SOME REMARKS ON NORMAL MULTIVARIATE REGRESSION*

STANLEY L. SCLOVE

(Received March 25, 1969)

Introduction

Stein [5] proved the minimax property of the maximum likelihood estimator for univariate normal multiple regression. The present paper is essentially an extension of this result to the case of multivariate normal multiple regression. There is some consideration of the prediction problem associated with regression, and an interesting modification of the usual prediction procedure is discussed.

1. Preliminaries and notation

Let $Z_1, \cdots, Z_N$ be independent random vectors, each distributed according to $\mathcal{N}(\mu, \Sigma)$, Σ being nonsingular. Let

$$Z_k=\begin{pmatrix} Y_k \\ X_k \end{pmatrix}, \quad \mu=\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \quad \text{and} \quad \Sigma=\begin{pmatrix} \Sigma_Y & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_X \end{pmatrix},$$

where Y_k and μ_Y are $q\times 1$, X_k and μ_X are $p\times 1$, Σ_Y is $q\times q$, Σ_X is $p\times p$, Σ_{YX} is $q\times p$, and $\Sigma_{XY}=\Sigma'_{YX}$. The conditional distribution of Y_k given $X_k=x$ is $\mathcal{N}(\alpha+\beta' x, \Sigma_{Y\cdot X})$, where $\Sigma_{Y\cdot X}=\Sigma_Y-\Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$, $\beta=\Sigma_X^{-1}\Sigma_{XY}$, and $\alpha=\mu_Y-\beta'\mu_X$. We consider the problem of estimating the pair (α, β) when the loss function is

$$\begin{aligned} L_1(\hat\alpha, \hat\beta; \mu, \Sigma)=&[(\hat\alpha-\alpha)+(\hat\beta-\beta)'\mu_X]'\Sigma_{Y\cdot X}^{-1}[(\hat\alpha-\alpha)+(\hat\beta-\beta)'\mu_X] \\ &+\operatorname{tr}\Sigma_X(\hat\beta-\beta)\Sigma_{Y\cdot X}^{-1}(\hat\beta-\beta)' . \end{aligned} \tag{1.1}$$

This loss function has the following interpretation. Suppose that we have a new independent observation X_0 and wish to predict the corresponding Y_0 when the loss function is $l(\hat Y_0; Y_0, \Sigma)=(\hat Y_0-Y_0)'\Sigma_{Y\cdot X}^{-1}(\hat Y_0-Y_0)$. By a computation similar to formula (2.9) of [5] (details are in [4]) one demonstrates

Fact 1. Suppose the prediction function $\hat Y_0$ has the form $\hat Y_0=\hat\alpha+$

* This work was supported in part by Office of Naval Research Contract Nonr-4259(08), NR-042-034 at Columbia University and Nonr-225(52), NR-342-022 at Stanford University.

$\hat{\beta}'X_0$, where $\hat{\alpha}$ and $\hat{\beta}$ are functions of $Z_1, \cdots, Z_N$. Then

$$E_{\mu,\Sigma}[l(\hat{Y}_0; Y_0, \Sigma) | Z_1, \cdots, Z_N] = L_1(\hat{\alpha}, \hat{\beta}; \mu, \Sigma) + q\,.$$

Let

$$\bar{Y} = \sum_{k=1}^{N} Y_k/N\,, \qquad \bar{X} = \sum_{k=1}^{N} X_k/N\,, \qquad S_Y = \sum_{k=1}^{N} (Y_k - \bar{Y})(Y_k - \bar{Y})'/N\,,$$

$$S_{XY} = \sum_{k=1}^{N} (X_k - \bar{X})(Y_k - \bar{Y})'/N\,, \qquad S_X = \sum_{k=1}^{N} (X_k - \bar{X})(X_k - \bar{X})'N$$

$$\text{and} \qquad S_{YX} = S'_{XY}\,.$$

We assume $N \geqq p+1$. The maximum likelihood estimators for β, α, Σ and $\Sigma_{Y \cdot X}$ are, respectively, $B = S_X^{-1} S_{XY}$, $a = \bar{Y} - B'\bar{X}$, $S = \begin{pmatrix} S_Y & S_{YX} \\ S_{XY} & S_X \end{pmatrix}$ and $S_{Y \cdot X} = S_Y - S_{YX} S_X^{-1} S_{XY}$. A sufficient statistic is $T = (a, B, S_{Y \cdot X}, \bar{X}, S_X)$.

In the sequel we employ some invariance notions. We consider groups $G = (G_1, G_2, G_3)$, where g_1 in G_1 operates on the sample space, g_2 in G_2 operates on the parameter space, and g_3 in G_3 operates on the space of estimates. Hence g in G has the form $g = (g_1, g_2, g_3)$. Since G_2 and G_3 will be taken to be groups induced by G_1, we shall in the sequel describe groups G by giving a typical element of G_1. If G is a group of transformations and $\mathscr{D}$ a given class of estimators, we denote by $\mathscr{D}_1(G, \mathscr{D})$ the class of estimators in $\mathscr{D}$ that are invariant under G.

2. Minimax property of the maximum likelihood estimator

The risk $R_1(a, B; \mu, \Sigma) = E_{\mu,\Sigma}[L_1(a, B; \mu, \Sigma)]$ can be computed as follows. Let

$$L_2(\hat{\beta}; \Sigma) = \operatorname{tr} \Sigma_X (\hat{\beta} - \beta) \Sigma_{Y \cdot X}^{-1} (\hat{\beta} - \beta)' \quad \text{and} \quad R_2(\hat{\beta}; \mu, \Sigma) = E_{\mu,\Sigma}[L_2(\hat{\beta}; \Sigma)]\,.$$

Straightforward computation (given in [4]) gives

Fact 2. If

$$\hat{\beta} = \hat{\beta}(S) \quad \text{and} \quad \hat{\alpha} = \bar{Y} - \hat{\beta}'\bar{X}\,, \tag{2.1}$$

then $R_1(\hat{\alpha}, \hat{\beta}; \mu, \Sigma) = (1 + 1/N) R_2(\hat{\beta}; \mu, \Sigma) + q/N$.

The maximum likelihood estimator (a, B) satisfies (2.1); hence

$$R_1(a, B; \mu, \Sigma) = \left(1 + \frac{1}{N}\right) R_2(B; \mu, \Sigma) + \frac{q}{N}\,. \tag{2.2}$$

The problem is invariant under the group described by the transformations

$$\begin{pmatrix} Y_k \\ X_k \end{pmatrix} \rightarrow \begin{pmatrix} KY_k+H'X_k+u \\ MX_k+v \end{pmatrix},$$

where $K(q\times q)$ and $M(p\times p)$ are nonsingular, H is $p\times q$, u is $q\times 1$, and v is $p\times 1$. The estimator B is invariant with respect to this group. Since the group operates transitively on the spaces of $\Sigma_{Y\cdot X}$ and Σ_X and the risk of an invariant estimator is constant on orbits, we can assume in the computation that $\Sigma_{Y\cdot X}=I_q$ and $\Sigma_X=I_p$. Letting E_0 denote the expectation at such parameter values, we have

$$R_2(B;\mu,\Sigma)=E_0[\mathrm{tr}\,(B-\beta)(B-\beta)']=\sum_{j=1}^{q}E_0[(b_j-\beta_j)'(b_j-\beta_j)]\,, \tag{2.3}$$

where $\beta=(\beta_1,\cdots,\beta_q)$ and $B=(b_1,\cdots,b_q)$. b_j is the maximum likelihood estimator of β_j. Again by invariance,

$$E_0[(b_j-\beta_j)'(b_j-\beta_j)]=E_{\mu,\Sigma}[(b_j-\beta_j)'\Sigma_X(b_j-\beta_j)/(\Sigma_{Y\cdot X})_{jj}]\,,$$

where $(\Sigma_{Y\cdot X})_{jj}$ is the (j,j)-th element of $\Sigma_{Y\cdot X}$. This expectation is the risk of b_j, which from [5] is seen to be $p/(N-p-2)$ if $N\geqq p+3$ and ∞ otherwise. Thus

$$R_2(B;\mu,\Sigma)=\begin{cases} \dfrac{qp}{N-p-2} & \text{if } N\geqq p+3 \\ \infty & \text{if } N\leqq p+2\,. \end{cases} \tag{2.4}$$

Using (2.2), we have

$$R_1(a,B;\mu,\Sigma)=\begin{cases} q\dfrac{N(p+1)-2}{N(N-p-2)} & \text{if } N\geqq p+3 \\ \infty & \text{if } N\leqq p+2\,. \end{cases} \tag{2.5}$$

Now, using Kiefer's general invariance theorem (Section 3 of [3]) and the Hunt-Stein method, we shall eventually prove the

THEOREM. *The pair* (a,B) *is minimax for the problem of estimating* (α,β) *when the loss is given by* (1.1).

The proof is similar to the one given for the univariate case in [5].

We have seen (2.5) that (a,B) has constant risk. Hence it suffices to prove that (a,B) is minimax over a subset of the parameter space. Choosing this subset so that $\Sigma_X=I_p$, $\Sigma_{Y\cdot X}=I_q$ and $\mu_X=0$, the loss function (1.1) becomes

$$(\hat\alpha-\alpha)'(\hat\alpha-\alpha)+\mathrm{tr}\,(\hat\beta-\beta)'(\hat\beta-\beta)\,. \tag{2.6}$$

Let $\Gamma=(\alpha,\beta')$; then (2.6) is $\mathrm{tr}\,(\hat\Gamma-\Gamma)'(\hat\Gamma-\Gamma)$, which is a strictly convex loss function. Hence

Fact 3. The class $\mathscr{D}^T$ of nonrandomized estimators based on the sufficient statistic T is complete.

Because of Fact 3, it suffices to show that (a, B) is minimax in $\mathscr{D}^T$.

LEMMA 1. *The only estimator in $\mathscr{D}^T$ that is invariant under the group G^* described by*

$$\begin{pmatrix} Y_k \\ X_k \end{pmatrix} \to \begin{pmatrix} KY_k + H'X_k + u \\ X_k \end{pmatrix},$$

where K is either I_q or $-I_q$, is (a, B).

PROOF OF LEMMA 1. (Omitted; formally the same as the proof for the univariate case in [5]. Details are in [4].)

By relating Theorem 8.6.1-4 of [2] to the problem of nonrandomized estimation in the natural way one obtains

Fact 4. Let G be a finite group of M elements which leaves invariant the problem of estimating a parameter ω. Suppose that the loss function is convex and the transformations g_3 are linear in the sense that

$$g_3(\hat{\omega}_1 + \hat{\omega}_2) = g_3(\hat{\omega}_1) + g_3(\hat{\omega}_2)\,. \tag{2.7}$$

Then given any estimator Ψ, the nonrandomized estimator

$$\Psi^*(x) = \sum_{g \in G} g_3^{-1}\Psi(g_1 x)/M \tag{2.8}$$

is invariant under G, and $\sup_{\omega} R(\Psi^*; \omega) \leqq \sup_{\omega} R(\Psi; \omega)$.

REMARK. Without condition (2.7), one could not always obtain a nonrandomized estimator Ψ^* that is invariant and statisfies (2.8).

Let $G^{(1)}$ be the group described by

$$\begin{pmatrix} Y_k \\ X_k \end{pmatrix} \to \begin{pmatrix} Y_k + H'X_k + u \\ X_k \end{pmatrix}.$$

Let $G^{(2)}$ be the two-element group $\{g^-, g^+\}$, where g^- is described by

$$g_1^- \begin{pmatrix} Y_k \\ X_k \end{pmatrix} = \begin{pmatrix} -Y_k \\ X_k \end{pmatrix}$$

and g^+ is the identity transformation. Then $G^* = G^{(2)} \circ G^{(1)} = \{g^{(2)} \circ g^{(1)} : g^{(i)}$ in $G^{(i)}$, $i = 1, 2\}$. Now we can obtain

LEMMA 2. *(a, B) is minimax in $\mathscr{D}_I(G^{(1)}, \mathscr{D}^T)$.*

PROOF OF LEMMA 2. Under $G^{(1)}$, $a \rightarrow a+u$, $B \rightarrow B+H$, and $\bar{X}$, S_X, and $S_{Y\cdot X}$ are invariant; or, letting $U=(a, B')$, $V=(\bar{X}, S_X, S_{Y\cdot X})$ and $F=(u, H')$, $U \rightarrow U+F$ and $V \rightarrow V$. We identify G and the group it induces on the space of U and V. Any g in $G^{(1)}$ operates trivially on the space of V: we write merely $g(U)$ instead of $g(U, V)$; furthermore we identify F and g and write $g(U)=U+g$ instead of $U+F$. The quotient group $G^*/G^{(1)}$ consists of the two cosets $C_-=\{g: g(U)=-U+g\}$, $C_+=\{g: g(u)=+U+g\}$. $G^{(1)}$ is a normal subgroup of G^*, and $G^*/G^{(1)}$ is isomorphic to $G^{(2)}$, the isomorphism being $C_+ \leftrightarrow g^+$, $C_- \leftrightarrow g^-$. Take G in Fact 4 to be $G^{(2)}$. Given any estimator Ψ in $\mathscr{D}_I(G^{(1)}, \mathscr{D}^T)$, the estimator Ψ^* obtained by averaging over the two-element group $G^{(2)}$ satisfies $\sup_{\mu, \Sigma} R(\Psi^*; \mu, \Sigma) \leqq \sup_{\mu, \Sigma} R(\Psi; \mu, \Sigma)$. Since Ψ is invariant under $G^{(1)}$, so is Ψ^*. By construction, Ψ^* is invariant under $G^{(2)}$. Hence Ψ^* is invariant under $G^*=G^{(2)} \circ G^{(1)}$. Thus any estimator that is minimax in $\mathscr{D}_I(G^*, \mathscr{D}^T)$ must be minimax in $\mathscr{D}_I(G^{(1)}, \mathscr{D}^T)$. But by Lemma 1, (a, B) is the *only* estimator in $\mathscr{D}_I(G^*, \mathscr{D}^T)$. Therefore (a, B) is minimax in $\mathscr{D}_I(G^{(1)}, \mathscr{D}^T)$.

Kiefer's theorem involves five assumptions which will be shown to hold for the group $G^{(1)}$.

LEMMA 3. (Kiefer's theorem). *Suppose that a group G leaves the problem invariant, that the five assumptions are satisfied, and that Ψ^* is minimax in $\mathscr{D}_I(G, \mathscr{D})$. Then Ψ^* is minimax in $\mathscr{D}$.*

PROOF OF THEOREM. We shall show that the five assumptions are satisfied for the problem at hand when $G=G^{(1)}$. Then the Theorem follows at once by taking $\mathscr{D}$ of Lemma 3 to be $\mathscr{D}^T$ and applying Fact 3 and Lemma 2.

To see that Assumption 1 is fullfilled, take $g_{U_1}(U_2)=U_2+U_1$, where U_1 and U_2 are arbitrary values of $U=(a, B')$. Then $g_{U_1}^{-1}(U_2)=U_2-U_1$. $G^{(1)}$ operates transitively on the space of U values: any two values are in the same orbit. Hence Assumption 1 becomes $g_{U_1}^{-1}(U_1)=g_{U_2}^{-1}(U_2)$ for any U_1, U_2. But for any U we have $g_U^{-1}(U)=U-U=0$, so this condition is met. Also, the estimator $\Psi(g, g_U^{-1}U)=\Psi(g+0)=\Psi(g)$ is in $\mathscr{D}^T$, as required.

Assumption 2 is satisfied because, letting $\Gamma=(\alpha, \beta')$, $g_{h_1U}^{-1}(h_3\hat{\Gamma})=h_3\hat{\Gamma}-h_1U=(\hat{\Gamma}+h)-(U+h)=\hat{\Gamma}-U=g_U^{-1}(\hat{\Gamma})$, for all h in G.

Assumption 3 is met because, by Lemma 2, (a, B) is minimax in $\mathscr{D}_I(G^{(1)}, \mathscr{D}^T)$.

Since g in $G^{(1)}$ maps (U, V) into $(U+F, V)$, where $F=(a, B')$ is a $q \times (p+1)$ matrix, $G^{(1)}$ is isomorphic to the additive group of a real linear space of dimension $q(p+1)$. The loss function is for each fixed estimate $(\hat{\alpha}, \hat{\beta})$ bounded on bounded sets in (α, β)-space and becomes infinite as the components of (α, β) go to infinity. Therefore Condition 4b of [3] for Assumption 4 holds.

In verifying Assumption 5, we again use the isomorphism with the additive group of a real linear space. Because of this, we can take μ of Assumption 5 to be Lebesgue measure and G_n to be the hypercube of side $2n$, centered at the origin, as in Condition 5b(1) of [3].

3. Extension

Consider the class of prediction functions of the form $\hat{Y}_0=\hat{\alpha}+\hat{\beta}'\hat{X}_0$ where $\hat{\alpha}=\hat{\alpha}(Z_1, \cdots, Z_N)$ and $\hat{\beta}=\hat{\beta}(Z_1, \cdots, Z_N)$. For such prediction functions we have Fact 1. Hence the risk of $\hat{Y}_0$ for predicting Y_0 is q more than $R_2(\hat{\alpha}, \hat{\beta}; \mu, \Sigma)$, and the problem of predicting Y_0 reduces to one of estimating (α, β). Hence the prediction function $a+B'X_0$ is minimax among prediction functions of the above form.

4. An example

Now let $\tilde{Y}_0=a+B'X_0=\bar{Y}+B'(X_0-\bar{X})$. Intuitively it seems that we might do better by replacing $\bar{X}$ by the updated mean

$$\bar{X}^*=\bar{X}+\frac{1}{N+1}(X_0-\bar{X}) .$$

Since

$$X_0-\bar{X}^*=X_0-\bar{X}-\frac{1}{N+1}(X_0-\bar{X})=\frac{N}{N+1}(X_0-\bar{X}) ,$$

the resulting prediction function $\tilde{\tilde{Y}}_0$ is

$$\tilde{\tilde{Y}}_0=\bar{Y}+B'(X_0-\bar{X}^*)=\bar{Y}+\frac{N}{N+1}B'(X_0-\bar{X}) .$$

Combining Facts 1 and 2, one obtains

Fact 5. If $\hat{Y}_0=\bar{Y}+\hat{\beta}(X_0-\bar{X})$, where $\hat{\beta}=\hat{\beta}(S)$, then $E_{\mu,\Sigma}[l(\hat{Y}_0; Y_0, \Sigma)]$ $=(1+1/N)[R_2(\hat{\beta}; \mu, \Sigma)+q]$.

Thus, such a prediction function can be assessed in terms of the performance of $\hat{\beta}$. In particular, comparison of $\tilde{Y}_0$ and $\tilde{\tilde{Y}}_0$ reduces to comparison of B and $(N/(N+1))B$. It is easy to see that the risk of any estimator kB, where k is a constant and $l=1-k$, is

$$R_2(kB; \mu, \Sigma)=k^2R_2(B; \mu, \Sigma)+l^2 \operatorname{tr} \Sigma_X\beta\Sigma_{Y\cdot X}^{-1}\beta' . \tag{4.1}$$

Using (4.1) and (2.4), we see that $R_2(kB;\,\mu,\,\Sigma)<R_2(B;\,\mu,\,\Sigma)$ if and only if

$$\operatorname{tr}\Sigma_X\beta\Sigma_{Y\cdot X}^{-1}\beta'<\frac{2-l}{l}\cdot\frac{qp}{N-p-2}\,. \tag{4.2}$$

When $l=1/(N+1)$, this is

$$\operatorname{tr}\Sigma_X\beta\Sigma_{Y\cdot X}^{-1}\beta'<(2N+1)\frac{qp}{N-p-2}\,. \tag{4.3}$$

Fact 6. Over the portion of the parameter space described by (4.3), the estimator $[N/(N+1)]B$ has lower risk than B.

REMARK. In the univariate case ($q=1$), letting $\bar{R}^2$ be the multiple correlation between Y and the X's, $\operatorname{tr}\Sigma_X\beta\Sigma_{Y\cdot X}^{-1}\beta'=\bar{R}^2/(1-\bar{R}^2)$. In this case (4.3) is equivalent to

$$\bar{R}^2<\frac{p+2Np}{N-2+2Np}\,. \tag{4.4}$$

For $(p,\,N)$ such that $N\geqq p+3$ (so that the risk of B is finite) the right hand side of (4.4) is always at least 2/3. The inequality (4.2) can be rewritten as

$$l<\frac{b(1-\bar{R}^2)}{a(1-\bar{R}^2)+\bar{R}^2} \tag{4.5}$$

where a and b are constants. This inequality suggests replacing the constant l by a function which with high probability satisfies (4.5). Indeed, Stein [5] has shown that estimators of the form $[1-l(R^2)]B$ where R is the sample multiple correlation and

$$l(R^2)=b(1-R^2)/[a(1-R^2)+R^2]$$

have everywhere lower risk than B for suitable constants a and b. Baranchik [1] showed that if we take $l(R^2)=c(1-R^2)/R^2$ the resulting estimator has everywhere lower risk than B. Here c can be any constant between 0 and $2(p-2)/(N-p+2)$. Note that $c(1-R^2)/R^2$ is inversely proportional to the F-statistic for testing the hypothesis $\beta=0$.

REFERENCES

[1] Baranchik, Alvin J. (1964). Multiple regression and estimation of the mean of a multivariate normal distribution, Tech. Report No. 51, Dept. of Statistics, Stanford University, Stanford, California.

[2] Blackwell, D. and Girshick, M. A. (1954). *Theory of Games and Statistical Decisions.* Wiley, New York.

[3] Kiefer, J. (1957). Invariance, minimax sequential estimation, and continuous time processes, *Ann. Math. Statist.*, 28, 573-601.

[4] Sclove, Stanley Louis (1967). Decision theoretic results for prediction and estimation in multivariate multiple regression, Ph. D. Thesis, Columbia University, New York, N.Y.

[5] Stein, Charles (1960). Multiple regression, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, (ed. I. Olkin), Stanford University Press, Stanford, California, 424-443.