COWLES FOUNDATION DISCUSSION PAPER NO. 28

Note: Cowles Foundation Discussion Papers are preliminary materials circulated privately to stimulate private discussion and critical comment. References in publications to Discussion Papers (other than mere acknowledgment by a writer that he has access to such unpublished material) should be cleared with the author to protect the tentative character of these papers.

Minimax Estimation for Linear Regressions*

Roy Radner

April 4, 1957

---

* Research undertaken by the Cowles Commission for Research in Economics under Contract Nonr-358(01), NR 047-006 with the Office of Navel Research

---

Minimax Estimation for Linear Regressions

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 considers 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 with no 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 covariances are bounded somehow, but one may not 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

In the estimation problems discussed in this note, the role of the coordinate system is best understood against a coordinate-free background. Let Y be an N-dimensional real vector space, let X be the space of all linear functionals on Y, and let P be a family of probability measures on the Borel sets* of Y.

---

* i.e., the field of sets generated by the class of sets of the form $\{ y \mid xy \leq c \}$, where x is in X, and c is any real number.

A random vector in Y will be denoted by η, fixed vectors in Y and X by lower case Roman letters, and the value of a linear functional x at a point y by xy.

It is assumed that, for every p in P,

$$E(x_1\eta)\,(x_2\eta) < \infty ,$$

for all x_1 and x_2 in X. The mean of the distribution p is that vector m in Y such that

$$E\, x\eta = xm$$

for all x in X, and will be denoted by m_p. The covariance of p is the symmetric non-negative bilinear form C_p on X defined by

$$C_p(x_1,x_2) = E[x_1(\eta - m_p)]\,[x_2(\eta - m_p)] .$$

Note that C_p need not be strictly positive definite.

The distribution p is said to be normal if $x\eta$ is normally distributed for every x in X.

Let f be a given linear functional in X, and suppose that one is required to estimate fm_p on the basis of a single observation of η. It is assumed that the loss due to incorrect estimation is the square of the error. In this note, minimax estimators of fm_p will be derived under various assumptions about P; all these assumptions have the following form in common:

Let M be a given linear subspace of Y, and let H be a given set of symmetric non-negative bilinear forms on X.

(2.1) For every probability measure p in P, the mean of p is in M and the covariance of p is in H.

(2.2) For every m in M and C in H, there is a normal p in P with mean m 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 an estimator (i.e., a measurable real-valued function on Y), then the risk, or expected loss, associated with using α is, for any p, given by

$$(2.3) \qquad r(\alpha,p) = E[\alpha(\eta) - fm_p]^2$$

$$= E[\alpha(\eta) - E\alpha(\eta)]^2 + [E\alpha(\eta) - fm_p]^2 .$$

In other words, the risk $r(\alpha,p)$ is the sum of the variance of α and the square of the bias of α. An estimator $\hat{\alpha}$ is minimax if, for every estimator α,

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

Within the set of all estimators, a particular subclass is of special significance, namely the class of all minimum variance linear unbiased estimators, or more briefly, Markoff estimators. Let C be any covariance in H; then the estimator a is said to be Markoff relative to C if

(2.4) a is in X (linearity)

(2.5) For every p in P, $Ea\eta = fm_p$ (unbiasedness)

(2.6) If b is any estimator satisfying (2.4) and (2.5), then for every p in P with covariance C, the risk using a is not greater than the risk using b, i.e.,

$$E(a\eta - fm_p)^2 \leqq E(b\eta - fm_p)^2 .$$

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

(2.5') The risk $E(a\eta - fm_p)^2$ is bounded as p varies in the class of all p in P that have covariance 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.

To see the equivalence of unbiasedness and bounded risk, for linear estimators, first observe that the risk for a linear estimator a is, by (2.3)

$$r(a,p) = C_p(a,a) + [(a - f)m_p]^2 . \tag{2.7}$$

The risk (2.6) is bounded in p if and only if the square of the bias (the second term on the right side of (2.6)) is bounded, which in turn is true if and only if the bias is zero for all m_p in M.

It follows immediately from (2.7) that for any estimator a satisfying (2.4) and (2.5), the risk is

$$r(a,p) = C_p(a,a). \tag{2.8}$$

Therefore, an estimator $\hat{a}$ is Markoff relative to C if and only if it minimizes $C(a,a)$ for a in the linear variety $(f + M^o)$, where M^o is the annihilator* of M.

---

* i.e., M^o is the set of all x in X such that $xm = 0$ for all m in M.

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

The main tool that will be used is Lemma 3 below, which is an immediate consequence of the following two lemmas. The first of these is a trivial and well-known fact about risk functions in general.

Lemma 1. If $r(\alpha,p) \leqq k$ for all p in P, and α is minimax against a subset P' of P on which $\sup_{p\epsilon P'} r(\alpha,p) = k$, then α is minimax against all of P.

The next lemma is a slight modification of a theorem of Hodges and Lehmann (theorem 6.5 of [3]).

Lemma 2. Let C be a given covariance, and let H consist of C alone; then the Markoff estimator relative to C is minimax.

Lemma 3. If $\hat{\alpha}$ is Markoff relative to $\hat{C}$ in H, and if $C(\hat{a},\hat{a}) \leqq \hat{C}(\hat{a},\hat{a})$ for every C in H, then $\hat{\alpha}$ is minimax.

Proof. By the hypothesis of the lemma, and equation (2.8),

$$r(\hat{a},p) \leqq \hat{C}(\hat{a},\hat{a}),$$

for every p in P. Let P' be the set of all p in P with covariance $\hat{C}$. By Lemma 2, $\hat{a}$ is Markoff against P', with minimax value $\hat{C}(\hat{a},\hat{a})$. Lemma 1 can now be applied to complete the proof.

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 Lemma 3 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 symmetric non-negative bilinear forms on X, 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.

3. The Case of Bounds in Terms of Characteristic Roots

In this section a minimax solution is derived for the estimation problem formulated in Section 2, when the covariances are bounded in certain ways in terms of their characteristic roots relative to a fixed inner product on X. The main result is that, in the class of cases considered, the Markoff estimator relative to the fixed inner product is minimax.

Let Q be a given, fixed, symmetric positive definite bilinear form on X. For any bilinear form C on X, the characteristic roots r_i of C relative to Q are defined, in the usual way, as the characteristic roots of the transformation T, from X into itself, defined by

$$C(x,z) = Q(x, Tz) ,$$

for all x and z in X.

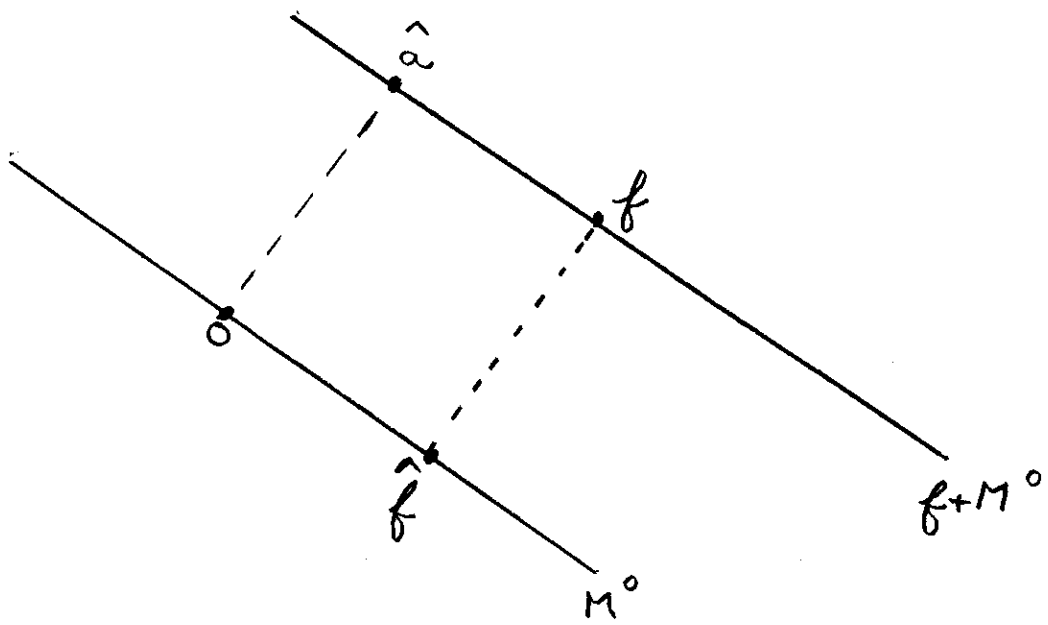
Suppose, further, that C is symmetric and non-negative **semi-definite;** then the characteristic roots of C are non-negative real numbers. For any number $q \geq 1$, the q-norm of C relative to Q is defined here to be

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

For $q = 1,2,$ and ∞, one gets the trace of T, the square root of the trace of T^2 (or the square root of the sum of squares of the elements of the matrix of T relative to any orthonormal coordinate system determined by Q), and the largest characteristic root of T, respectively.

Theorem 1. Let Q be a fixed symmetric positive definite bilinear form on X, let q and k be given such that $1 \leq q \leq \infty$, and $k > 0$, and let H be the set of all covariances C such that $N(C;q,Q) \leq k$; then for the estimation problem described in Section 2, the Markoff estimator $\hat{a}$ relative to the covariance Q is minimax, and the minimax loss is $kQ(\hat{a},\hat{a})$.

Proof. Let $\hat{a}$ minimize $Q(a,a)$ in $f + M^o$ (i.e., $\hat{a}$ is Markoff relative to Q); $\hat{a}$ is equal to $(f - \hat{f})$, where $\hat{f}$ is the Q-orthogonal projection of f into M^o. Hence $\hat{a}$ is Q-orthogonal to M^o (see figure).



Let R be the Q-orthogonal projection onto the (one-dimensional) subspace spanned by $\hat{a}$, and let B be the corresponding bilinear form defined by

$$B(x,z) \equiv Q(x,Rz) \ . \tag{3.1}$$

Note that $N(B;q,Q) = 1$, and hence that kB is in H.

Since $\hat{a}$ is orthogonal to M^o, $R(a) = R(\hat{a}) = \hat{a}$ for every a in $f + M^o$, and

$$B(a,a) = Q(\hat{a},\hat{a}) . \tag{3.2}$$

In particular, $\hat{a}$ is Markoff relative to B, and to kB.

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

$$\begin{aligned} C(\hat{a},\hat{a}) \leqq r\, Q(\hat{a},\hat{a}) &= N(C;\infty,Q)\, B(\hat{a},\hat{a}) \\ &\leqq N(C;q,Q)\, B(\hat{a},\hat{a}) \\ &\leqq kB(\hat{a},\hat{a}) . \end{aligned} \tag{3.3}$$

The theorem now follows from Lemma 3, equation (3.3) 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 other values of q. However the Markoff estimator $\hat{a}$ of Theorem 1 is the only linear minimax estimator, which can be seen as follows. A minimax linear estimator must have bounded risk, i.e., be in $f + M^o$. Let a be in $f + M^o$ and different from $\hat{a}$, let S be the Q-orthogonal projection onto the subspace spanned by a, and let A be the bilinear form defined by

$$A(x,z) = Q(x,Sz) .$$

It follows that

$$kA(a,a) = kQ(a,a) > kQ(\hat{a},\hat{a}) = kB(\hat{a},\hat{a}) ,$$

i.e., the risk for a against the covariance kA is greater than the risk for $\hat{a}$ against kB .

4. The Case of Bounds on the Variances of Fixed Linear Functions.

In this section a minimax solution is derived for the estimation problem of Section 2 for the case in which the class of covariances is delimited by bounding the variances of given linear functions of the random vector, i.e., by choosing a coordinate system in Y and bounding the variances of the coordinates of η .

Theorem 2. Let $b_1,\ldots,b_N$ be a given basis in X, let $k_1,\ldots,k_N$ be N given positive numbers, let H be the set of all covariances C such that $C(b_i,b_i) \leq k_i^2$ for $i = 1,\ldots,N$, and let $\hat{c}$ be the minimum of $\sum_i k_i|\lambda_i|$ subject to the constraint that $\sum_i \lambda_i b_i$ be in $(f + M^o)$; then any $\hat{a} = \sum_i \hat{\lambda}_i b_i$ in $(f + M^o)$ for which $\sum_i k_i|\hat{\lambda}_i| = \hat{c}$ is a minimax estimator for the problem of Section 2, and $\hat{c}^2$ is the minimax loss.

Proof. There is no loss of generality in assuming that $k_i = 1$ for every i . Let

$$S = \left\{x \middle| x = \sum_i \lambda_i b_i, \sum_i k_i |\lambda_i| \leq \hat{c}\right\} .$$

S is convex, and the intersection of S and $(f + M^o)$ is contained in the boundary of S . Hence, by the Hahn - Banach Theorem (see [2], p. 69, for example),

there is a $\hat{y}$ in Y such that

$$(4.1)\qquad \begin{cases} a\hat{y} = \hat{c}, & \text{for all } a \text{ in } (f + M^{o}), \\ x\hat{y} \leq \hat{c}, & \text{for all } x \text{ in } S \end{cases}$$

It follows that, for every i, $b_i\hat{y} \leq 1$, since $\hat{c}b_i$ is in S .

Let $\hat{C}$ be the covariance defined by

$$\hat{C}(x,z) = (x\hat{y})\ (z\hat{y});$$

then $\hat{C}(a,a) = \hat{c}^2$ for all a in $(f + M^{o})$. In particular $\hat{a}$ is Markoff against $\hat{C}$. On the other hand, for every C in H it follows, using Schwarz' Inequality, that

$$\begin{aligned} C(\hat{a},\hat{a}) &= \sum_{i,j} \hat{\lambda}_i\hat{\lambda}_j\, C(b_i,b_j) \\ &\leq \sum_{i,j} |\hat{\lambda}_i\hat{\lambda}_j| C(b_i,b_i)^{1/2}\ C(b_j,b_j)^{1/2} \\ &\leq (\textstyle\sum_i |\hat{\lambda}_i|)^2 = \hat{c}^2 = \hat{C}(\hat{a},\hat{a}) . \end{aligned}$$

Lemma 3 can now be applied to complete the proof of Theorem.

Note that if a is in $(f + M^{o})$ but not in S, then, by an argument similar to that used in Theorem 2, there is a covariance A of rank 1 in H such that $A(a,a) > \hat{c}^2$. Hence Theorem 2 characterizes all the linear minimax estimators.

5. Examples

1. Suppose that the random variables $\eta_1,\ldots,\eta_N$ each have the same mean m_p, which is to be estimated, and assume that the sum of the variances of the η_i is not greater than k. To apply Theorem 1, take Y to be N-dimensional Euclidean space, $X = Y$, $xy = \sum_i x_i y_i$, M the set of all y such that $y_1 = y_2 = \ldots = y_N$, M^o the set of all x such that $\sum x_i = 0$, f any vector for which $\sum f_i = 1$ (e.g., $(1,0,\ldots,0)$), Q the form corresponding to the Euclidean metric (i.e., $Q(x,z) = \sum_i x_i z_i$), and $q = 1$. It follows that a minimax estimate of m_p is the arithmetic mean of $\eta_1,\ldots,\eta_N$, i.e., $\hat{a} = (\frac{1}{N},\ldots,\frac{1}{N})$, and the minimax loss is $k\sum_i \hat{a}_i^2 = \frac{k}{N}$. This minimax estimator is, of course, the Markoff estimator for the situation in which it is known that the η_i are independent, with equal variances.

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

2. Consider the estimation problem of Example 1, except now assume that the variance of η_i is not greater than k_i^2, $i = 1,\ldots,N$. 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 $k_1 = \ldots = k_N$, then any linear unbiased estimator is minimax.

3. Suppose it is required to estimate the slope 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 $Y = X =$ N-dimensional Euclidean space (when N is the sample size), $xy = \sum x_i y_i$, and $E\eta_i = d + et_i$, where $t_1,\ldots,t_N$ are the values of the "independent variable," and d and e are unknown. A bounded risk linear estimator a must satisfy

$$\text{(5.2)} \qquad \begin{cases} \sum a_i = 0 \\ \sum a_i t_i = 1 . \end{cases}$$

By Theorem 2, any 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

$$\text{(5.3)} \qquad \hat{a}_i = \begin{cases} \dfrac{-1}{t_N - t_1}, & \text{for } i = 1, \\ \dfrac{1}{t_N - t_1}, & \text{for } i = N, \\ 0, & \text{otherwise;} \end{cases}$$

and the minimax loss is $\dfrac{4k^2}{(t_N - t_1)^2}$. 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 the coordinates of η is not greater than k. As in Example 1, this corresponds to taking $Q(x,z) = \Sigma x_i z_i$, and assuming that the trace of the covariance, relative to Q, is not greater than k. Suppose further that $t_i = i - 1$ (e.g., think of t_i as successive times). By Theorem 1 the usual least squares estimate $\frac{\Sigma(\eta_i - \eta)(t_i - t)}{\Sigma(t_i - t)^2}$ is a minimax estimate of e, and the minimax loss is $\frac{k}{\sum_i (t_i - t)^2}$.

Now consider the transformation (taking successive differences)

$$\xi_i = \begin{cases} \eta_i , & \text{for } i = 1, \\ \eta_i - \eta_{i-1} , & \text{for } i = 2, \ldots, N . \end{cases} \tag{5.4}$$

The means of the ξ_i are

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

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

$$\frac{1}{N-1} \sum_2^N \xi_i = \frac{\eta_N - \eta_1}{N-1} ,$$

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

References

[1] Aitken, A. C., "On least squares and the linear combination of observations," Proceedings of the Royal Society of Edinburgh, Vol. 55 (1935), 42-48

[2] Bourbaki, N., Espaces Vectoriel Topologiques, Paris: Hermann and Company, 1953

[3] Hodges, J. L. and Lehmann, E. L., "Some problems in minimax point estimation," Annals of Mathematical Statistics, Vol. 21 (1950), 182-197

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