing the Shrinkage Constants

The method of choosing the constants a_i that is outlined in Theorem 3 is not the method that can provide the maximum amount of improvement in the condition number. However, if $M(a,\lambda)$ is positive for this choice, then there is room for a more substantial improvement in conditioning without forfeiting minimaxity. If we again set $a_{max} = a_1 = 1$, then the greatest improvement in conditioning will be realized when $a_{min} = a_p$ is as large as possible. A development similar to that of Theorem 3 will show that for $a_{max} = 1$ and $a_{min} = a$, $M(a,\lambda)$ is maximized by choosing $a_i = \min(1, a\lambda_i^2/\lambda_p^2)$. For this choice of the a_i's we have

$$M(a,\lambda) = \frac{1}{a}\left(\frac{\lambda_p}{\lambda_1}\right)^2 + (p-3) - \left(\frac{\lambda_p^2}{a}\right) \sum_{i=2}^{p-1} \left(a\lambda_i^2/\lambda_p^2) - 1\right)^{+} \lambda_i^{-2} \quad . \tag{5.4}$$

Thus, the greatest improvement in conditioning (while still preserving minimaxity) is realized by choosing a to be the largest value for which (5.4) is positive. We can find this value of a by noticing that for $(\lambda_p/\lambda_j)^2 < a < (\lambda_p/\lambda_j + 1)^2$,

$$M(a,\lambda) = \frac{1}{a}\left(\frac{\lambda_p}{\lambda_1}\right)^2 + (p - j - 2) \quad , \tag{5.5}$$

which is decreasing in a and, moreover, has no sign changes over the specified interval. Therefore, we need merely locate the smallest value of j that makes (5.5) negative, and set $a = (\lambda_p/\lambda_{j-1})^2 = a_p$, $a_i = 1$ for $i = 1, \cdots, p-1$.

This choice of a_i's allows a_p to be as large as possible (relative to a_1), while still preserving minimaxity. Thus, this choice can produce the greatest improvement in conditioning of any minimax estimator considered here. It is important to note, however, that since the condition number of these estimators is a function of the data, the choice of a_i's that can potentially

produce the smallest condition number may not produce the smallest observed condition number. This is illustrated in Section 6.

There is one more point to be considered in choosing the constants a_i. Although $M(a,\lambda)$ is scale free, both the minimax condition (3.5) and the risk of the estimator are not scale free. Once the structure of the a_i's has been chosen, they then should be scaled in some optimal way. Although no absolute best scale factor is known, the following argument provides some guidelines.

From the proof of Theorem 1 (Casella (1980)), it can be shown that the maximum increase in risk of the estimator (3.3) over the risk of $\hat{\beta}$ is proportional to

$$\max(a_i/\lambda_i^2)[\max(a_i/\lambda_i) - 2(m/(m+2))M(a,\lambda)] \quad . \tag{5.6}$$

Now let each a_i be multiplied by the scale factor a. Then (5.6) becomes

$$a\ \max(a_i/\lambda_i^2)[a\ \max(a_i/\lambda_i) - 2(m/(m+2))M(a,\lambda)] \quad . \tag{5.7}$$

Our object is to minimize (5.7), thus producing the estimator with the smallest upper bound on the increase in risk. The minimum value of (5.7) is attained when

$$a = \frac{mM(a,\lambda)}{(m+2)\max(a_i/\lambda_i)} \quad . \tag{5.8}$$

Substituting (5.8) back into (5.7) shows that the maximum increase in risk is proportional to

$$-\max(a_i/\lambda_i^2)(m/(m+2))^2M^2(a,\lambda) \quad . \tag{5.9}$$

Thus, choosing the scale in this way seems, in the absence of any other information, to be an optimal choice.

5.3 Other Considerations

For any given set of data, a ridge estimator in the class considered can be forced to be minimax, or forced to improve conditioning. The case in which both goals can be accomplished has been characterized in this section. However, a word should be said about the case in which both goals cannot be achieved simultaneously, and how to proceed in such a case.

If $\sum_{i=1}^{p} \lambda_i^{-2} < (\lambda_1 \lambda_p)^{-1}$, then any minimax estimator in the class considered cannot guarantee improved conditioning, and any estimator which guarantees improved conditioning will not be minimax. The experimenter must then decide which goal is more important, and use the estimator that best suits his needs. As will be seen in the example, the condition $\sum_{i=1}^{p} \lambda_i^{-2} < (\lambda_1 \lambda_p)^{-1}$, while being a sign of ill-conditioning, does not signify gross ill-conditioning. Thus, a minimax estimator might still be appropriate, if the original condition number is acceptable. If it is the case that the original condition number is unacceptable, the wisest course would seem to be the forfeiture of minimaxity to achieve numerical stability.

6. An Example

It is a difficult task to characterize the types of eigenstructures that will admit a minimax estimator which also improves conditioning. It is clear that the situation is more favorable when $\lambda_{max}/\lambda_{min}$ is small, but the relative positions of the other eigenvalues are also important. As an example, consider the Acetylene Data, analyzed by Marquardt and Snee (1975). The data consisted of 16 observations on one response and three predictor variables. Before any computations were done, the means were removed from the variables (a procedure which improves conditioning) and all variables were standardized. A full nine-term quadratic model was fitted to the data, and the eigenvalues are given in Table 2. It can be seen that the problem is

Table 2. Minimax Biasing Factors for the Acetylene Data

Eigenvalues	I. a_i's that maximize $M(a,\lambda)$	II. a_i's that provide greatest improvement in conditioning while preserving minimaxity
4.2	1	1
2.16	1	1
1.14	1	1
1.04	1	1
.38	1	1
.05	1	1
.014	$(0.14)^2/(.0001 \times 4.2) = .467$	1
.005	$(.005)^2/(.0001 \times 4.2) = .060$	1
.0001	$(.0001)/4.2 = .000024$	$(.0001/.005)^2 = .0004$

highly ill-conditioned, with $\lambda_1/\lambda_p = 42,000$. Also, $\Sigma\lambda_i^{-1} - 2\lambda_i^{-1} = -9,703$ and $\Sigma\lambda_i^{-2} - 2\lambda_p^{-2} = -54,489$ so neither a spherically symmetric estimator, nor an estimator which shrinks each coordinate equally, can be minimax. However, Theorem 3 shows that $\max M(a,\lambda) = 1.17$, so that there exists a minimax estimator which will improve conditioning. The choice of a_i's that obtain $M(a,\lambda) = 1.17$ is given in Table 2 as Choice I. In accordance with the discussion in Section 5.2, we scale each a_i by multiplying it by $(m/(m+2))M(a,\lambda)/\max(a_i/\lambda_i) = (6/8)(1.17)/(33.4) = .026$. For these a_i's the condition number of the estimator is given by

$$\max_{i,j} \frac{\lambda_i + a_i/t}{\lambda_j + a_i/t} = \frac{\lambda_1}{\lambda_p}\left(\frac{t + .026/\lambda_1}{t + (.026)(.00024)/\lambda_p}\right) ,$$

where $t = \hat{\sigma}^2/\hat{\beta}'X'X\hat{\beta}$. For $t = 0$, the condition number is 4,167, which is the smallest condition number attainable with this estimator. While this is still quite a large condition number, it does represent a 90.1% decrease over the original least squares condition number. For the observed data, $t = .0004$, which results in an observed condition of 4408, an 89.5% decrease over least squares.

The choice of a_i's that is in column II represents the factors that will yield the greatest potential improvement in conditioning without sacrificing minimaxity. For this choice of a_i's we use the scale factor $(m/(m+2))M(a,\lambda)/\max_i(a_i/\lambda_i) = (618)(.138)/200 = .0005$, which yields a condition number of

$$\frac{\lambda_1}{\lambda_p}\left(\frac{t + .0005/\lambda_1}{t + .0005 \times .0004/\lambda_p}\right) .$$

At $t = 0$ the condition number is 2,500, a 94% improvement over least squares, while for the observed data, the condition number is 9,083, a 78% improvement.

Thus while this choice of constants guaranteed the greatest potential improvement, they did not provide the greatest actual improvement.

The condition numbers of these estimators are increasing functions of t, with a maximum value of λ_1/λ_p . Since $t = \hat{\sigma}^2/\hat{\beta}'X'X\hat{\beta}$, we can expect maximum improvement when the least squares estimator provides a good fit to the data.

Marquart and Snee (1975) also investigated a reduced five-term model which was better conditioned than the original nine-term model. The eigenvalues of the five-term model are (2.37, 1.08, .855, .690, .01). For these eigenvalues, however, $\sum_1^4 \lambda_i^{-2} - (\lambda_1\lambda_p)^{-1} < 0$, and thus no minimax estimator in the class considered here can guarantee improved conditioning. So although the five-term model is better conditioned ($\lambda_1/\lambda_p = 237$), the structure of the eigenvalues will not allow improved conditioning and minimaxity. This may be partly due to the fact that the five-term model has only one small eigenvalue, while the nine-term model has four small eigenvalues. One should realize, however, that a condition number of 237 is not that bad, and there should be hesitation in replacing the least squares estimator in such cases.

It has, therefore, been illustrated that it is not the case that, as the condition number increases, the possibility of improving conditioning with a minimax estimator decreases. There does not seem to be any simple relationship between minimaxity and condition number, but rather the structure of the entire set of eigenvalues must be considered.

7. Summary and Comments

In ill-conditioned problems there is a dichotomy between the two original goals of ridge regression; the improvement of numerical conditioning and the reduction of mean square error. Multicollinear predictor variables will tend

to produce eigenvalues that satisfy $\lambda_i^2 > \lambda_1\lambda_p$ for some values of i between 2 and p . If enough of the eigenvalues satisfy this condition, all minimax estimators in the class considered here will worsen conditioning, and all estimators which improve conditioning will not be minimax.

Explicit conditions have been given that insure the existence of a minimax ridge estimator which improves conditioning. Also, methods for constructing such estimators were presented. It was also seen, through example, that substantial improvement in conditioning can be achieved by minimax estimators. Moreover, the relationship between the eigenvalues (rather than just the condition number) is the important factor in determining the existence of condition-improving minimax estimators. Thus, a large condition number alone does not preclude the existence of condition-improving minimax estimators.

Minimax estimators are the only estimators which can guarantee uniform improvement in risk over the least squares estimator. The requirement of minimaxity is, in general, a rather strong requirement to force upon an estimator. In the context of a regression problem, to achieve minimaxity one is forced to alter the structure of the covariance matrix to make it as symmetric as possible. When this is done using the ridge regression technique the matrix inverted in the calculation of the estimator can become more asymmetric than the original least squares matrix.

If, for a given data set, a minimax estimator cannot improve conditioning, there is no clear way to proceed. Perhaps the best course is to examine the condition number of a minimax estimator (which could be larger than that of the least squares estimator), and also examine the maximum risk of a ridge estimator which improves conditioning (the maximum risk will be larger than that of the least squares estimator). Based on this additional information, the estimator which best suits the experimenter's needs can be chosen.

References

1. Berger, James O. (1976). Tail Minimaxity in Location Vector Problems and Its Applications. Ann. Statist. 4, 33-50.

2. Bock, M. E. (1975). Minimax Estimators of the Mean of a Multivariate Normal Distribution. Ann. Statist. 3, 209-218.

3. Brown, L. (1971). Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems. Ann. Math. Statist. 42, 855-903.

4. Casella, George. (1980). Minimax Ridge Regression Estimation. Ann. Statist. 8, 1036-1056.

5. Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. Technometrics 21, 215-223.

6. Hemmerle, W. J., and Brantle, T. F. (1978). Explicit and Constrained Generalized Ridge Estimation. Technometrics 20, 109-120.

7. Hoerl, A. E., and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12, 55-68.

8. Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1975). Ridge Regression: Some Simulations. Comm. in Statist. 6, 473-484.

9. Lawless, J. F., and Wang, P. (1976). A Simulation Study of Ridge and Other Regression Estimators. Comm. in Statist. 5, 307-323.

10. Marquart, D. W., and Snee, R. E. (1975). Ridge Regression in Practice. Amer. Statist. 12, 3-19.

11. McDonald, G., and Galarneau, D. (1975). A Monte Carlo Evaluation of Some Ridge-Type Estimators. J. Amer. Statist. Assoc. 70, 407-416.

12. Obenchain, R. L. (1977). Classical F-Tests and Confidence Regions for Ridge Regression. Technometrics 19, 429-440.

13. Stewart, G. W. (1973). Introduction to Matrix Computations. Academic Press, New York.

14. Strawderman, W. E. (1978). Minimax Adaptive Generalized Ridge Regression Estimators. J. Amer. Statist. Assoc. 73, 623-627.

15. Thisted, Ronald A. (1976). Ridge Regression, Minimax Estimation, and Empirical Bayes Methods. Ph.D. Thesis, Stanford University.

16. Vinod, H. D. (1976). Application of New Ridge Regression Methods to a Study of Bell System Scale Economies. J. Amer. Statist. Assoc. 71, 835-841.