Ann. Inst. Statist. Math.
38 (1986), Part A, 319-329

CRITERIA FOR SELECTION OF RESPONSE VARIABLES AND THE ASYMPTOTIC PROPERTIES IN A MULTIVARIATE CALIBRATION

RYUEI NISHII

(Received Aug. 7, 1984; revised May 16, 1985)

Summary

Let a set of p responses $\boldsymbol{y}=(y_1, \cdots, y_p)'$ has a multivariate linear regression on a set of q explanatory variables $\boldsymbol{x}=(x_1, \cdots, x_q)'$. Our aim is to select the most informative subset of responses for making inferences about an unknown $\boldsymbol{x}$ from an observed $\boldsymbol{y}$. Under normality on $\boldsymbol{y}$, two selection methods, based on the asymptotic mean squared error and on the Akaike's information criterion, are proposed by Fujikoshi and Nishii (1986, *Hiroshima Math. J.*, **16**, 269–277). In this paper, under a mild condition we will derive the cross-validation criterion and obtain the asymptotic properties of the three procedures.

1. Introduction

Consider a linear relationship between p response variables $\boldsymbol{y}=(y_1, \cdots, y_p)'$ and explanatory variables $\boldsymbol{x}=(x_1, \cdots, x_q)'$ such as

$$\boldsymbol{y}=\boldsymbol{\alpha}+\beta'\boldsymbol{x}+\boldsymbol{e}\,, \tag{1.1}$$

where $\boldsymbol{\alpha}: p\times 1$ is the vector of unknown parameters, $\beta: q\times p$ is the matrix of unknown parameters satisfying $\operatorname{rank}\beta=q\leqq p$, and $\boldsymbol{e}: p\times 1$ is an error vector having mean zero vector and unknown covariance matrix Σ. In Sections 1 and 2, we will assume that $\boldsymbol{e}$ is normally distributed. Suppose responses $\boldsymbol{y}_r$ to given $\boldsymbol{x}_r$ $(r=1, \cdots, N)$ are independently observed. Set $Y=[\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N]': N\times p$, $X=[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]': N\times q$ and $E=[\boldsymbol{e}_1, \cdots, \boldsymbol{e}_N]': N\times p$. Then we have the following multivariate linear relationship

$$Y=\boldsymbol{1}\boldsymbol{\alpha}'+X\beta+E\,,$$

Key words: AIC, calibration, cross-validation, multivariate linear regression, variable selection.

where $\mathbf{1}=(1, \cdots, 1)': N\times 1$. For simplicity we assume $X'\mathbf{1}=\mathbf{0}$. The commonly used estimates of $\boldsymbol{\alpha}$, β and Σ are

$$\hat{\boldsymbol{\alpha}}=\boldsymbol{a}=\bar{\boldsymbol{y}}\,, \qquad \hat{\beta}=B=(X'X)^{-1}X'Y$$

$$\text{and} \qquad \hat{\Sigma}=S=\frac{1}{n}Y'\left\{I_N-\frac{1}{N}\mathbf{1}\mathbf{1}'-X(X'X)^{-1}X'\right\}Y\,,$$

where $n=N-q-1>p$ and $\bar{\boldsymbol{y}}=N^{-1}\sum \boldsymbol{y}_r$. The problem of calibration is to make inference about $\boldsymbol{x}$ which corresponds to a new reading $\boldsymbol{y}$. The classical estimate of $\boldsymbol{x}$ is given by

$$\hat{\boldsymbol{x}}=(BS^{-1}B')^{-1}BS^{-1}(\boldsymbol{y}-\bar{\boldsymbol{y}})\,. \tag{1.2}$$

In the case $p=q=1$, this problem of the straight-line calibration has been discussed by many authors, e.g., Shukla [9] and Lwin and Maritz [5]. In the multivariate case, see Williams [10] and Brown [2]. The problem of selecting responses is important since all of responses may not be informative. Brown [2] proposed a procedure which is based on the test of additional information by Rao [7]. Alternatively Fujikoshi and Nishii [4] proposed two procedures. One is based on the asymptotic mean squared error and the other is based on Akaike's information criterion by Akaike [1]. In this paper we propose the third procedure based on the cross-validation criterion. Asymptotic properties of these three procedures are obtained without the assumption of normality.

2. Definition of the true model

If all parameters $\boldsymbol{\alpha}$, β and Σ are known, a natural estimate of $\boldsymbol{x}$ is given by

$$\hat{\boldsymbol{x}}=(\beta\Sigma^{-1}\beta')^{-1}\beta\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\alpha})\,.$$

If the last column of $\beta\Sigma^{-1}$ equals to zero vector, the last response variable y_p is of no use for estimating $\boldsymbol{x}$, in other words, y_p is not informative. Hereafter we suppose that only $y_1, \cdots, y_{p_0}$ are informative. We say $\boldsymbol{y}_{j_0}=(y_1, \cdots, y_{p_0})'$ or $j_0=\{1, \cdots, p_0\}$ be the true model. For $j=\{j_1, \cdots, j_{k(j)}\}$ being a subset of $\{1, \cdots, p\}$, we define a vector of size $k(j)$,

$$\boldsymbol{y}_j=(y_{j_1}, \cdots, y_{j_{k(j)}})'$$

where $k(j)$ denotes the size of j. Then the classical estimate defined in (1.2) under the model j is given by

$$\hat{\boldsymbol{x}}(j)=(B_jS_{jj}^{-1}B_j')^{-1}B_jS_{jj}^{-1}(\boldsymbol{y}_j-\bar{\boldsymbol{y}}_j)\,, \tag{2.1}$$

where $B_j: q\times k(j)$, $S_{jj}: k(j)\times k(j)$ and $\bar{\boldsymbol{y}}_j: k(j)\times 1$ denote the submatrices and the subvector of B, S and $\bar{\boldsymbol{y}}$ specified by j respectively. Therefore it is necessary that $k(j)\geqq q$.

Let J be some family of subsets of $\{1, \cdots, p\}$. Then the problem of variable selection may be regarded as how to select the most informative subset j from J. Now we assume that the family J includes the true model j_0.

Remark. For $j\supset j_0\supset l$, it holds that

$$\beta\Sigma^{-1}\beta'=\beta_j\Sigma_{jj}^{-1}\beta_j=\beta_0\Sigma_{00}^{-1}\beta_0\geqq\beta_l\Sigma_{ll}^{-1}\beta_l' \quad \text{and} \quad \text{trace}\,(\beta\Sigma^{-1}\beta')>\text{trace}\,(\beta_l\Sigma_{ll}^{-1}\beta_l')\,,$$

where $\beta_j: q\times k(j)$ and $\Sigma_{jj}: k(j)\times k(j)$ denote the submatrices of β and Σ respectively, $\beta_0=\beta_{j_0}: q\times p_0$, $\Sigma_{00}=\Sigma_{j_0j_0}: p_0\times p_0$ and $k(j_0)=p_0$.

Hereafter we will make the following assumption.

ASSUMPTION 1. Let $G_N=N^{-1}X'X: q\times q$. Then G_N converges to some positive definite matrix G as N tends to infinity, i.e., $\lim_{N\to\infty} G_N= G>0$.

Fujikoshi and Nishii [4] obtained the stochastic expansion of $(\hat{\boldsymbol{x}}(j)-\boldsymbol{x})'\Delta(\hat{\boldsymbol{x}}(j)-\boldsymbol{x})$ up to $O_p(n^{-1})$ and an estimate of its expectation as $(1+N^{-1})M(j)$ where Δ is a positive definite matrix of order q and

$$M(j)=\frac{n(n-1)}{\{n-k(j)+q\}\{n-k(j)+q-1\}}\,\text{trace}\,\{\Delta(B_jS_{jj}^{-1}B_j')^{-1}\}\,.$$

Their procedure is to select a model j so as to minimize $M(j)$. We denote the selected model by $\hat{j}_M$.

The second procedure is based on Akaike's information criterion. The criterion is analyzed by Fujikoshi [3] in the context of discriminant analysis and the same discussion is possible for our problem. The maximum likelihood under the model j is obtained by the maximum likelihood subject to the constraints $\beta_{j^*}=\beta_j\Sigma_{jj}^{-1}\Sigma_{jj^*}$ where j^* is the complement of j with respect to the full model $\bar{j}_p=\{1,\cdots, p\}$. Let $A(j)=\text{AIC}\,(j)-\text{AIC}\,(\bar{j}_p)$. Then $A(j)$ is given by

$$A(j)=N\log\frac{|Y'(I_N-N^{-1}\mathbf{1}\mathbf{1}')Y||S_{jj}|}{|S||Y_j'(I_N-N^{-1}\mathbf{1}\mathbf{1}')Y_j|}-2q\{p-k(j)\}\,,$$

where $Y_j: N\times k(j)$ is the submatrix of $Y: N\times p$. We can select the subset of responses which minimizes $A(j)$ and denote such index set by $\hat{j}_A$.

3. Derivation of cross-validation

In Sections 1 and 2, error vectors $\boldsymbol{e}_r$ are assumed to be i.i.d. as $N_p[\boldsymbol{0}, \Sigma]$. But to derive cross-validation, it is enough to assume

ASSUMPTION 2. Let $\bar{\boldsymbol{e}}_r=\Sigma^{-1/2}\boldsymbol{e}_r$ $(r=1, \cdots, N)$. Then $\bar{E}=[\bar{\boldsymbol{e}}_1, \cdots, \bar{\boldsymbol{e}}_N]'$: $N\times p$ is an array of N random samples from a p-variate random vector $\bar{\boldsymbol{e}}=(\bar{e}_1, \cdots, \bar{e}_p)'$ such that $\mathcal{E}[\bar{e}_t]=0$, $\mathcal{E}[\bar{e}_t^2]=1$, $\mathcal{E}[\bar{e}_t^4]=\mu_4$, $\mathcal{E}[\bar{e}_t^8]<\infty$ $(t=1, \cdots, p)$ and all moments up to 8th order of $\bar{e}_1, \cdots, \bar{e}_p$ are given as if they are independently distributed, e.g., $\mathcal{E}[\bar{e}_1\bar{e}_2^2]=\mathcal{E}[\bar{e}_1]\mathcal{E}[\bar{e}_2^2]=0$.

Let $\tilde{\boldsymbol{x}}_r$ be the estimate of $\boldsymbol{x}_r$ obtained by replacing B, S and $\bar{\boldsymbol{y}}$ in (2.1) by B_{-r}, S_{-r} and $\bar{\boldsymbol{y}}_{-r}$, respectively. Here B_{-r}, S_{-r} and $\bar{\boldsymbol{y}}_{-r}$ are the estimates of parameters obtained by not using the r-th item. The stochastic expansion of $\tilde{\boldsymbol{x}}_r-\boldsymbol{x}_r$ are given by

$$\tilde{\boldsymbol{x}}_r-\boldsymbol{x}_r=\boldsymbol{u}_r+o_p(n^{-1}) ,$$

where $\boldsymbol{u}_r=c_r^{(1)}W\boldsymbol{v}_r+c_r^{(2)}W(X'X)^{-1}\boldsymbol{x}_r$, $W=(BS^{-1}B')^{-1}$, $\boldsymbol{v}_r=\boldsymbol{y}_r-\bar{\boldsymbol{y}}-B'\boldsymbol{x}_r$, $c_r^{(1)}=1-\frac{1}{n}\boldsymbol{v}_r'S^{-1}B'WBS^{-1}\boldsymbol{v}_r+\frac{1}{n}\boldsymbol{v}_r'S^{-1}\boldsymbol{v}_r+\boldsymbol{x}_r'(X'X)^{-1}WBS^{-1}\boldsymbol{v}_r$ and $c_r^{(2)}=\boldsymbol{v}_r'S^{-1}B'\cdot WBS^{-1}\boldsymbol{v}_r-\boldsymbol{v}_r'S^{-1}\boldsymbol{v}_r$. By the same argument we can construct the estimate $\tilde{\boldsymbol{x}}_r(j)$ under the model j and obtain

$$\tilde{\boldsymbol{x}}_r(j)-\boldsymbol{x}_r=\boldsymbol{u}_r(j)+o_p(n^{-1}) .$$

Using the matrix Δ used in the definition of $M(j)$, we have

THEOREM 1. *The stochastic expansion of* $\sum_{r=1}^{N}\boldsymbol{u}_r'(j)\Delta\boldsymbol{u}_r(j)$ *up to* $O_p(1)$ *is given by*

$$C(j)=n\ \mathrm{trace}\,(\Delta W_j)+2\ \mathrm{trace}\,(\Delta H_jD_jH_j') ,$$

where $W_j=(B_jS_{jj}^{-1}B_j')^{-1}: q\times q$, $H_j=W_jB_jS_{jj}^{-1}V_j': q\times N$, $D_j=D_j^{(1)}-D_j^{(2)}+D^{(3)}: N\times N$, $D_j^{(1)}=\frac{1}{n}\,\mathrm{diag}\,[V_jS_{jj}^{-1}V_j']$, $D_j^{(2)}=\frac{1}{n}\,\mathrm{diag}\,[V_jS_{jj}^{-1}B_j'W_jB_jS_{jj}^{-1}V_j']$, $D^{(3)}=\mathrm{diag}\left[\frac{1}{N}\boldsymbol{1}\boldsymbol{1}'+X(X'X)^{-1}X'\right]$ *and* $V_j=\left\{I_N-\frac{1}{N}\boldsymbol{1}\boldsymbol{1}'-X(X'X)^{-1}X'\right\}Y_j$: $N\times k(j)$. *Here* diag$\,[A]$ *denotes a diagonal matrix whose diagonals are the same as those of* A.

The first term of $C(j)$ can be considered as a bias caused by fitting model j since trace$\,(\Delta W_j)<$trace$\,(\Delta W_l)$ for $j\supset l$. The second term is the complexity of the selected model. When both models j and l $(j\supset l)$ are close to the true model, it holds that trace$(\Delta H_jD_jH_j')>$trace$(\Delta H_lD_lH_l')$. We also select the subset of responses minimizing $C(j)$

and denote such index set by $\hat{j}_C$.

4. Asymptotic distributions of $\hat{j}_M$, $\hat{j}_C$ and $\hat{j}_A$

To obtain asymptotic distributions of three criteria, we need the following assumption, besides Assumptions 1 and 2.

ASSUMPTION 3. Let $\tau_N = \max\{x_r'(X'X)^{-1}x_r \mid r=1, \cdots, N\}$. Then τ_N converges to zero as N tends to infinity.

The following lemma is useful to examine the asymptotic behaviour of $\hat{j}_M$ and $\hat{j}_C$. The proof is placed in Appendix 1.

LEMMA 1. *Let* $J_0 = \{j \in J \mid j \supseteq j_0\}$ *and* $W_0 = W_{j_0}$. *Then for* $j \in J_0$, *we have*

$$nW_0 - nW_j \xrightarrow{L} \Xi(\Xi^{-1}+G^{-1})^{1/2} Z L_j' L_j Z' (\Xi^{-1}+G^{-1})^{1/2} \Xi ,$$

where $\Xi = (\beta\Sigma^{-1}\beta')^{-1}$, $Z: q\times p_1$ $(p_1 = p - p_0)$ *is a random matrix whose all elements are i.i.d. as* $N(0, 1)$.

Here $L_j: (k(j)-p_0)\times p_1$ is an incidence matrix of the two sets $j - j_0$ and $\{p_0+1, \cdots, p\}$ for $j \in J_0$, i.e., $(s, j_{p_0+s} - p_0)$-elements of L_j are given by one $(s=1, \cdots, k(j)-p_0)$ and all other elements are given by zero. For example $L_j = (I_{t-p_0}; 0): (t-p_0)\times p_1$ when $j = \{1, \cdots, p_0, \cdots, t\}$, $t > p_0$.

THEOREM 2. *Let* $J_1 = \{j \in J \mid j \not\supseteq j_0\}$. *Under Assumptions* 1–3, *we have*

(i) *For any* j *in* J_1, $\lim_{N\to\infty} \Pr\{\hat{j}_M = j\} = 0$.

(ii) *Let* $\Omega = (\Xi^{-1}+G^{-1})^{1/2}\Xi\Delta\Xi(\Xi^{-1}+G^{-1})^{1/2}$. *Then for any* j *in* J_0,

$$\lim_{N\to\infty} \Pr\{\hat{j}_M = j\} = \Pr[\text{trace}\{(L_l'L_l - L_j'L_j)Z'\Omega Z\} \leq 2\{k(l)-k(j)\}\,\text{trace}(\Delta\Xi) \ \textit{for}\ l \in J_0] . \tag{4.1}$$

PROOF. (i) Let j be in J_1. Then by Remark following (2.1), we have

$$p\text{-}\lim_{N\to\infty} M(j) = \Pr[\text{trace}\{\Delta(\beta_j\Sigma_{jj}^{-1}\beta_j')^{-1}\} > \text{trace}\{\Delta(\beta_0\Sigma_{00}^{-1}\beta_0')^{-1}\}] = p\text{-}\lim_{N\to\infty} M(j_0) .$$

Hence $\Pr\{\hat{j}_M = j\} \leq \Pr\{M(j) \leq M(j_0)\} = o(1)$.

(ii) Let j be in J_0. Then by Lemma 1 and Remark it holds

$$\begin{aligned} nM(j_0) - nM(j) &= n\,\text{trace}\{\Delta(W_0 - W_j)\} + 2(p_0 - q)\,\text{trace}(\Delta W_0) \\ &\quad - 2(k(j)-q)\,\text{trace}(\Delta W_j) + o_p(1) \\ &= \text{trace}(L_j'L_jZ'\Omega Z) + 2(p_0 - k(j))\,\text{trace}(\Delta\Xi) + o_p(1) . \end{aligned}$$

This completes the proof.

The following lemma will be proved in Appendix 2.

LEMMA 2. *The second term of* $C(j)$ *is asymptotically evaluated as*

$$p\text{-}\lim_{N\to\infty} H_j D_j H_j' = (k(j)+1)\Xi$$

for $j \in J_0$, *where* H_j *and* D_j *are defined in Theorem* 1.

THEOREM 3. *Theorem* 2 *remains valid if* $\hat{j}_M$ *is replaced by* $\hat{j}_C$.

PROOF. (i) The leading term of $\frac{1}{n}C(j)$ is trace (ΔW_j), which is also the leading term of $M(j)$. Therefore the similar discussion shows that $\lim_{N\to\infty} \Pr\{\hat{j}_C = j\} = 0$ for $j \in J_1$.

(ii) Let j be in J_0. Then by Remark, Lemmas 1 and 2, it holds that

$$\begin{aligned} C(j_0) - C(j) &= n \text{ trace } \{\Delta(W_0 - W_j)\} + 2(p_0+1) \text{ trace } \{\Delta(\beta_0 \Sigma_{00}^{-1} \beta_0')^{-1}\} \\ &\quad - 2(k(j)+1) \text{ trace } \{\Delta(\beta_j \Sigma_{jj}^{-1} \beta_j')^{-1}\} + o_p(1) \\ &= \text{trace } (L_j' L_j Z' \Omega Z) + 2(p_0 - k(j)) \text{ trace } (\Delta \Xi) + o_p(1) . \end{aligned}$$

For $\hat{j}_A$, Fujikoshi [3] obtained a theorem on the asymptotic distribution in the context of discriminant analysis under Assumption 1 and normality. Although Assumptions 2 and 3 are weaker than the normality assumption, his theorem still remains valid.

THEOREM 4. *Let* $K(j_0)=0$ *and* $K(j)=\tilde{\Sigma}^{-1/2} L_j' (L_j \tilde{\Sigma}^{-1} L_j')^{-1} L_j \tilde{\Sigma}^{-1/2} : p_1 \times p_1$, *where* $j \in J_0$, $\tilde{\Sigma} = \Sigma_{11} - \Sigma_{10} \Sigma_{00}^{-1} \Sigma_{01} : p_1 \times p_1$ *and* $\Sigma = \begin{bmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{bmatrix}$. *Then we have*

(i) *For any* j *in* J_1, $\lim_{N\to\infty} \Pr\{\hat{j}_A = j\} = 0$.

(ii) *For any* j *in* J_0,

$$\lim_{N\to\infty} \Pr\{\hat{j}_A = j\} = \Pr[\text{trace}[\{K(l) - K(j)\} Z'Z] \leqq 2q\{k(l) - k(j)\} \quad \text{for } l \in J_0] , \tag{4.2}$$

where Z *is defined in Lemma* 1.

Note that the formula (4.1) depends on Δ, G and the matrix of unknown parameters $\Xi = (\beta \Sigma^{-1} \beta')^{-1}$. After a suitable orthogonal transformation, we have

$$\text{trace } (L_j' L_j Z' \Omega Z) = \sum_{t=1}^{r} \sum_{s=1}^{q} \omega_s z_{st}^2 ,$$

with $r = k(j) - p_0$. Here $\omega_s > 0$ are the eigenvalues of the unknown matrix Ω and z_{st} are i.i.d. as $N(0, 1)$. Thus it is not easy to reduce (4.1) into a simple form. On the other hand, (4.2) depends only on

$\tilde{\Sigma}=\Sigma_{11}-\Sigma_{10}\Sigma_{00}^{-1}\Sigma_{01}$, and trace $(K(j)Z'Z)$ has the chi-square distribution with $q(k(j)-p_0)$ degrees of freedom. When J consists of hierarchic models $J=\{\bar{j}_q, \cdots, \bar{j}_p\}$ with $\bar{j}_t=\{1, \cdots, t\}$, the exact formula of (4.2) is obtained by Fujikoshi [3], which is essentially due to Shibata [8]. When J is all subsets such as $J=\{j\subseteq \bar{j}_p | k(j)\geqq q\}$ and $\tilde{\Sigma}$ is diagonal, (4.2) is simplified as

$$\lim_{N\to\infty} \Pr\{\hat{j}_A=j\}=[\Pr\{\chi_q^2\geqq 2q\}]^{k(j)-p_0}[\Pr\{\chi_q^2<2q\}]^{p-k(j)},$$

for $j\in J_0$. This formula is essentially due to Nishii [6].

5. Asymptotic properties in general case

Up to this point, our study is restricted to the case when the regression model has a constant term $\boldsymbol{\alpha}$, which corresponds to an explanatory variable 1. New we generalize (1.1) to

$$\boldsymbol{y}=\alpha'\boldsymbol{x}_0+\beta'\boldsymbol{x}+\boldsymbol{e}, \tag{5.1}$$

where α and β are $q_0\times p$ and $q\times p$ matrices of full rank respectively. Our problem here is to estimate $\boldsymbol{x}$ when $\boldsymbol{y}: p\times 1$ is observed and $\boldsymbol{x}_0: q_0\times 1$ is given. This situation is typical in the missing-data problem. We define $X_0: N\times q_0=[\boldsymbol{x}_{01}, \cdots, \boldsymbol{x}_{0N}]'$ and $X: N\times q=[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]'$. For simplicity suppose $X_0'X=0$, rank $X_0=q_0$ and rank $X=q$. Natural estimates of α, β and Σ are:

$$\hat{\alpha}=A=(X_0'X_0)^{-1}X_0'Y,\quad \hat{\beta}=B=(X'X)^{-1}X'Y \quad \text{and}$$
$$\hat{\Sigma}=S=\frac{1}{m}Y'\{I_N-X_0(X_0'X_0)^{-1}X_0'-X(X'X)^{-1}X'\}Y,$$

where $m=N-q_0-q>p$. Then the estimate of the unknown vector of explanatory variables $\boldsymbol{x}$ is given by

$$\hat{\boldsymbol{x}}=(BS^{-1}B')^{-1}BS^{-1}(\boldsymbol{y}-A'\boldsymbol{x}_0).$$

In this situation Assumption 1 is modified as

ASSUMPTION 4. The matrices $N^{-1}X_0'X_0$ and $G_N=N^{-1}X'X$ converges to positive definite matrices as N tends to infinity respectively, say $G=\lim_{N\to\infty} G_N$.

By suitable analogue we can derive the criteria as

$$M^*(j)=\frac{m(m-1)}{\{m-k(j)+q\}\{m-k(j)+q-1\}}\text{ trace}\{\Delta(B_jS_{jj}^{-1}B_j')^{-1}\},$$

$$C^*(j)=m\text{ trace}\{\Delta(B_jS_{jj}^{-1}B_j')^{-1}\}+2\text{ trace}(\Delta H_j^*D_j^*H_j^{*\prime}),$$

$$A^*(j)=N\log\frac{|S+B'G_NB||S_{jj}|}{|S||S_{jj}+B_j'G_NB_j|}-2q\{p-k(j)\},$$

where $H_j^*=(B_jS_{jj}^{-1}B_j')^{-1}B_jS_{jj}^{-1}V_j^*: q\times N$ and $V_j^*=\{I_N-X_0(X_0'X_0)^{-1}X_0'-X(X'X)^{-1}X'\}Y_j: N\times k(j)$. Here D_j^* is obtained by replacing V_j to V_j^* in D_j described in Theorem 1. We say $\hat{j}_{M^*}=j$ when $M^*(j)=\min_{l\in J} M^*(l)$, and $\hat{j}_{M^*}$ and $\hat{j}_{A^*}$ are defined in the similar way. Finally we make

ASSUMPTION 5. The value $\max\{x_{0r}'(X_0'X_0)^{-1}x_{0r}+x_r'(X'X)^{-1}x_r \mid r=1, \cdots, N\}$ converges to zero as N tends to infinity.

Assumptions 4 and 5 are not so restrictive and they are satisfied when $X_0=\mathbf{1}$. It is not so difficult to check that the following theorem holds true.

THEOREM 6. *Under Assumptions* 2, 4 *and* 5, *the asymptotic distributions of* $\hat{j}_{M^*}$, $\hat{j}_{C^*}$ *and* $\hat{j}_{A^*}$ *are same as those of* $\hat{j}_M$, $\hat{j}_C$ *and* $\hat{j}_A$ *obtained by Theorems* 2, 3 *and* 4 *respectively.*

Acknowledgements

The author wishes to thank the referee for his careful reading and valuable comments.

HIROSHIMA UNIVERSITY

REFERENCES

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, 267-281, (eds. B. N. Petrov and F. Czáki), Akademiai Kiadó, Budapest.

[2] Brown, P. J. (1982). Multivariate calibration, *J. R. Statist. Soc.*, B, **44**, 287-321.

[3] Fujikoshi, Y. (1983). A criterion for variable selection in multiple discriminant analysis, *Hiroshima Math. J.*, **13**, 203-214.

[4] Fujikoshi, Y. and Nishii, R. (1986). Selection of variables in a multivariate inverse regression problem, *Hiroshima Math. J.*, **16**, 269-277.

[5] Lwin, T. and Maritz, J. S. (1982). An analysis of the linear-calibration controversy from the perspective of compound estimation, *Technometrics*, **24**, 235-242.

[6] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression, *Ann. Statist.*, **12**, 2, 758-765.

[7] Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*, Wiley, New York.

[8] Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion, *Biometrika*, **63**, 117-126.

[9] Shukla, G. K. (1972). On the problem of calibration, *Technometrics*, **14**, 547-553.

[10] Williams, E. J. (1959). *Regression Analysis*, Wiley, New York.

Appendix 1

Proof of Lemma 1

To prove Lemma 1, we use the following

PROPOSITION. *Let $\bar{E}=[\bar{E}_0, \bar{E}_1]=[\bar{e}_{rs}]: N\times p$ ($\bar{E}_0$: $N\times p_0$) be an array of N random samples defined in Assumption 2, and let $X: N\times q$ be a matrix of full rank satisfying Assumption 1. Define $\zeta=N^{-1/2}\bar{E}_0'\bar{E}_1: p_0\times p_1$ ($p_1=p-p_0$) and $\xi=(X'X)^{-1/2}X'\bar{E}_1: q\times p_1$. Then all elements of random matrices ζ and ξ are asymptotically i.i.d. as $N(0, 1)$.*

PROOF. Consider the joint characteristic function of ζ and ξ,

$$\phi(T, U)=\mathcal{E}[\exp\{i \operatorname{trace}(T'\zeta+U'\xi)\}],$$

where $T=[t_{kl}]: p_0\times p_1$ and $U=[\boldsymbol{u}_1, \cdots, \boldsymbol{u}_{p_1}]: q\times p_1$. The first four cumulants of trace$(T'\zeta+U'\xi)$ are given by

$$\kappa_1=0, \quad \kappa_2=\operatorname{trace}(T'T+U'U), \quad \kappa_3=0 \quad \text{and}$$

$$\kappa_4=\frac{1}{N}\mu_4^2\sum_{k=1}^{p_0}\sum_{l=1}^{p_1}t_{kl}^4+\mu_4\sum_{r=1}^{N}\sum_{s=1}^{p_1}\{\boldsymbol{x}_r'(X'X)^{-1/2}\boldsymbol{u}_s\}^4. \tag{A.1}$$

By Assumption 2 and Schwarz's inequality, the second term of the right hand side in (A.1) is dominated by

$$\mu_4\sum_{s=1}^{p_1}(u_s'u_s)^2[\max\{\boldsymbol{x}_r'(X'X)^{-1}\boldsymbol{x}_r \mid r=1, \cdots, N\}]^2=O(\tau_N^2),$$

where τ_N is defined in Assumption 3. This implies

$$\log\phi(T, U)=-\frac{1}{2}\operatorname{trace}(T'T+U'U)+O(N^{-1})+O(\tau_N^2),$$

completing the proof.

PROOF OF LEMMA 1. Let $\Gamma=\begin{bmatrix}\Gamma_{00} & 0\\ \Gamma_{10} & \Gamma_{11}\end{bmatrix}: p\times p$ be a matrix such that Γ_{11} is a lower triangular matrix of order p_1 and $\Gamma\Sigma\Gamma'=I_p$. Define $\bar{\beta}=\beta\Gamma'$, $\bar{B}=B\Gamma'$, $\bar{S}=\Gamma S\Gamma'$, $\bar{E}=[\bar{E}_0, \bar{E}_1]$ and $\zeta=(X'X)^{-1/2}X'\bar{E}_1$. For $j\in J_0$, $\bar{\beta}_j$ is a submatrix of $\bar{\beta}$, $\bar{\beta}_0=\bar{\beta}_{j_0}$ and so on. Since j_0 is the true model, $\bar{\beta}=[\bar{\beta}_0, 0]$. Thus

$$\bar{B}_j=[\bar{B}_0, (X'X)^{-1/2}\bar{E}_1L_j]: q\times k(j) \quad \text{and}$$

$$\begin{aligned}\bar{B}_j\bar{S}_{jj}^{-1}\bar{B}_j'=&\bar{B}_0\bar{S}_{00}^{-1}\bar{B}_0'\\ &+\{\bar{B}_0\bar{S}_{00}^{-1}\bar{S}_{01}-(X'X)^{-1/2}\bar{E}_1\}L_j\bar{S}_{jj\cdot 0}^{-1}L_j'\{\bar{S}_{10}\bar{S}_{00}^{-1}\bar{B}_0'-\bar{E}_1'(X'X)^{-1/2}\}\end{aligned}$$

$$=\bar{B}_0\bar{S}_{00}^{-1}\bar{B}_0'+\frac{1}{N}(\bar{\beta}_0\zeta-G^{-1/2}\xi)L_jL_j'(\bar{\beta}_0\zeta-G^{-1/2}\xi)'+o_p\left(\frac{1}{N}\right),$$

where L_j is defined in Lemma 1, $\zeta=N^{-1/2}\bar{E}_0'\bar{E}_1$ and $S_{jj\cdot 0}=S_{jj}-S_{j0}S_{00}^{-1}S_{0j}$. Thus we have

$$n(\bar{B}_0\bar{S}_{00}^{-1}\bar{B}_0')^{-1}-n(\bar{B}_j\bar{S}_{jj}^{-1}\bar{B}_j')^{-1}=\Xi(\bar{\beta}_0\zeta-G^{-1/2}\xi)L_j'L_j(\bar{\beta}_0\zeta-G^{-1/2}\xi)'\Xi+o_p(1)$$
$$\xrightarrow{L} \Xi(\Xi^{-1}+G^{-1})^{1/2}ZL_j'L_jZ'(\Xi^{-1}+G^{-1})^{1/2}\Xi,$$

where $Z:q\times p_1$ is a random matrix whose all elements are i.i.d. as $N(0,1)$.

Appendix 2

Proof of Lemma 2

Obviously, $\bar{S}_{jj}$ and $\bar{B}_j\bar{S}_{jj}^{-1}\bar{B}_j'$ converge in probability to $I_{k(j)}$ and $\bar{\beta}_j\bar{\beta}_j'$ respectively. Define

$$T_1: p\times p=\bar{E}'\,\mathrm{diag}\,[\bar{E}(\bar{E}'\bar{E})^{-1}\bar{E}']\bar{E},$$

$$T_2: p\times p=\frac{1}{n}\bar{E}'\,\mathrm{diag}\,[\bar{E}\bar{\beta}'(\bar{\beta}\bar{\beta}')^{-1}\beta\bar{E}']\bar{E} \quad \text{and}$$

$$T_3: p\times p=\bar{E}'\left[\frac{1}{N}11'+X(X'X)^{-1}X'\right]\bar{E}.$$

To prove Lemma 2, it is sufficient to show that

(i) $T_1 \xrightarrow{P} (p-1+\mu_4)I_p$,

(ii) $T_2 \xrightarrow{P} (q-1+\mu_4)\bar{\beta}'(\bar{\beta}\bar{\beta}')^{-1}\bar{\beta}$ and

(iii) $T_3 \xrightarrow{P} (q+1)I_p$.

(i) Let $W=(w_{kl})=(N^{-1}\bar{E}'\bar{E})^{-1}$. Then $W\xrightarrow{P} I_p$. The k-th diagonal element of T_1 is

$$(T_1)_{kk}=\frac{1}{N}\left[w_{kk}\sum_{r=1}^{N}\bar{e}_{rk}^4+\sum_{s\neq k}w_{ss}\sum_{r=1}^{N}\bar{e}_{rk}^2\bar{e}_{rs}^2+2\sum_{s\neq k}w_{sk}\sum_{r=1}^{N}\bar{e}_{rs}\bar{e}_{rk}^3\right.$$
$$\left.+\sum_{s\neq t\neq k\neq s}w_{st}\sum_{r=1}^{N}\bar{e}_{rk}^2\bar{e}_{rs}\bar{e}_{rt}\right].$$

By law of large numbers, $(T_1)_{kk}\xrightarrow{P}\mu_4+p-1$. Similar discussion leads us to $(T_1)_{kl}\xrightarrow{P}0$ for $k\neq l$.

(ii) Since $\mathrm{rank}\,\bar{\beta}=q$, there exists a non-singular matrix $T:q\times q$ and an orthogonal matrix $\Gamma=[\Gamma_0;\Gamma_1]$ of order $p(\Gamma_0:p\times p_0)$ such that $\bar{\beta}=[T;0]\Gamma'=T\Gamma_0'$. Transform $\bar{E}$ by Γ as $W=\bar{E}\Gamma=[W_0;W_1]:N\times p$. Then

$$\bar{\beta}'(\bar{\beta}\bar{\beta}')^{-1}\bar{\beta}=\Gamma_0\Gamma_0' \quad\text{and}\quad T_2=\Gamma[W_0;W_1]'\,\mathrm{diag}\,[W_0W_0'][W_0;W_1]\Gamma'.$$

By similar arguments as in the proof of (i), it holds that

$$T_2 \xrightarrow{P} (q-1+\mu_4)\Gamma_0\Gamma_0'=(q-1+\mu_4)\bar{\beta}'(\bar{\beta}\bar{\beta}')^{-1}\bar{\beta} .$$

(iii) The (k, l)-element of T_3 is given by

$$(T_3)_{kl}=\sum_{r=1}^{N}\left\{\frac{1}{N}\delta_{kl}+\boldsymbol{x}_r'(X'X)^{-1}\boldsymbol{x}_r\right\}\bar{e}_{rk}\bar{e}_{rl} ,$$

where δ_{kl} is Kronecker's delta. When $k=l$, $\mathcal{E}[(T_3)_{kk}]=q+1$ and the variance of $(T_3)_{kk}$ is equal to $\sum_{r=1}^{N}\left\{\frac{1}{N}+\boldsymbol{x}_r'(X'X)^{-1}\boldsymbol{x}_r\right\}^2\leqq(q+1)(N^{-1}+\tau_N)=O(\tau_N)$. Similarly we have $(T_3)_{kl} \xrightarrow{P} 0$ when $k\neq l$.