

UNIVERSITY OF NORTH CAROLINA  
Department of Statistics  
Chapel Hill, N. C.

Mathematical Sciences Directorate  
Air Force Office of Scientific Research  
Washington 25, D. C.

AFOSR Report No. **892**

SOME PROPERTIES OF THE LEAST SQUARES ESTIMATOR  
IN REGRESSION ANALYSIS WHEN THE 'INDEPENDENT' VARIABLES ARE STOCHASTIC

by

P. K. Bhattacharya  
University of North Carolina

June, 1961

Contract No. 49(638)-213

Qualified requestors may obtain copies of this report from the ASTIA Document Service Center, Arlington Hall Station, Arlington 12, Virginia. Department of Defense contractors must be established for ASTIA services, or have their "need-to-know" certified by the cognizant military agency or have their project or contract.

Institute of Statistics  
Mimeograph Series No. 292

## 1. Introduction and Summary.

In the classical linear estimation set-up, we have

$$(1) \quad E \underline{y} = X \underline{\theta},$$

where  $\underline{y}' = (y_1, \dots, y_n)$  is a random vector whose components are uncorrelated and have equal variance,  $\underline{\theta}' = (\theta_0, \theta_1, \dots, \theta_p)$  is a vector whose elements are unknown constants and  $X$  is a matrix of  $n$  rows and  $p+1$  columns,  $n \geq p+1$ , which has full rank and whose elements are known constants. Plackett [2] gives a historical note on the least squares estimator

$$\hat{\underline{\theta}} = (X'X)^{-1} X'y$$

of  $\underline{\theta}$ , for which the following property is well-known, --

(I) Each component  $\hat{\theta}_j$  of  $\hat{\underline{\theta}}$  is the uniformly minimum variance, unbiased, linear (in  $y$ 's) estimator of the corresponding component  $\theta_j$  of  $\underline{\theta}$ .

It can also be easily seen that

(II) For a quadratic loss function for the estimator of each component  $\theta_j$  of  $\underline{\theta}$ , the least squares estimators have uniformly minimum risk among the class of all linear (in  $y$ 's) estimators with bounded risk.

Lehmann and Hodges [1] have shown that if we do not restrict ourselves to estimators which are linear in  $y$ 's, then the least squares estimators have the following weaker property, --

(III) If the loss in estimating the true vector  $\underline{\theta}$  by another vector  $\underline{\beta}$  is  $(\underline{\theta} - \underline{\beta})' (\underline{\theta} - \underline{\beta})$ , then the least squares estimator is minimax among the class of all estimators of  $\underline{\theta}$  if there exists a number  $v$  such that

$\text{Var } y_i \leq v, i = 1, \dots, n$ , and the family of distributions  $\mathcal{F}$  of  $(y_1, \dots, y_n)$  contains the sub-family  $\mathcal{F}_0$  of all independent normal distributions of  $(y_1, \dots, y_n)$  which satisfy (1) for some  $\underline{\theta}$ , and have  $\text{Var } y_i = v, i = 1, \dots, n$ .

In many situations, however,  $(y, x_1, \dots, x_p)$  follows a  $(p+1)$  - variate distribution on which observations are made and the method of least squares is

applied to estimate the linear regression of  $y$  on  $x_1, \dots, x_p$ , regarding the  $x$  - observations to be non-stochastic. This problem differs from the classical problem of linear estimation in two respects, --

(a) Instead of (1) the model should be

$$E \int \underline{y} \mid X \int = X \underline{\theta},$$

where the elements of the  $X$  matrix are stochastic, and

(b) the problem being that of estimating the entire regression function rather than the coefficients of the linear regression individually, the loss function should be defined in the space of all linear functions of  $x_1, \dots, x_p$ . For reasons given in section 2, the loss in estimating the true regression function  $\phi(x_1, \dots, x_p)$  by another function  $\psi(x_1, \dots, x_p)$  may be considered to be of the form,

$$\int \int \phi(x_1, \dots, x_p) - \psi(x_1, \dots, x_p) \int^2 d F(x_1, \dots, x_p),$$

where  $F$  is the distribution function of  $(x_1, \dots, x_p)$ .

For the above loss function, it is shown under certain conditions that if the class of estimates which are linear in  $y$ 's and have bounded risk is non-empty, then the estimate obtained by the method of least squares belongs to this class and has uniformly minimum risk in this class. A necessary and sufficient condition on  $F(x_1, \dots, x_p)$  is obtained for this class to be non-empty, which unfortunately is not easy to verify in particular cases and is violated in a very simple situation. However, by a sequential modification of the sampling scheme, this condition may always be satisfied at the cost of an arbitrarily small increase in the expected sample size. It is also shown under certain further conditions on the family of admissible distributions that the least squares estimator is minimax in the class of all estimators.

## 2. Formal statement of the problem.

Let  $y, x_1, \dots, x_p$  be real-valued random variables whose joint distribution satisfies the following conditions, --

Condition (i). The distribution function  $F$  of  $x_1, \dots, x_p$  is such that

\*(a) for every non-null  $\underline{a}' = (a_0, a_1, \dots, a_p)$ , the set

$$S_{\underline{a}} = \{(x_1, \dots, x_p) : a_0 + a_1 x_1 + \dots + a_p x_p = 0\}$$

has probability measure zero.

(b)  $\mu_{jj'} = E(x_j x_{j'})$  is finite,  $j, j' = 0, 1, \dots, p$ ,  $x_0 \equiv 1$ .

Condition (ii).  $E\sqrt{y|x_1, \dots, x_p}$  is a linear function of  $x_1, \dots, x_p$ , say

$$\phi(x_1, \dots, x_p) = \theta_0 + \theta_1 x_1 + \dots + \theta_p x_p.$$

Condition (iii).  $V\sqrt{y|x_1, \dots, x_p} = \sigma^2 < \infty$ , which does not depend on

$$(x_1, \dots, x_p).$$

$(y_i, x_{1i}, \dots, x_{pi})$ ,  $i=1, \dots, n$ ,  $n \geq p+1$ , are mutually independent observations on  $(y, x_1, \dots, x_p)$ . The problem is to estimate the regression function

$$\phi(x_1, \dots, x_p) = \theta_0 + \theta_1 x_1 + \dots + \theta_p x_p,$$

or in other words the row vector  $\underline{\theta}' = (\theta_0, \theta_1, \dots, \theta_p)$ , where the loss involved in estimating  $\underline{\theta}$  by  $\underline{\beta}$  is,

$$\begin{aligned} (2) \quad W(\underline{\theta}, \underline{\beta}) &= \int \int (\theta_0 + \theta_1 x_1 + \dots + \theta_p x_p) - (\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p) \Big|^2 dF(x_1, \dots, x_p) \\ &= \sum_{j=0}^p \sum_{j'=0}^p \mu_{jj'} (\theta_j - \beta_j) (\theta_{j'} - \beta_{j'}) \\ &= (\underline{\theta}' - \underline{\beta}') M (\underline{\theta} - \underline{\beta}), \end{aligned}$$

where

$$M = \begin{bmatrix} 1 & \mu_{01} & \dots & \mu_{0p} \\ \mu_{01} & \mu_{11} & \dots & \mu_{1p} \\ \dots & \dots & \dots & \dots \\ \mu_{0p} & \mu_{1p} & \dots & \mu_{pp} \end{bmatrix}.$$

It follows from condition (i) that  $M$  is positive definite, and

$$0 < W(\underline{\theta}, \underline{\beta}) < \infty$$

---

\*This condition is satisfied if  $x$  has a continuous distribution and  $x_j = x^j$  or if  $x_1, \dots, x_p$  are  $\sin x, \sin 2x, \dots, \cos x, \cos 2x, \dots$ .

for all  $\underline{\beta} \neq \underline{\theta}$ .

The loss function (2) is motivated by the following consideration. Suppose we are required to predict the value of  $y$  associated to a random observation made on  $(x_1, \dots, x_p)$  subject to a quadratic loss. If the true regression function

$$\phi(x_1, \dots, x_p) = \theta_0 + \theta_1 x_1 + \dots + \theta_p x_p$$

were known, our prediction rule would be

$$y(x_1, \dots, x_p) = \theta_0 + \theta_1 x_1 + \dots + \theta_p x_p,$$

and the risk of the procedure would be  $\sigma^2$ . If however, we use the prediction rule

$$y'(x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p,$$

then the risk would be  $\sigma^2 + W(\underline{\theta}, \underline{\beta})$ .

### 3. Optimum property of the least squares estimator in the class of linear estimators with bounded risk.

In this section we shall restrict our attention only to those procedures which are linear in  $y$ 's and have bounded risk. Let us denote the class of all such procedures by  $\mathcal{C}_1$ . Then an estimator  $\underline{t} \in \mathcal{C}_1$  if and only if

$$(3) \quad \frac{\underline{t}}{p+1 \times 1} = \frac{L}{p+1 \times n} \frac{\underline{y}}{n \times 1}$$

and  $r(\underline{\theta}, \underline{t}) = E [W(\underline{\theta}, \underline{t})]$

is a bounded function of  $\theta_0, \theta_1, \dots, \theta_p$ , where each element  $l_{ji}$  of  $L$  is a function of  $x_{11}, \dots, x_{1n}, \dots, x_{p1}, \dots, x_{pn}$ , and  $\underline{y}' = (y_1, \dots, y_n)$ . Let

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & \dots & x_{pn} \end{bmatrix}$$

Then the estimator obtained by the method of least squares is,

$$\underline{t}^* = (X'X)^{-1} Xy.$$

Since  $n \geq p+1$ , it follows from condition (ia) that the matrix  $X$  has rank  $p+1$  with probability 1 and therefore  $\underline{t}^*$  can be uniquely determined for almost all samples.

In what follows,  $E_{y|x} [U(x,y)]$  stands for  $E_y [U(x,y) | x]$ .

We shall first show that if an estimator  $\underline{t} \in \mathcal{C}_1$ , then for almost all  $X$ , the conditional expectation of each component of  $\underline{t}$  given  $X$  should be equal to the corresponding component of  $\underline{\theta}$ , whatever  $\underline{\theta}$  may be.

For any  $\underline{t} \in \mathcal{C}_1$ , it follows from (3) that

$$E_{y|x}(t_j) = \sum_{i=1}^n l_{ji}(\theta_0 + \theta_1 x_{1i} + \dots + \theta_p x_{pi}) = \theta_j + \sum_{j'=0}^p h_{jj'} \theta_{j'}, \quad \text{say,}$$

where  $h_{jj'}$  are functions of  $X$ . Let  $H$  be a  $(p+1) \times (p+1)$  matrix which has  $h_{jj'}$  in its  $j$ th row and  $j'$ th column. To every  $\underline{t} \in \mathcal{C}_1$ , there corresponds such a matrix  $H$ , and to show this correspondence, we shall use the notation

$H_{\underline{t}}$ . Thus if  $\underline{t} \in \mathcal{C}_1$ , then

$$(4) \quad E_{y|x}(\underline{t}) - \underline{\theta} = H_{\underline{t}} \underline{\theta}.$$

We now prove

Lemma 1. If  $\underline{t} \in \mathcal{C}_1$ , then

$$E_{y|x}(\underline{t}) \equiv \underline{\theta}$$

for almost all  $X$ .

Proof. By virtue of (4) it will be enough to show that if

$$P[H_{\underline{t}} \neq 0] > 0,$$

then  $\underline{t} \notin \mathcal{C}_1$ . Let  $\underline{t}$  satisfy (3); then we shall show that  $r(\underline{\theta}, \underline{t})$  is unbounded. We have

$$\begin{aligned} r(\underline{\theta}, \underline{t}) &= E_{X,y} [(\underline{t}' - E_{y|x}(\underline{t}')) M (\underline{t} - E_{y|x}(\underline{t}))] \\ &\quad + E_X [ \{ E_{y|x}(\underline{t}') - \underline{\theta}' \} M \{ E_{y|x}(\underline{t}) - \underline{\theta} \} ] \end{aligned}$$

$$\begin{aligned} &\geq E_X \left[ \{ E_{\underline{y}|X}(\underline{t}') - \underline{\theta}' \}' M \{ E_{\underline{y}|X}(\underline{t}) - \underline{\theta} \} \right] \\ &= E_X \left[ \underline{\theta}' H_{\underline{t}}' M H_{\underline{t}} \underline{\theta} \right]. \end{aligned}$$

If  $P[\underline{H}_{\underline{t}} \neq \underline{0}] > 0$ , some element of  $H_{\underline{t}}$  is non-zero with positive probability. Let that element be one in the  $j_0$ th column of  $H_{\underline{t}}$ . Let

$$A_{j_0} = \{ \underline{\theta} : \theta_j = 0 \text{ for } j \neq j_0 \}.$$

Then for  $\underline{\theta} \in A_{j_0}$ ,

$$\underline{\theta}' H_{\underline{t}}' M H_{\underline{t}} \underline{\theta} = \theta_{j_0}^2 g(X)$$

where  $P[g(X) \geq 0] = 1$  and  $P[g(X) > 0] > 0$ , since  $M$  is positive definite. Hence there exists  $\delta > 0$  such that

$$P(\delta) = P[g(X) > \delta] > 0.$$

Then for  $\underline{\theta} \in A_{j_0}$ ,

$$r(\underline{\theta}, \underline{t}) \geq E_X \left[ \underline{\theta}' H_{\underline{t}}' M H_{\underline{t}} \underline{\theta} \right] \geq \theta_{j_0}^2 \delta P(\delta),$$

and for any given  $c$ ,  $r(\underline{\theta}, \underline{t})$  can be made greater than  $c$  by choosing  $\underline{\theta}$  in  $A_{j_0}$  with

$$|\theta_{j_0}| > \sqrt{c/\delta P(\delta)}.$$

This completes the proof.

The following corollary is immediate.

Corollary. If  $\underline{t} \in \mathcal{C}_1$ , then

$$(5) \quad r(\underline{\theta}, \underline{t}) = \sigma^2 E_X \sum_{j=0}^p \sum_{j'=0}^p \mu_{jj'} \sum_{i=1}^n l_{ji} l_{j'i}$$

where  $l_{ji}$  are the elements of the matrix  $L$  through which  $\underline{t}$  is defined and they satisfy

$$(6) \quad \sum_{i=1}^n l_{ji} x_{ji} = 1, \quad \sum_{i=1}^n l_{ji} x_{j'i} = 0 \text{ for } j \neq j' = 0, 1, \dots, p$$

for almost all  $X$ .

It follows from the above corollary that for  $\underline{t} \in \mathcal{C}_1$ ,  $r(\underline{\theta}, \underline{t})$  does not depend on  $\underline{\theta}$ . Therefore, if we minimize the right side of (5) with respect to  $l_{ji}$ 's subject to the conditions in (6), the resulting matrix  $\hat{L}$  will define an estimator

$$\hat{\underline{t}} = \hat{L} \underline{y}$$

for which

$$r(\underline{\theta}, \hat{\underline{t}}) \leq r(\underline{\theta}, \underline{t})$$

for all  $\underline{\theta}$  and for arbitrary  $\underline{t} \in \mathcal{C}_1$ .

Now the equations giving  $\hat{L}$  are

$$(7) \quad M \hat{L} = \Delta X'$$

$$(8) \quad \hat{L} X = I$$

a.e., where  $\Delta$  is a  $(p+1) \times (p+1)$  matrix of constants and  $I$  is the  $(p+1) \times (p+1)$  unit matrix. Solving (7) and (8) we get

$$\hat{L} = (X'X)^{-1} X',$$

and the resulting estimator

$$\hat{\underline{t}} = (X'X)^{-1} X' \underline{y} = \underline{t}^*, \text{ a.e.,}$$

where  $\underline{t}^*$  is the estimator obtained by the method of least squares.

But we still do not know whether  $\underline{t}^* \in \mathcal{C}_1$ . However, it can be easily seen that unless  $\mathcal{C}_1$  is empty,  $\underline{t}^* \in \mathcal{C}_1$ . We thus have

Theorem 1. If  $\mathcal{C}_1$  is non-empty, then the least squares estimator  $\underline{t}^* \in \mathcal{C}_1$ , and

$$r(\underline{\theta}, \underline{t}^*) \leq r(\underline{\theta}, \underline{t})$$

for all  $\underline{\theta}$  and for arbitrary  $\underline{t} \in \mathcal{C}_1$  with a strict inequality holding if

$$P[\underline{t} = \underline{t}^*] < 1.$$

If we denote by  $\mathcal{C}_2$  the class of estimators which are linear in  $y$ 's, then the following corollary is immediate from the fact that  $r(\underline{\theta}, \underline{t}^*)$  is a constant for all  $\underline{\theta}$ .

Corollary. If  $\mathcal{C}_1$  is non-empty, then the least squares estimator  $\underline{t}^*$  is the unique minimax estimator in  $\mathcal{C}_2$ .

The optimum property of  $\underline{t}^*$  thus depends on the non-emptiness of  $\mathcal{C}_1$  which can be characterized in terms of  $F(x_1, \dots, x_p)$ . We have seen that  $\mathcal{C}_1$  is non-empty if and only if  $r(\underline{\theta}, \underline{t}^*)$  is finite. Clearly,

$$r(\underline{\theta}, \underline{t}^*) = \sigma^2 E_X \text{tr} [(X'X)^{-1} M].$$

Let us denote by  $B$  the  $(p+1) \times (p+1)$  matrix of which the  $(j, j')$ th element is the expectation of the corresponding element in  $(X'X)^{-1}$  and for any matrix  $A$ , let us call the largest of the absolute values of its elements, the norm  $\|A\|$  of  $A$ . Then a necessary and sufficient condition for  $\mathcal{C}_1$  being non-empty is,

Condition (ic).  $\|B\| < \infty$ .

Thus if the joint distribution of  $(y, x_1, \dots, x_p)$  satisfies condition (ic) along with (ia), (ib), (ii) and (iii), then we can delete the phrase "if  $\mathcal{C}_1$  is non-empty" in the statements of theorem 1 and its corollary.

#### 4. Minimax property of $\underline{t}^*$ in the class of all estimators.

Let  $F$  be the unknown marginal distribution of  $(x_1, \dots, x_p)$  and let  $\mathcal{G}$  be the family of admissible distributions of  $(y, x_1, \dots, x_p)$  which have  $F$  as the marginal distribution of  $(x_1, \dots, x_p)$ . In this section we assume  $F$  to satisfy conditions (ia), (ib) and (ic) and the family  $\mathcal{G}$  to satisfy condition (ii) and

Condition (iv). 
$$v = \sup_{G \in \mathcal{G}} \sup_{x_1, \dots, x_p} V_G [y | x_1, \dots, x_p] < \infty,$$

where  $V_G [y | x_1, \dots, x_p]$  stands for the conditional variance of  $y$  given  $x_1, \dots, x_p$  under  $G$ .

Condition (v).  $\mathcal{G}$  includes the class  $\mathcal{G}_0$  of all  $G$  obtained by taking the product of the distribution  $F$  of  $(x_1, \dots, x_p)$  with a

conditional distribution of  $y$  given  $x_1, \dots, x_p$  which is normal with mean satisfying (ii) with some  $\underline{\theta}$  and with variance  $v$ .

Under these conditions we shall prove that  $\underline{t}^*$  is a minimax estimate for  $\underline{\theta}$ . We shall require the following lemma to prove the minimax property.

Lemma 2. Let  $A(z)$  be a mapping from an arbitrary space  $Z$  to the space of all  $k \times k$  non-singular matrices such that  $\|A(z)\|$  and  $\|A(z)^{-1}\|$  are both bounded. Let  $U$  be a fixed  $k \times k$  matrix and  $\{b_m\}$  a sequence of real numbers converging to zero. Then,

- (a)  $\text{Det } \underline{A}(z) + b_m \underline{U}$  converges to  $\text{Det } A(z)$  uniformly in  $z$ .
- (b) There exists an integer  $m_0$  such that  $\underline{A}(z) + b_m \underline{U}^{-1}$  exists for  $m \geq m_0$  and for all  $z \in Z$ .
- (c)  $\|\underline{A}(z) + b_m \underline{U}^{-1} - A(z)^{-1}\|$  converges to zero uniformly in  $z$ , where  $\underline{A}(z) + b_m \underline{U}^{-1}$  is defined to be the matrix each of whose elements is  $\max \int \text{Sup}_{z \in Z} \|A(z)\|, \text{Sup}_{z \in Z} \|A(z)^{-1}\| + 1$  if  $\text{Det } \underline{A}(z) + b_m \underline{U} = 0$ .

Proof. Suppose  $\max \int \text{Sup}_{z \in Z} \|A(z)\|, \text{Sup}_{z \in Z} \|A(z)^{-1}\|, \|U\| = c$ .

Then (a) follows from the fact that

$$|\text{Det } \underline{A}(z) + b_m \underline{U} - \text{Det } A(z)| \leq |b_m| c^k \cdot k!, \text{ if } |b_m| \leq 1.$$

Since  $|\text{Det } A(z)^{-1}| \leq c^k \cdot k!$ ,  $|\text{Det } A(z)| \geq 1/c^k \cdot k!$ . Also, it follows from (a) that there exists an integer  $m_0$  such that for  $m \geq m_0$

$$\text{Det } A(z) - 1/2c^k \cdot k! \leq \text{Det } \underline{A}(z) + b_m \underline{U} \leq \text{Det } A(z) + 1/2c^k \cdot k!$$

for all  $z \in Z$ . Hence for  $m \geq m_0$  and for all  $z \in Z$ ,

$$|\text{Det } \underline{A}(z) + b_m \underline{U}| \geq 1/2c^k \cdot k!,$$

and therefore (b) follows.

(c) is proved as soon as we apply (a) to the determinants of  $\{A(z) + b_m U\}$ ,  $m = 1, 2, \dots$  and all their cofactors.

Theorem 2. Under conditions (ia), (ib), (ic), (ii), (iv) and (v), the least squares estimator  $\underline{t}^*$  is minimax for  $\underline{\theta}$  in the class of all estimators.

Proof. We shall first prove that

$$(9) \quad \sup_{G \in \mathcal{G}_0} r(G, \underline{t}^*) \leq \sup_{G \in \mathcal{G}_0} r(G, \underline{t})$$

for all  $\underline{t}$ .

Since  $G \in \mathcal{G}_0$  is completely described by  $\underline{\theta}$ , we can write  $r(\underline{\theta}, \underline{t})$  instead of  $r(G, \underline{t})$  for arbitrary  $\underline{t}$  where  $\underline{\theta}$  corresponds to  $G$ , and variation over  $\mathcal{G}_0$  is the same as variation over  $\underline{\theta}$ . So we shall show that

$$(10) \quad \sup_{\underline{\theta}} r(\underline{\theta}, \underline{t}^*) \leq \sup_{\underline{\theta}} r(\underline{\theta}, \underline{t})$$

for all  $\underline{t}$ . Suppose there exists  $\hat{\underline{t}}$  such that

$$(10) \quad \sup_{\underline{\theta}} r(\underline{\theta}, \hat{\underline{t}}) = \sup_{\underline{\theta}} r(\underline{\theta}, \underline{t}^*) - \epsilon, \epsilon > 0.$$

We shall argue up to a contradiction to (10) and (9) will follow.

Conditions (ib) and (ic) implies that  $E \operatorname{tr} \int (X'X)^{-1} \underline{M} \int$  exists.

Hence we can find a constant  $c(\epsilon)$  such that

$$\left| \int \operatorname{tr} \int (X'X)^{-1} \underline{M} \int \prod_{i=1}^n dF(x_{1i}, \dots, x_{pi}) - E \operatorname{tr} \int (X'X)^{-1} \underline{M} \int \right| < \epsilon/2v.$$

$$\{X: \|(X'X)\| \leq c(\epsilon), \|(X'X)^{-1}\| \leq c(\epsilon)\}$$

Denote the set  $\{X: \|(X'X)\| \leq c(\epsilon), \|(X'X)^{-1}\| \leq c(\epsilon)\}$  by  $R(\epsilon)$ .

Consider the sequence  $\{\xi_m\}$  of a priori distributions of  $\underline{\theta}$  such that

$$d\xi_m(\underline{\theta}) = (2\pi m)^{-(p+1)/2} \exp \int - \frac{1}{2m} \underline{\theta}' \underline{\theta} \int \prod_{j=0}^p d\theta_j.$$

Since the joint probability differential of  $(y, x_1, \dots, x_p)$  under  $G$

is

$$g_{\underline{\theta}}(\underline{y}|X) = \prod_{i=1}^n dy_i \prod_{i=1}^n dF(x_{1i}, \dots, x_{pi})$$

where

$$g_{\underline{\theta}}(\underline{y}|X) = \text{const.} \exp \int - \frac{1}{2v} (\underline{y} - X \underline{\theta})' (\underline{y} - X \underline{\theta}) \int,$$

and  $\underline{\theta}$  corresponds to  $G$ , we have

$$\begin{aligned}
 r(\underline{\xi}_m, \hat{\underline{t}}) &= \int_{\underline{\theta}} \int_X \int_{\underline{y}} (\underline{\theta} - \hat{\underline{t}})' M(\underline{\theta} - \hat{\underline{t}}) g_{\underline{\theta}}(\underline{y}|X) \prod_{i=1}^n dy_i \prod_{i=1}^n \pi dF(x_{1i}, \dots, x_{pi}) d\underline{\xi}_m(\underline{\theta}) \\
 &= \int_X \int_{\underline{y}} \int_{\underline{\theta}} (\underline{\theta} - \hat{\underline{t}})' M(\underline{\theta} - \hat{\underline{t}}) g_{\underline{\theta}}(\underline{y}|X) d\underline{\xi}_m(\underline{\theta}) \prod_{i=1}^n dy_i \prod_{i=1}^n \pi dF(x_{1i}, \dots, x_{pi}) \\
 &\geq \int_{X \in R(\epsilon)} \int_{\underline{y}} \int_{\underline{\theta}} (\underline{\theta} - \hat{\underline{t}})' M(\underline{\theta} - \hat{\underline{t}}) g_{\underline{\theta}}(\underline{y}|X) d\underline{\xi}_m(\underline{\theta}) \prod_{i=1}^n dy_i \prod_{i=1}^n \pi dF(x_{1i}, \dots, x_{pi}) \\
 &= \text{Const.} \int_{X \in R(\epsilon)} \int_{\underline{y}} \int_{\underline{\theta}} (\underline{\theta} - \hat{\underline{t}})' M(\underline{\theta} - \hat{\underline{t}}) \\
 &\quad \times \exp\left[-\frac{1}{2v} \left\{ \underline{\theta} - (X'X + \frac{v}{m} I)^{-1} X' \underline{y} \right\}' (X'X + \frac{v}{m} I) \left\{ \underline{\theta} - (X'X + \frac{v}{m} I)^{-1} X' \underline{y} \right\} \right] \\
 &\quad \times \prod_{j=0}^p \pi d\theta_j \cdot \exp\left[-\frac{1}{2v} \underline{y}' \underline{y} + \frac{1}{2v} \underline{y}' X (X'X + \frac{v}{m} I)^{-1} X' \underline{y} \right] \prod_{i=1}^n dy_i \prod_{i=1}^n \pi dF(x_{1i}, \dots, x_{pi})
 \end{aligned}$$

for sufficiently large  $m$ , since the hypotheses of lemma 2 are ensured for  $X \in R(\epsilon)$  and hence there exists  $m_0$  such that  $(X'X + \frac{v}{m} I)^{-1}$  exists for all  $m > m_0$  and for all  $X \in R(\epsilon)$ . Now for such large values of  $m$  and for any given  $X$  and  $\underline{y}$ ,

$$\begin{aligned}
 &\int_{\underline{\theta}} (\underline{\theta} - \hat{\underline{t}})' M(\underline{\theta} - \hat{\underline{t}}) \\
 &\quad \times \exp\left[-\frac{1}{2v} \left\{ \underline{\theta} - (X'X + \frac{v}{m} I)^{-1} X' \underline{y} \right\}' (X'X + \frac{v}{m} I) \left\{ \underline{\theta} - (X'X + \frac{v}{m} I)^{-1} X' \underline{y} \right\} \right] \\
 &\quad \times \prod_{j=0}^p \pi d\theta_j \\
 &\geq \int_{\underline{\theta}} \left\{ \underline{\theta} - (X'X + \frac{v}{m} I)^{-1} X' \underline{y} \right\}' M \left\{ \underline{\theta} - (X'X + \frac{v}{m} I)^{-1} X' \underline{y} \right\} \\
 &\quad \times \exp\left[-\frac{1}{2v} \left\{ \underline{\theta} - (X'X + \frac{v}{m} I)^{-1} X' \underline{y} \right\}' (X'X + \frac{v}{m} I) \left\{ \underline{\theta} - (X'X + \frac{v}{m} I)^{-1} X' \underline{y} \right\} \right] \\
 &\quad \times \prod_{j=0}^p \pi d\theta_j
 \end{aligned}$$

(since  $M$  is positive definite)

∴ If we define

$$\underline{t}_m = (X'X + \frac{v}{m} I)^{-1} X'y \quad \text{if } (X'X + \frac{v}{m} I)^{-1} \text{ exists}$$

$$= (X'X)^{-1} X'y \quad \text{otherwise,}$$

then  $r(\underline{\xi}_m, \hat{\underline{t}})$

$$\geq \int_{X \in R(\epsilon)} \int_{\underline{y}} \int_{\underline{\theta}} (\underline{t}_m - \underline{\theta})' M(\underline{t}_m - \underline{\theta}) g_{\underline{\theta}}(\underline{y} | X) d\underline{\xi}_m(\underline{\theta}) \int$$

$$\times \prod_{i=1}^n \pi dy_i \prod_{i=1}^n dF(x_{1i}, \dots, x_{pi})$$

$$= \int_{X \in R(\epsilon)} \int_{\underline{\theta}} \int_{\underline{y}} (\underline{t}_m - \underline{\theta})' M(\underline{t}_m - \underline{\theta}) g_{\underline{\theta}}(\underline{y} | X) \prod_{i=1}^n \pi dy_i \int$$

$$\times d\underline{\xi}_m(\underline{\theta}) \prod_{i=1}^n dF(x_{1i}, \dots, x_{pi})$$

$$= \int_{X \in R(\epsilon)} \int_{\underline{\theta}} \int_{\underline{y}} \{ (X'X + \frac{v}{m} I)^{-1} X'y - (X'X + \frac{v}{m} I)^{-1} X'X \underline{\theta} \}' \cdot$$

$$\cdot M \{ (X'X + \frac{v}{m} I)^{-1} X'y - (X'X + \frac{v}{m} I)^{-1} X'X \underline{\theta} \} \times g_{\underline{\theta}}(\underline{y} | X) \prod_{i=1}^n \pi dy_i$$

$$+ \{ (X'X + \frac{v}{m} I)^{-1} X'X \underline{\theta} - \underline{\theta} \}' M \{ (X'X + \frac{v}{m} I)^{-1} X'X \underline{\theta} \} d\underline{\xi}_m(\underline{\theta}) \prod_{i=1}^n \pi dF(x_{1i}, \dots, x_{pi})$$

$$= v \int_{X \in R(\epsilon)} \text{tr} \int (X'X + \frac{v}{m} I)^{-1} X'X (X'X + \frac{v}{m} I)^{-1} M \int \prod_{i=1}^n \pi dF(x_{1i}, \dots, x_{pi})$$

$$+ m \int_{X \in R(\epsilon)} \text{tr} \int \{ (X'X + \frac{v}{m} I)^{-1} X'X - I \} \cdot \{ (X'X + \frac{v}{m} I)^{-1} X'X - I \}' M \int$$

$$\times \prod_{i=1}^n \pi dF(x_{1i}, \dots, x_{pi})$$

$$= v \int_{X \in R(\epsilon)} \text{tr} \int (X'X + \frac{v}{m})^{-1} M \int \prod_{i=1}^n \pi dF(x_{1i}, \dots, x_{pi}) = \rho_m(\epsilon) \quad \text{say.}$$

It follows from lemma 2 that

$$\lim_{m \rightarrow \infty} \rho_m(\epsilon) = v \int_{X \in R(\epsilon)} \text{tr} \underline{\Gamma} (X'X)^{-1} \underline{M} \underline{\Gamma} dF$$

.. For sufficiently large  $m$ ,

$$\rho_m(\epsilon) > v \int_{X \in R(\epsilon)} \text{tr} \underline{\Gamma} (X'X)^{-1} \underline{M} \underline{\Gamma} dF - \epsilon/2$$

$$> v E \text{tr} \underline{\Gamma} (X'X)^{-1} \underline{M} \underline{\Gamma} - \epsilon$$

.. For such large values of  $m$ ,

$$r(\hat{\xi}_m, \hat{\underline{t}}) > v E \text{tr} \underline{\Gamma} (X'X)^{-1} \underline{M} \underline{\Gamma} - \epsilon$$

But for all  $\underline{\theta}$ ,

$$r(\underline{\theta}, \underline{t}^*) = v E \text{tr} \underline{\Gamma} (X'X)^{-1} \underline{M} \underline{\Gamma} = \sup_{\underline{\theta}} r(\underline{\theta}, \underline{t}^*)$$

Now if (10) is true, then

$$\sup_{\underline{\theta}} r(\underline{\theta}, \hat{\underline{t}}) = v E \text{tr} \underline{\Gamma} (X'X)^{-1} \underline{M} \underline{\Gamma} - \epsilon < r(\hat{\xi}_m, \hat{\underline{t}})$$

for sufficiently large  $m$ , which is impossible. Hence (9) holds.

Since  $\mathcal{G}_0 \subset \mathcal{G}$ ,

$$\sup_{G \in \mathcal{G}_0} r(G, \underline{t}) \leq \sup_{G \in \mathcal{G}} r(G, \underline{t})$$

for arbitrary  $\underline{t}$ . Thus,

$$\sup_{G \in \mathcal{G}_0} r(G, \underline{t}^*) \leq \sup_{G \in \mathcal{G}} r(G, \underline{t}) \text{ for all } \underline{t}.$$

$$\text{Now } \sup_{G \in \mathcal{G}_0} r(G, \underline{t}^*) = v E \text{tr} \underline{\Gamma} (X'X)^{-1} \underline{M} \underline{\Gamma} = \sup_{G \in \mathcal{G}} r(G, \underline{t}^*).$$

$$\text{Hence } \sup_{G \in \mathcal{G}} r(G, \underline{t}^*) \leq \sup_{G \in \mathcal{G}} r(G, \underline{t}) \text{ for all } \underline{t}.$$

##### 5. Remarks.

a) Condition (ic) is not satisfied in general. In fact, it can be easily

seen that for  $p = 1$  and for a normal distribution of  $x_1$ , this condition is satisfied if and only if  $n \geq 4$ , though  $\underline{t}^*$  can be uniquely determined for almost all samples of size 2 or more.

b) It is obvious that when condition (ic) is not satisfied, the least squares estimator is inadmissible.

c) Nothing is known about the uniqueness of  $\underline{t}^*$  as a minimax estimator for  $\underline{\theta}$  or even its admissibility when the conditions of theorem 2 are satisfied. According to Stein's [3] conjecture,  $\underline{t}^*$  may be shown to be inadmissible.

d) No simple way of verifying condition (ic) is known. However, under the following modification of the sampling scheme, this condition is always satisfied.

Suppose  $(y, x_1, \dots, x_p)$  follows a  $(p+1)$ -variate distribution. Let us choose and fix a constant  $c$ , however large. We then call the independent observations  $(y_i, x_{1i}, \dots, x_{pi})$ ,  $i=1, \dots, n \geq p+1$ , on  $(y, x_1, \dots, x_p)$  having risk of order  $c$  or less if and only if  $\|(X'X)^{-1}\| \leq c$ . Then our sampling scheme is as follows:

Sampling Scheme. Choose and fix a positive constant  $c$ . Make  $n$  independent observations on  $(y, x_1, \dots, x_p)$ . If the  $x$ -observations have risk of order  $c$  or less, stop sampling; if not, reject the observations and repeat the procedure till a set of observations having risk of order  $c$  or less is obtained, which is called the set of effective observations up to a risk of order  $c$ .

For any  $c$ , the effective observations up to a risk of order  $c$  can be considered to be observations on a process  $(y_i', x_{1i}', \dots, x_{pi}')$ ,  $i=1, \dots, n$  for which  $y_1', \dots, y_n'$  given  $x_{1i}', \dots, x_{pi}'$ ,  $i=1, \dots, n$ , are mutually independent, the regression function of  $y_1'$  on  $x_1', \dots, x_p'$  and the conditional variance of  $y_1'$  given  $x_1', \dots, x_p'$  are the same as those for  $(y, x_1, \dots, x_p)$ , while the marginal distribution of  $(x_1', \dots, x_p')$  satisfies condition (ic). Hence if  $(y, x_1, \dots, x_p)$  satisfy condition (ii), then  $(y', x_1', \dots, x_p')$  also satisfies condition (ii), and similarly for condition (iii) or (iv) or (v). It can also be noticed that we

have never made use of the independence of  $(x_{1i}, \dots, x_{pi})$ ,  $i=1, \dots, n$  in course of our analysis; all that we required was the independence of  $(y_1, \dots, y_n)$  given  $(x_{1i}, \dots, x_{pi})$ ,  $i=1, \dots, n$  and this property is preserved in the process  $(y'_i, x'_{1i}, \dots, x'_{pi})$ ,  $i=1, \dots, n$ . Thus we see that under conditions (ia), (ib), (ii) and (iii), the least squares estimator  $\underline{t}^*$  obtained from a set of effective observations up to a risk of some order  $c$  belongs to  $\mathcal{C}_1$ , and is the unique estimator for  $\underline{\theta}$  having uniformly minimum risk among all members of  $\mathcal{C}_1$  obtained from the same set of observations. Also, under conditions (ia), (ib), (ii), (iv) and (v), the above estimator is minimax for  $\underline{\theta}$  in the class of all estimators obtained from the same set of observations. Under this sampling scheme, the sample size becomes a random variable with expectation greater than  $n$  but since  $(X'X)^{-1}$  exists with probability 1 by virtue of condition (ia), the increase in the expected sample size over  $n$  can be made arbitrarily small by taking  $c$  sufficiently large. Also, since no observation on  $y$  is necessary in order to decide about the effectiveness of a set of  $(y_i, x_{1i}, \dots, x_{pi})$ ,  $i=1, \dots, n$ , and in many practical situations, an observation on  $(x_1, \dots, x_p)$  is much less costly than the observation on the associated  $y$ , the real increase in the cost of observations due to the introduction of the above modification in the sampling scheme may be much less than what appears at first sight.

The author wishes to thank Professor Wassily Hoeffding and Professor S. N. Roy for making some helpful suggestions.

#### REFERENCES

- [1] Lehmann, E. L. and Hodges, J. L., Jr. (1950). Ann. Math. Stat., Vol. 21, p.182.
- [2] Plackett, R. L. (1950). Biometrika, Vol 36, p. 458.
- [3] Stein, Charles (1956). Proc. Third Berk. Symp. on Math. Stat. and Prob., Vol. I, p. 197.