SOME PROPERTIES OF THE LEAST SQUARES ESTIMATOR IN REGRESSION ANALYSIS WHEN THE PREDICTOR VARIABLES ARE STOCHASTIC[1]

By P. K. Bhattacharya

University of North Carolina

1. Introduction and summary. In the classical linear estimation set-up, we have

$$E\mathbf{y} = X\boldsymbol{\theta}, \tag{1}$$

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

(I) Each component $\hat{\theta}_j$ of $\hat{\boldsymbol{\theta}}$ is the estimator with uniformly minimum variance among all unbiased linear estimators of the corresponding component θ_j of $\boldsymbol{\theta}$.

It can also be easily seen that

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

Properties (I) and (II) hold for model (1) with uncorrelated and homoscedastic $y_1, \cdots, y_n$. Hodges and Lehmann [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 $\boldsymbol{\theta}$ by another $\boldsymbol{\beta}$ is $(\boldsymbol{\theta} - \boldsymbol{\beta})'(\boldsymbol{\theta} - \boldsymbol{\beta})$, then the least squares estimator is minimax among the class of all estimators of $\boldsymbol{\theta}$ if there exists a number v such that Var $y_i \leqq v$, $i = 1, \cdots, n$, and the family of distributions $\mathcal{F}$ of $(y_1, \cdots, y_n)$ contains the sub-family $\mathcal{F}_0$ of all independent normal distributions of $(y_1, \cdots, y_n)$ which satisfy (1) for some $\boldsymbol{\theta}$, and have Var $y_i = v$, $i = 1, \cdots, n$.

In many situations, however, $(y, x_1, \cdots, x_p)$ 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 $x_1, \cdots, x_p$, regarding the x-observations to be non-stochastic. This problem differs from the classical problem of linear estimation because instead of (1) the model is $E[\mathbf{y} \mid X] = X\boldsymbol{\theta}$, where the elements of the X matrix are stochastic.

For reasons given in Section 2, the loss in estimating the true regression function $\phi(x_1, \cdots, x_p)$ by another function $\psi(x_1, \cdots, x_p)$ is considered to be of

Received August 9, 1961; revised May 5, 1962.

[1] This research was supported by the Air Force Office of Scientific Research.

the form,

$$\int [\phi(x_1, \cdots, x_p) - \psi(x_1, \cdots, x_p)]^2 \, dF(x_1, \cdots, x_p),$$

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

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 $F(x_1, \cdots, x_p)$ is obtained for this class to be non-empty, which unfortunately is not easy to verify in particular cases. 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.

For the case of multivariate normal distribution of $(y, x_1, \cdots, x_p)$, Stein [3] has considered this problem under a loss function similar to the one given above. He has shown the minimax property of the least squares estimates (which also happen to be the maximum likelihood estimates in a multivariate normal model) for the regression coefficients, and has raised many interesting questions about the admissibility of these estimates.

2. Formal statement of the problem. Let $y, x_1, \cdots, x_p$ be real-valued random variables with joint distribution function G. We assume for simplicity that G is the product of a completely specified distribution F of $x_1, \cdots, x_p$ with an unknown conditional distribution of y given $x_1, \cdots, x_p$. Let G belong to a family of distributions $\mathcal{G}$. We further assume that F and $\mathcal{G}$ satisfy the following conditions:

CONDITION (i). F is such that

(a)[2] for every non-null $(a_0, a_1, \cdots, a_p)$,

$$\Pr [a_0 + a_1x_1 + \cdots + a_px_p = 0] = 0.$$

(b) $E(x_j^2) < \infty, j = 1, \cdots, p$.

CONDITION (iia) For every $G \in \mathcal{G}$, $E_G[y \mid x_1, \cdots, x_p]$ is a linear function of $x_1, \cdots, x_p$, say $\phi(x_1, \cdots, x_p) = \theta_0 + \theta_1x_1 + \cdots + \theta_px_p$, the row vector $\theta' = (\theta_0, \theta_1, \cdots, \theta_p)$ depending on G.

CONDITION (iib) The set of all θ corresponding to all $G \in \mathcal{G}$, equals the $(p + 1)$-dimensional Euclidean space.

CONDITION (iii) There exists a constant σ^2 such that $V_G[y \mid x_1, \cdots, x_p] = \sigma^2$ for all $G \in \mathcal{G}$ and for all $(x_1, \cdots, x_p)$.

In the statement of the above conditions as well as in what follows, $E_G[y \mid x_1, \cdots, x_p]$ and $V_G[y \mid x_1, \cdots, x_p]$ stand for the conditional expectation and the conditional variance respectively, of y given $x_1, \cdots, x_p$ under G.

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

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

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

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

$$\begin{aligned} W(\boldsymbol{\theta}, \boldsymbol{\beta}) &= \int [(\theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p) \\ &\qquad - (\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)]^2 \, dF(x_1, \cdots, x_p) \\ &= \sum_{j=0}^{p} \sum_{j'=0}^{p} \mu_{jj'}(\theta_j - \beta_j)(\theta_{j'} - \beta_{j'}) \\ &= (\boldsymbol{\theta}' - \boldsymbol{\beta}')M(\boldsymbol{\theta} - \boldsymbol{\beta}), \end{aligned} \tag{2}$$

where

$$M = \begin{bmatrix} 1 & \mu_{01} & \cdots & \mu_{0p} \\ \mu_{01} & \mu_{11} & \cdots & \mu_{1p} \\ & \cdots & & \\ \mu_{0p} & \mu_{1p} & \cdots & \mu_{pp} \end{bmatrix}, \mu_{jj'} = E(x_j x_{j'}), j, j' = 0, 1, \cdots, p, x_0 \equiv 1.$$

It follows from condition (i) that M is positive definite, and $0 < W(\boldsymbol{\theta}, \boldsymbol{\beta}) < \infty$ for all $\boldsymbol{\beta} \neq \boldsymbol{\theta}$.

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

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

were known, our prediction rule would be

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

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

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

then the risk is $\sigma^2 + W(\boldsymbol{\theta}, \boldsymbol{\beta})$.

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 $\mathbf{t} \in \mathcal{C}_1$ if and only if

$$\underset{(p+1)\times 1}{\mathbf{t}} = \underset{(p+1)\times n}{\mathbf{L}} \; \underset{n\times 1}{\mathbf{y}} \tag{3}$$

and $\rho(\boldsymbol{\theta}, \mathbf{t}) = E[W(\boldsymbol{\theta}, \mathbf{t})]$ is a bounded function of $\theta_0, \theta_1, \cdots, \theta_p$, where $\mathbf{y}' =$

$(y_1, \cdots, y_n)$, and each element l_{ji} of L is a function of $x_{11}, \cdots, x_{1n}, \cdots, x_{p1}, \cdots, x_{pn}$. Let

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

Then the estimator obtained by the method of least squares is, $\mathbf{t}^* = (X'X)^{-1}X'\mathbf{y}$. Since $n \geqq p + 1$, it follows from condition (ia) that the matrix X has rank $p + 1$ with probability 1 and therefore $\mathbf{t}^*$ can be uniquely determined for almost all samples.

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

We shall first show that if an estimator $\mathbf{t} \varepsilon \mathcal{C}_1$, then for almost all X, the conditional expectation of each component of $\mathbf{t}$ given X should be equal to the corresponding component of $\boldsymbol{\theta}$, whatever $\boldsymbol{\theta}$ may be. For any $\mathbf{t} \varepsilon \mathcal{C}_1$, it follows from (3) that

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

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

$$E_{\mathbf{y}|X}(\mathbf{t}) - \boldsymbol{\theta} = H_{\mathbf{t}}\boldsymbol{\theta}. \tag{4}$$

We now prove

LEMMA 1. *If* $\mathbf{t} \varepsilon \mathcal{C}_1$, *then* $E_{\mathbf{y}|X}(\mathbf{t}) \equiv \boldsymbol{\theta}$ *for almost all* X.

PROOF. By virtue of (4) it will be enough to show that if $P[H_{\mathbf{t}} \neq 0] > 0$, then $\mathbf{t} \varepsilon \mathcal{C}_1$. Let $\mathbf{t}$ satisfy (3); then we shall show that $\rho(\boldsymbol{\theta}, \mathbf{t})$ is unbounded. We have

$$\begin{aligned} \rho(\boldsymbol{\theta}, \mathbf{t}) &\geqq E_X[(E_{\mathbf{y}|X}(\mathbf{t}') - \boldsymbol{\theta}')M(E_{\mathbf{y}|X}(\mathbf{t}) - \boldsymbol{\theta})] \\ &= E_X[\boldsymbol{\theta}'H_{\mathbf{t}}'MH_{\mathbf{t}}\boldsymbol{\theta}]. \end{aligned}$$

If $P[H_{\mathbf{t}} \neq 0] > 0$, some element of $H_{\mathbf{t}}$ is non-zero with positive probability. Let that element be one in the j_0th column of $H_{\mathbf{t}}$. Let

$$A_{j_0} = \{\boldsymbol{\theta} : \theta_j = 0 \quad \text{for} \quad j \neq j_0 \quad \text{and} \quad \theta_{j_0} \neq 0\}.$$

Then for $\boldsymbol{\theta} \varepsilon A_{j_0}$,

$$\boldsymbol{\theta}'H_{\mathbf{t}}'MH_{\mathbf{t}}\boldsymbol{\theta} = \theta_{j_0}^2 g(X)$$

where $P[g(X) \geqq 0] = 1$ 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 $\boldsymbol{\theta} \varepsilon A_{j_0}$,

$$\rho(\boldsymbol{\theta}, \mathbf{t}) \geqq E_X[\boldsymbol{\theta}'H_{\mathbf{t}}'MH_{\mathbf{t}}\boldsymbol{\theta}] \geqq \theta_{j_0}^2 \delta P(\delta),$$

and for any given c, $\rho(\theta, \mathbf{t})$ can be made greater than c by choosing θ in A_{j_0} with $|\theta_{j_0}| > [c/\delta P(\delta)]^{\frac{1}{2}}$. This completes the proof.

The following corollary is immediate.

COROLLARY. *If* $\mathbf{t} \in \mathcal{C}_1$, *then*

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

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

$$\sum_{i=1}^{n} l_{ji} x_{ji} = 1, \qquad \sum_{i=1}^{n} l_{ji} x_{j'i} = 0 \quad \text{for} \quad j \neq j' = 0, 1, \cdots, p, \tag{6}$$

for almost all X.

It follows from the above corollary that for $\mathbf{t} \in \mathcal{C}_1$, $\rho(\theta, \mathbf{t})$ does not depend on θ. Therefore, if we minimize the right side of (5) with respect to l_{ji}'s subject to the conditions in (6), the resulting matrix $\hat{L} = (X'X)^{-1}X'$ will define an estimator $\hat{\mathbf{t}} = \hat{L}\mathbf{y} = (X'X)^{-1}X'\mathbf{y}$ for which $\rho(\theta, \hat{\mathbf{t}}) \leqq \rho(\theta, \mathbf{t})$ for all θ and for arbitrary $\mathbf{t} \in \mathcal{C}_1$. But $\hat{\mathbf{t}} = \mathbf{t}^*$, a.e.

It can be easily seen that unless $\mathcal{C}_1$ is empty, $\mathbf{t}^* \in \mathcal{C}_1$. We thus have

THEOREM 1. *If* $\mathcal{C}_1$ *is non-empty, then the least squares estimator* $\mathbf{t}^* \in \mathcal{C}_1$, *and* $\rho(\theta, \mathbf{t}^*) \leqq \rho(\theta, \mathbf{t})$ *for all* θ *and for arbitrary* $\mathbf{t} \in \mathcal{C}_1$ *with a strict inequality holding if* $P[\mathbf{t} = \mathbf{t}^*] < 1$.

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 $\rho(\theta, \mathbf{t}^*)$ is a constant for all θ.

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

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

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

Hence a necessary and sufficient condition for $\mathcal{C}_1$ being non-empty is

CONDITION (ic). $E \operatorname{tr} [(X'X)^{-1}M] < \infty$.

This is a condition on the marginal distribution F of $(x_1, \cdots, x_p)$, and it implies condition (ib). Thus under conditions (ia), (ic), (iia), (iib) and (iii), we can state Theorem 1 and its corollary without the qualifying clause "if $\mathcal{C}_1$ is non-empty".

4. Minimax property of t^* in the class of all estimators. In this section we assume F to satisfy conditions (ia) and (ic) and the family $\mathcal{G}$ to satisfy conditions (iia) and

CONDITION (iv). There exists a number v such that $V_G[y \mid x_1, \cdots, x_p] \leqq v$ for all $G \in \mathcal{G}$ and for all $x_1, \cdots, x_p$.

CONDITION (v). $\mathcal{G}$ includes the class $\mathcal{G}_0$ of all G obtained by taking the product of the distribution F of $(x_1, \cdots, x_p)$ with a conditional distribution of y given

$x_1, \cdots, x_p$ which is normal with mean satisfying (iia) for some θ and with variance v.

Under these conditions we shall prove that $\mathbf{t}^*$ is a minimax estimate for θ. 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 uniformly bounded. Let U be a fixed $k \times k$ matrix and $\{b_m\}$ a sequence of real numbers converging to zero. Then,*

(a) Det $[A(z) + b_m U]$ *converges to* Det $A(z)$ *uniformly in z.*

(b) *There exists an integer m_0 such that $[A(z) + b_m U]^{-1}$ exists for $m \geqq m_0$ and for all $z \varepsilon Z$.*

(c) $\|[A(z) + b_m U]^{-1} - A(z)^{-1}\|$ *converges to zero uniformly in z, where $[A(z) + b_m U]^{-1}$ is defined arbitrarily when* Det $[A(z) + b_m U] = 0$.

PROOF. Suppose max $[\mathrm{Sup}_{z \varepsilon Z} \|A(z)\|, \mathrm{Sup}_{z \varepsilon Z} \|A(z)^{-1}\|, \|U\|] = c$. Then (a) follows from the fact that

$$|\text{Det } [A(z) + b_m U] - \text{Det } A(z)| \leqq |b_m| c^k (2^k - 1) \cdot k!, \quad \text{if} \quad |b_m| \leqq 1.$$

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

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

for all $z \varepsilon Z$. Hence for $m \geqq m_0$ and for all $z \varepsilon Z$,

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

and therefore (b) follows.

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

THEOREM 2. *Under conditions* (ia), (ic), (iia), (iv) *and* (v), *the least squares estimator $\mathbf{t}^*$ is minimax for θ in the class of all estimators.*

PROOF. We shall first prove that

$$\mathrm{Sup}_{G \varepsilon \mathcal{G}_0} r(G, \mathbf{t}^*) \leqq \mathrm{Sup}_{G \varepsilon \mathcal{G}_0} r(G, \mathbf{t}) \qquad \text{for all } \mathbf{t}. \tag{7}$$

Since there is a one-one correspondence between $\mathcal{G}_0$ and the $(p + 1)$-dimensional Euclidean space (in which θ takes its values), we can write $\rho(\theta, \mathbf{t})$ instead of $r(G, \mathbf{t})$ for each $\mathbf{t}$ where θ corresponds to G. Also since for each θ, $\rho(\theta, \mathbf{t}) = vE \text{ tr } [(X'X)^{-1}M]$, it will be enough to show that for all $\mathbf{t}$,

$$\mathrm{Sup}_\theta \rho(\theta, \mathbf{t}) \geqq vE \text{ tr } [(X'X)^{-1}M].$$

Suppose there exists $\hat{\mathbf{t}}$ such that

$$\mathrm{Sup}_\theta \rho(\theta, \hat{\mathbf{t}}) = vE \text{ tr } [(X'X)^{-1}M] - \epsilon, \qquad \epsilon > 0.$$

We shall contradict this by showing that for any given $\epsilon > 0$, we can make

$$\rho(\xi_m, \hat{\mathbf{t}}) > vE \text{ tr } [(X'X)^{-1}M] - \epsilon \tag{8}$$

by choosing m sufficiently large, where $\{\xi_m\}$ is a sequence of *a priori* distributions of θ defined as follows

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

Choose and fix $\epsilon > 0$. Since $E \operatorname{tr}\left[(X'X)^{-1}M\right]$ exists, we can find a constant $c(\epsilon)$ such that

$$\left| \int_{X \epsilon R(\epsilon)} \operatorname{tr}[(X'X)^{-1}M] \prod_{i=1}^{n} dF(x_{1i}, \cdots, x_{pi}) - E \operatorname{tr}\left[(X'X)^{-1}M\right] \right| < \epsilon/2v,$$

where $R(\epsilon)$ is the set $\{X : \|(X'X)\| \leqq c(\epsilon), \|(X'X)^{-1}\| \leqq c(\epsilon)\}$. Now,

$$\rho(\xi_m, \hat{t}) = \int_{\Theta} \left[\int_X \int_y (\theta - \hat{t})'M(\theta - \hat{t}) g_\theta(y \mid X) \cdot \prod_{i=1}^{n} dy_i \prod_{i=1}^{n} dF(x_{1i}, \cdots, x_{pi}) \right] d\xi_m(\theta)$$

where $g_\theta(y \mid X) = \text{const.} \exp\left[-(1/2v)(y - X\theta)'(y - X\theta)\right]$.

Since the integral in the above expression is non-negative, we get

$$\rho(\xi_m, \hat{t}) \geqq \int_{X \epsilon R(\epsilon)} \int_y \left[\int_\Theta (\theta - \hat{t})'M(\theta - t) g_\theta(y \mid X)\, d\xi_m(\theta) \right] \cdot \prod_{i=1}^{n} dy_i \prod_{i=1}^{n} dF(x_{1i}, \cdots, x_{pi}).$$

Again if $(X'X + (v/m)I)^{-1}$ exists, $g_\theta(y \mid X)\, d\xi_m(\theta)$ can be written as

$$\text{const.} f_{1m}(\theta \mid X, y) f_{2m}(X, y) \prod_{j=0}^{p} d\theta_j ,$$

where

$$f_{1m}(\theta \mid X, y) = \exp\left[-1/(2v)\{\theta - (X'X + (v/m)I)^{-1}X'y\}'(X'X + (v/m)I) \cdot \{\theta - (X'X + (v/m)I)^{-1}X'y\}\right]$$

and

$$f_{2m}(X, y) = \exp\left[-(1/2v)y'y + (1/2v)y'X(X'X + (v/m)I)^{-1}X'y\right].$$

But $(X'X + (v/m)I)^{-1}$ exists for all m greater than or equal to some integer m_0, and for all $X \varepsilon R(\epsilon)$, since the hypotheses of Lemma 2 are ensured for $X \varepsilon R(\epsilon)$. Hence we have

$$\rho(\xi_m, \hat{t}) \geqq \text{const.} \int_{X \epsilon R(\epsilon)} \int_y \left[\int_\Theta (\theta - \hat{t})'M(\theta - \hat{t}) f_{1m}(\theta \mid X, y) \cdot \prod_{j=0}^{p} d\theta_j \right] f_{2m}(X, y) \prod_{i=1}^{n} dy_i \prod_{i=1}^{n} dF(x_{1i}, \cdots, x_{pi})$$

for $m \geqq m_0$. Now since M is positive definite, for values of $m \geqq m_0$ and for any given $X \varepsilon R(\epsilon)$ and $\mathbf{y}$,

$$\int_{\Theta} (\boldsymbol{\theta} - \hat{\mathbf{t}})' M (\boldsymbol{\theta} - \hat{\mathbf{t}}) f_{1m}(\boldsymbol{\theta} \mid X, \mathbf{y}) \prod_{j=0}^{p} d\theta_j$$

$$\geqq \int_{\Theta} \{\boldsymbol{\theta} - (X'X + (v/m)I)^{-1} X'\mathbf{y}\}'$$

$$\cdot M \{\boldsymbol{\theta} - (X'X + (v/m)I)^{-1} X'\mathbf{y}\} f_{1m}(\boldsymbol{\theta} \mid X, \mathbf{y}) \prod_{j=0}^{p} d\theta_j .$$

Hence,

$$\rho(\xi_m, \hat{\mathbf{t}}) \geqq \text{const.} \int_{X \varepsilon R(\epsilon)} \int_{\mathbf{y}} \left[\int_{\Theta} \{\boldsymbol{\theta} - (X'X + (v/m)I)^{-1} X'\mathbf{y}\}' \right.$$

$$\left. \cdot M \{\boldsymbol{\theta} - (X'X + (v/m)I)^{-1} X'\mathbf{y}\} f_{1m}(\boldsymbol{\theta} \mid X, \mathbf{y}) \prod_{j=0}^{p} d\theta_j \right]$$

$$\cdot f_{2m}(X, \mathbf{y}) \prod_{i=1}^{n} dy_i \prod_{i=1}^{n} dF(x_{1i}, \cdots, x_{pi}).$$

After some simplifications, the right side of the last inequality reduces to

$$v \int_{X \varepsilon R(\epsilon)} \text{tr}\, [(X'X + (v/m)I)^{-1} M] \prod_{i=1}^{n} dF(x_{1i}, \cdots, x_{pi}) = K_m(\epsilon) \text{ say.}$$

It follows from Lemma 2 that

$$\lim_{m \to \infty} K_m(\epsilon) = v \int_{X \varepsilon R(\epsilon)} \text{tr}\, [(X'X)^{-1} M] \prod_{i=1}^{n} dF(x_{1i}, \cdots, x_{pi}).$$

Therefore, for sufficiently large m,

$$K_m(\epsilon) > v \int_{X \varepsilon R(\epsilon)} \text{tr}\, [(X'X)^{-1} M] \prod_{i=1}^{n} dF(x_{1i}, \cdots, x_{pi}) - \epsilon/2$$

$$> vE \text{ tr}\, [(X'X)^{-1} M] - \epsilon.$$

Hence, for such large values of m, (8) holds. This proves (7). To complete the proof of the theorem, we have only to note that since $\mathcal{G}_0 \subset \mathcal{G}$,

$$\text{Sup}_{G \varepsilon \mathcal{G}_0} r(G, \mathbf{t}) \leqq \text{Sup}_{G \varepsilon \mathcal{G}} r(G, \mathbf{t})$$

for arbitrary $\mathbf{t}$, whereas

$$\text{Sup}_{G \varepsilon \mathcal{G}_0} r(G, \mathbf{t}^*) = vE \text{ tr}\, [(X'X)^{-1} M] = \text{Sup}_{G \varepsilon \mathcal{G}} r(G, \mathbf{t}^*).$$

5. Remarks.

(a) The results in the previous sections are proved under the assumption that the marginal distribution F of $(x_1, \cdots, x_p)$ is completely specified. Now suppose F belongs to a family $\mathfrak{F}$ of p-variate distribution functions. Then the family

$\mathcal{G}$ of distributions of $(y, x_1, \cdots, x_p)$ can be expressed as the union of disjoint sub-families $\mathcal{G}(F)$, $F \varepsilon \mathcal{F}$, where each $G \varepsilon \mathcal{G}(F)$ has F as the marginal distribution of $(x_1, \cdots, x_p)$. It can be easily seen that if conditions (ia) and (ic) are satisfied by each $F \varepsilon \mathcal{F}$ and if conditions (iia), (iib) and (iii) are satisfied by $\mathcal{G}$, then Theorem 1 and its corollary hold. Also, Theorem 2 remains valid if condition (ic) is replaced by

CONDITION (id). $\text{Sup}_{F \varepsilon \mathcal{F}} E_F \text{ tr } [(X'X)^{-1} M_F] < \infty$, where M_F is the matrix defined in Section 2 in which $\mu_{jj'} = E_F(x_j x_{j'})$, and if condition (v) holds for each $\mathcal{G}(F)$, $F \varepsilon \mathcal{F}$.

(b) Condition (ic) is not satisfied in general and no simple way of verifying this condition is known. When this condition is not satisfied, $\mathbf{t}^*$ is not only inadmissible but is the worst possible estimate. In such cases, or even when the condition cannot be verified due to difficulties in analysis, it is very dangerous to use the method of least squares. It is not known whether there exists an estimator of $\boldsymbol{\theta}$ with bounded risk when condition (ic) is not satisfied. Even if such an estimator exists, it has to be non-linear in y's.

(c) Under the following sequential modification of the sampling scheme, condition (ic) is always satisfied. The procedure given below is very crude but if the loss function is the sum of two components, one given by (2) and another proportional to the sample size, it has a bounded risk function.

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

Sampling Scheme. Choose and fix a positive constant c. Make n independent observations on $(y, x_1, \cdots, x_p)$. If the x-observations have risk of order c, stop sampling; if not, reject the observations and repeat the procedure till a set of observations having risk of order c 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 $(y_i', x_{1i}', \cdots, x_{pi}')$, $i = 1, \cdots, n$ for which $y_1', \cdots, y_n'$ given $x_{1i}', \cdots, x_{pi}'$, $i = 1, \cdots, n$, are mutually independent, the regression function of y' on $x_1', \cdots, x_p'$ and the conditional variance of y' given $x_1', \cdots, x_p'$ are the same as those for $(y, x_1, \cdots, x_p)$, while the marginal distribution of $(x_1', \cdots, x_p')$ satisfies condition (ic). Hence if $(y, x_1, \cdots, x_p)$ satisfy conditions (iia) and (iib), then $(y', x_1', \cdots, x_p')$ also satisfies conditions (iia) and (iib), and similarly for condition (iii) or (iv) or (v). It can also be noticed that we have never made use of the independence of $(x_{1i}, \cdots, x_{pi})$, $i = 1, \cdots, n$ in the course of our analysis; all that we required was the independence of $(y_1, \cdots, y_n)$ given $(x_{1i}, \cdots, x_{pi})$, $i = 1, \cdots, n$ and this property is preserved in the process $(y_i', x_{1i}', \cdots, x_{pi}')$, $i = 1, \cdots, n$. Thus we see that under conditions (ia), (ib), (iia), (iib) and (iii), the least squares estimator $\mathbf{t}^*$ 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 $\boldsymbol{\theta}$ having uniformly minimum risk among

all members of $\mathcal{C}_1$ obtained from the same set of observations. Also, under conditions (ia), (ib), (iia), (iv) and (v), the above estimator is minimax for θ 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.

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

REFERENCES

[1] HODGES, J. L., JR. and LEHMANN, E. L. (1950). Some problems in minimax point estimation. *Ann. Math. Statist.* **21** 182–197.

[2] PLACKETT, R. L. (1949). A historical note on the method of least squares. *Biometrika* **36** 458–460. Stanford University Press. Stanford, Calif.

[3] STEIN, CHARLES (1960). Multiple regression. *Contributions to Probability and Statistics (Essays in Honor of Harold Hotelling)*. 424–443.