MINIMAX ESTIMATION FOR LINEAR REGRESSIONS[1]

By R. Radner

University of California, Berkeley

1. Introduction and Summary. When estimating the coefficients in a linear regression it is usually assumed that the covariances of the observations on the dependent variable are known up to multiplication by some common positive number, say c, which is unknown. If this number c is known to be less than some number k, and if the set of possible distributions of the dependent variable includes "enough" normal distributions (in a sense to be specified later) then the minimum variance linear unbiased estimators of the regression coefficients (see [1]) are minimax among the set of all estimators; furthermore these minimax estimators are independent of the value of k. (The risk for any estimator is here taken to be the expected square of the error.) This fact is closely related to a theorem of Hodges and Lehmann ([3], Theorem 6.5), stating that if the observations on the dependent variable are assumed to be independent, with variances not greater than k, then the minimum variance linear estimators corresponding to the assumption of equal variances are minimax.

For example, if a number of observations are assumed to be independent, with common (unknown) mean, and common (unknown) variance that is less than k; and if, for every possible value of the mean, the set of possible distributions of the observations includes the normal distribution with that mean and with variance equal to k; then the sample mean is the minimax estimator of the mean of the distribution.

The assumption of independence with common unknown variance is, of course, essentially no less general than the assumption that the covariances are known up to multiplication by some common positive number, since the latter situation can be reduced to the former by a suitable rotation of the coordinate axes (provided that the original matrix of covariances is non-singular).

This note consideres the problem of minimax estimation, in the general "linear regression" framework, when less is known about the covariances of the observations on the "dependent variable" than in the traditional situation just described. For example, one might not be sure that these observations are independent, nor feel justified in assuming any other specific covariance structure. It is immediately clear that, from a minimax point of view, one cannot get along without any prior information at all about the covariances, for in that case the risk of every estimator is unbounded. In practice, however, one is typically willing to grant that the covarainces are bounded somehow, but one may not

Received June 4, 1957; revised May 5, 1958.

[1] Research undertaken by the Cowles Commission for Research in Economics under Contract Nonr-358(01), NR 047-006 with the Office of Naval Research.

have a very precise idea of the nature of the bound. One is therefore led to look for different ways of bounding the covariances, in the hope that the minimax estimators are not too sensitive to the bound.

Unfortunately, in the directions explored here, the minimax estimator is sensitive to the "form" of the bound, although once the form has been chosen the minimax estimator does not depend on the "magnitude" of the bound. This result thus provides an instance in which the minimax principle is not too effective against the difficulties due to vagueness of the statistical assumptions of a problem, although this is a type of situation in which it has often been successful (see Savage in [4], pp. 168–9).

In this note, two ways of bounding the covariances are considered. The first is equivalent to choosing a coordinate system for the "dependent variables," and placing a bound on the characteristic roots of the matrix of covariances of the coordinates, in terms of one of a certain class of metrics (e.g., placing a bound on the trace on the covariance matrix, or on its largest characteristic root). The second way consists of choosing a coordinate system, and then placing a bound on the variance of each coordinate.

In the first situation, the minimum variance linear unbiased estimator corresponding to the case of uncorrelated coordinates, with equal variances, turns out to be minimax; this minimax estimator is, in general, different for different choices of coordinate system, but does not depend on the "magnitude" of the bound. Also, the minimax loss typically decreases at the rate of the reciprocal of the sample size.

In the second situation, the minimax procedures derived here involve ignoring most of the observations, and applying a linear unbiased estimator to the rest. Again, the minimax procedure depends upon the choice of coordinate system; furthermore, in this case the minimax loss typically either does not approach zero with increasing sample size, or does so much more slowly than the reciprocal of the sample size.

Thus the minimax estimator appears to be less unsatisfactory in the first situation than in the second, but in both cases it depends upon the choice of coordinate system, which is a disadvantage if there is no "natural" coordinate system intrinsic to the regression problem being considered.

Section 2 below presents the formulation of the problem, and a basic lemma. Sections 3 and 4 explore the two ways of bounding the covariances just mentioned. Some examples are given in Section 5. I am indebted to R. R. Bahadur, L. J. Savage, and G. Debreu for their helpful comments.

2. Problem formulation and a basic lemma. Let y be a random N-dimensional column vector, with a distribution p that is known to be in some family P of distributions. Let $m_p = Ey$ denote the mean of the distribution p, and suppose that one is required to estimate the value of $f'm_p$ on the basis of a single observation on y, where f is given. It is assumed that the loss due to incorrect estimation is the square of the error. In this note minimax estimators of $f'm_p$ will be de-

rived under two different assumptions about P; both assumptions have the following form:

Let T be given $N \times M$ matrix; let C_p denote the covariance of p, i.e.,

$$C_p = E(y - m_p)(y - m_p)';$$

and let H be a given set of $N \times N$ covariance matrices.

(2.1) For every p in P, the mean $m_p = Tx$ for some M-dimensional vector x, and C_p is in H.

(2.2) For every x, and every C in H, there is a *normal* distribution in P with mean Tx and covariance C.

The assumption that P includes normal distributions is a natural one, since normality can rarely be ruled out as preposterous.

If α is any estimator, then the risk, or expected loss, associated with using α is, for any p, given by

$$\begin{aligned} r(\alpha, p) &= E[\alpha(y) - f'm_p]^2 \\ &= E[\alpha(y) - E\alpha(y)]^2 + [E\alpha(y) - f'm_p]^2. \end{aligned} \tag{2.3}$$

An estimator $\hat{\alpha}$ is minimax if, for every estimator α,

$$\sup_{p \in P} r(\hat{\alpha}, p) \leqq \sup_{p \in P} r(\alpha, p).$$

Because of the convexity of the risk function, it is not necessary to consider randomized estimators (see [3], Theorem 3.2).

Relative to a given covariance C, an estimator α is said to be *minimum variance linear unbiased*, or more briefly, *Markoff*, if

$$\alpha(y) = a'y \quad \text{(linearity)}. \tag{2.4}$$

$$\text{For every } p \text{ in } P,\ Ea'y = m_p \quad \text{(unbiasedness)}. \tag{2.5}$$

(2.6) If β is any estimator satisfying (2.4) and (2.5), then for every p in P with covariance C, $r(\alpha, p) \leqq r(\beta, p)$.

The significance of the Markoff estimators in this problem is that, in both cases considered in this note, there is a Markoff estimator, relative to some C in H, that is minimax.

It follows from (2.1) that a linear estimator a is unbiased if and only if $T'a = T'f$; and from (2.3) that the risk for a linear unbiased estimator is $a'C_pa$. Therefore, a linear estimator a is Markoff relative to C if and only if it minimizes $a'Ca$ subject to the constraint $T'a = T'f$.

It might be noted here that it follows from (2.3) that the standard definition of a Markoff estimator given above is equivalent to another one in which condition (2.5) (unbiasedness) is replaced by the following (bounded risk):

(2.5′) The risk $E(a'y - f'm_p)^2$ is bounded as p varies in the class of all p in P that have covariance C, for any given C.

The idea of replacing the constraint of unbiasedness by the constraint of bounded risk is close to the minimax spirit, and seems to be due to L. J. Savage.

The main tool that will be used is the following lemma, which is closely related to a theorem of Hodges and Lehman ([3], theorem 6.5), and is stated here without proof.

LEMMA. *If $\hat{a}$ is Markoff relative to $\hat{C}$ in H, and if $\hat{a}'C\hat{a} \leqq \hat{a}'\hat{C}\hat{a}$ for every C in H, then $\hat{a}$ is minimax.*

In the "classical" situation to which the general Markoff theorem on least squares is applied (see, for example, Aitken [1]), it is assumed that the covariance of the distribution p is known up to multiplication by a positive constant, i.e., that the covariance is cC, where C is known but c is not. If it is further assumed that c is bounded by some number k, then it follows immediately from the Lemma that the Markoff estimator relative to kC is minimax. Note that the Markoff estimator is independent of k.

On the other hand, if nothing at all is known about the covariance of p, i.e., if H is taken to be the class of *all* $N \times N$ covariance matrices, then the risk for every estimator is unbounded. To get a finite minimax value, the class H must be "bounded" in some sense, and the next two sections explore two directions in which such a bound can be defined. In each case it should be borne in mind that postulated assumptions are thought of as applying *after, possibly, an appropriate transformation of the coordinate system.*

3. The case of bounds in terms of characteristic roots. In this section minimax estimators are derived for the problem formulated in Section 2, when the covariances are bounded in certain ways in terms of their characteristic roots.

For any covariance matrix C, let r_i denote its characteristic roots (these will be non-negative real numbers). For any number $q \geqq 1$, the *q-norm* of C is defined here to be

$$N(C; q) = \left(\sum_i r_i^q\right)^{1/q}.$$

For $q = 1, 2$, and ∞, one gets the trace of C, the square root of the sum of squares of the elements of C, and the largest characteristic root of C, respectively. Note that for the identity matrix I, $N(I; q) = N^{1/q}$.

THEOREM 1. *Let q and k be given such that $1 \leqq q \leqq \infty$ and $k > 0$, and let H be the set of all covariances C such that $N(C; q) \leqq k$; then for the estimation problem described in Section 2, the Markoff estimator $\hat{a}$ relative to the identity matrix[2] is minimax, and the minimax loss is $k\hat{a}'\hat{a}$.*

PROOF. The idea of the proof is to show that the covariance of rank one that

[2] Strictly speaking, relative to the identity times an appropriate constant, since the identity may not be in H.

concentrates all the variance in the direction of $f'y$ is least favorable. Let $B = \hat{a}\hat{a}'/\hat{a}'\hat{a}$. Note that $N(B; q) = 1$. Since $\hat{a}$ is that unbiased linear estimator with minimum length, any unbiased linear estimator is of the form $\hat{a} + d$, where $\hat{a}'d = 0$. Hence for all unbiased linear estimators b,

$$b'Bb = \hat{a}'B\hat{a} = \hat{a}'\hat{a};$$

in particular, $\hat{a}$ is Markoff with respect to B, and to kB.

Let C be any covariance in H, and let r be its largest characteristic root; then

$$\hat{a}'C\hat{a} \leqq r\hat{a}'\hat{a} = N(C; \infty)\hat{a}'\hat{a} \leqq N(C; q)\hat{a}'\hat{a} \leqq k\hat{a}'B\hat{a}. \tag{3.1}$$

The theorem now follows from the lemma, equation (3.1), and the fact that $\hat{a}$ is Markoff relative to kB.

For the case $q = 1$, it can be shown that the minimax estimator is not unique, but it is not known whether it is unique for $q > 1$. However, the Markoff estimator $\hat{a}$ of Theorem 1 is the only linear minimax estimator, which can be seen as follows. A linear minimax estimator d must have bounded risk, and therefore must be unbiased. Suppose d is different from $\hat{a}$, and let $D = dd'/d'd$; then

$$kd'Dd = kd'd > ka'a,$$

i.e., the risk for d against the covariance kD is greater than the minimax risk.

Note that it follows immediately from Theorem 1, that if the characteristic roots of the covariance matrices in H are defined relative to any fixed symmetric positive definite matrix Q, then the Markoff estimator relative to Q will be minimax.

4. The case of bounds on the variances of the coordinates. In this section minimax estimators are found for the problem of Section 2 in the case in which the class H of covariances is delimited by bounding the variances of given linear functions of the random vector, in other words, by choosing a particular coordinate system and bounding the variances of the colrdinates.

THEOREM 2. *Let $k_1, \cdots, k_N$ be N given positive numbers; let H be the set of covariances C such that $c_{ii} \leqq k_i^2$ for $i = 1, \cdots, N$; then any $\hat{a}$ that minimizes $\sum_i k_i \mid a_i \mid$ subject to $T'_a = T'f$ (unbiasedness) is a minimax estimator for the problem of Section 2, and $\hat{c}^2 = (\sum_i k_i \mid a_i \mid)^2$ is the minimax loss.*

PROOF. There is no loss of generality in assuming that $k_i = 1$ for every i. As in Theorem 1, one is led to look for a least favorable covariance matrix among those of rank 1.

Let U be the set of linear unbiased estimators; for any C in H and b in U,

$$b'Cb = \sum_{ij} b_i b_j c_{ij} \leqq \sum_{ij} \mid b_i b_j \mid (c_{ii} c_{jj})^{\frac{1}{2}} \leqq \sum_{ij} \mid b_i b_j \mid = \left(\sum_i \mid b_i \mid\right)^2. \tag{4.1}$$

Let $\hat{a}$ be any vector that minimizes $\sum_i \mid a_i \mid$ in U, and let $\hat{c} = \sum_i \mid \hat{a}_i \mid$. By equation (4.1), and the lemma, the present theorem is proved if a vector $\hat{e}$ can be found such that (1) $\hat{a}$ is Markoff against $\hat{E} = \hat{e}\hat{e}'$; (2) $\hat{E}$ is in H, i.e., $\hat{e}_i^2 = 1$ for every i; and (3) the risk for $\hat{a}$ against E equals $\hat{c}^2$.

To this end, let S be the set of all vectors b such that $\sum_i \mid b_i \mid \leqq c$. S is a bounded convex polyhedron, and the intersection of S with U is contained in

the boundary of S, by the definition of c. Hence there is a hyperplane supporting S that contains U, i.e., there is a vector $\hat{e}$ such that

$$b'\hat{e} = \hat{c}, \text{ for all } b \text{ in } U, \tag{4.2}$$

$$b'\hat{e} \leqq \hat{c}, \text{ for all } b \text{ in } S \tag{4.3}$$

(see, for example, [2], p. 4).

By (4.2), e satisfies conditions (1) and (3) above. By the definition of S, any vector with one coordinate equal in absolute value to $\hat{c}$, and all other coordinates zero, is in S. Hence, by (4.3), $\hat{c} \mid \hat{e}_i \mid \leqq \hat{c}$, for every i, so that $\hat{e}_i^2 \leqq 1$ for every i; thus condition (2) above is also satisfied, which completes the proof.

Note that Theorem 2 characterizes all the linear minimax estimators, which is easily seen by an argument similar to that which follows Theorem 1.

5. Examples.

1. Suppose that the random variables $y_1, \cdots, y_N$ each have the same mean x, which is to be estimated, and assume that the sum of the variances of the y_i is not greater than k. To apply Theorem 1, Take T to be the $N \times 1$ matrix whose elements are all equal to 1, f to be vector for which $\sum f_i = 1$ (e.g., $[1, 0, \cdots, 0]$), and $q = 1$. It follows that a minimax estimate of $f'm_p = x$ is the arithmetic mean of $y_1, \cdots, y_N$, i.e., $\hat{a} = (1/N, \cdots, 1/N)$, and the minimax loss is $k \sum_i \hat{a}_i^2 = k/N$. This minimax estimator is, of course, the Markoff estimator for the situation in which it is known that the y_i are independent, with equal variances.

The same result would be obtained if it were assumed that the variance of any linear combination $\sum b_i y_i$ such that $\sum b_i^2 = 1$ is not greater than k (the case $q = \infty$).

2. Consider the estimation problem of Example 1, except now assume that the variance of y_i is not greater than k_i^2, $i = 1, \cdots, N$. By Theorem 2, a minimax estimator is given by

$$\hat{a}_i = \begin{cases} 1, & \text{for that } i \text{ for which } k_i \text{ is minimum,} \\ 0, & \text{otherwise,} \end{cases} \tag{5.1}$$

and the minimax loss is $\min_i k_i^2$. Note that in this example the minimax loss is independent of the sample size N, except insofar as $\min_i k_i$ depends upon N. If $k_1 = \cdots = k_N$, then any linear unbiased estimator is minimax.

3. Suppose it is required to estimate the slope e in the linear regression of one variable on another, and it is assumed that the variance of the "dependent variable" is not greater than k^2. To apply Theorem 2, take

$$T' = \begin{bmatrix} 1, \cdots, 1 \\ t_1, \cdots, t_N \end{bmatrix} \quad \text{and} \quad x' = (d, e),$$

where $t_1, \cdots, t_N$ are the values of the "independent variable," and d and e are unknown. A bounded risk (unbiased) linear estimator a must satisfy

$$\begin{aligned} \sum a_i &= 0, \\ \sum a_i t_i &= 1. \end{aligned} \tag{5.2}$$

By Theorem 2, any $\hat{a}$ that minimizes $\sum | a_i |$ subject to equation (5.2) is a minimax estimator of e. Without loss of generality, t_N can be taken to be the largest value of t_i, and t_1 the smallest; then it is not hard to show that the unique solution of the above minimization problem is

$$a_i = \begin{cases} \dfrac{-1}{t_N - t_1}, & \text{for} \quad i = 1, \\ \dfrac{1}{t_N - t_1}, & \text{for} \quad i = N, \\ 0, & \text{otherwise;} \end{cases} \tag{5.3}$$

and the minimax loss is $4k^2/(t_N - t_1)^2$. In other words, a minimax estimate of e is obtained by taking the slope of the line passing through the "extreme" points (y_1, t_1) and (y_N, t_N).

4. Consider the estimation problem of Example 3, but assume that the sum of the variances of $y_1, \cdots, y_N$ is not greater than k. As in Example 1, this corresponds to taking $q = 1$ in Theorem 1. By Theorem 1 the usual least squares estimate $\sum [(y_i - \bar{y})(t_i - \bar{t})]/(t_i - \bar{t})^2$ is a minimax estimate of e, and the minimax loss is $k/\sum (t_i - \bar{t})^2$.

Suppose further that $t_i = i - 1$ (e.g., think of t_i as successive times), and consider the transformation (taking successive differences)

$$z_i = \begin{cases} y_1, & \text{for } i = 1, \\ y_i - y_{i-1}, & \text{for } 1 = 2, \cdots, N. \end{cases} \tag{5.4}$$

The means of the z_i are

$$Ez_i = \begin{cases} d, & \text{for } i = 1, \\ e, & \text{for } i = 2, \cdots, N. \end{cases} \tag{5.5}$$

Now assume that the sum of the variances of the *new variables* z_i is not greater than k; then by Theorem 1 a minimax estimate of e is

$$\frac{1}{N-1} \sum_2^N z_i = \frac{y_N - y_1}{N-1},$$

and the minimax loss is $k/(N-1)$, a different result from that obtained before making the transformation (5.4).

REFERENCES

[1] A. C. Aitken, "On least squares and the linear combination of observations," *Proc. Roy. Soc. Edinburgh*, Vol. 55 (1935).

[2] T. Bonnesen and W. Fenchel, *Theorie der Konvexen Körper*, J. Springer, Berlin, 1934.

[3] J. L. Hodges and E. L. Lehmann, "Some problems in minimax point estimation," *Ann. Math. Stat.*, Vol. 21 (1950), 182-197.

[4] L. J. Savage, *The Foundations of Statistics*, John Wiley & Sons, New York, 1954.