Biometrika (1989), **76**, 2, *pp.* 369–74
Printed in Great Britain

A strongly consistent procedure for model selection in a regression problem

BY C. RADHAKRISHNA RAO AND YUEHUA WU
Center for Multivariate Analysis, Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A.

SUMMARY

We consider the multiple regression model $Y_n = X_n\beta + E_n$, where Y_n and E_n are n-vector random variables, X_n is an $n \times m$ matrix and β is an m-vector of unknown regression parameters. Each component of β may be zero or nonzero, which gives rise to 2^m possible models for multiple regression. We provide a decision rule for the choice of a model which is strongly consistent for the true model as $n \to \infty$. The result is proved under certain mild conditions, for instance without assuming normality of the distribution of the components of E_n.

Some key words: AIC; BIC; GIC; Linear regression; Model selection; Variable selection.

1. INTRODUCTION

Consider the multiple regression model

$$Y_n = X_n\beta + E_n, \tag{1·1}$$

where Y_n and E_n are n-vectors, $\beta = (\beta_1, \ldots, \beta_m)'$ is an m-vector parameter and $X_n = (x_{1n} : \ldots : x_{mn}) = (x^{(1)} : \ldots : x^{(n)})'$ is the $n \times m$ design matrix. Let, for an index set $j = \{j_1, \ldots, j_k\}$ $(1 \leqslant j_1 < \ldots < j_k \leqslant m)$,

$$X_n^j = (x_{j_1 n} : \ldots : x_{j_k n}), \quad \beta_{(j)} = (\beta_{j_1}, \ldots, \beta_{j_k})$$

and define model or hypothesis j by H_j: $\beta_i \neq 0$ $(i \in j)$ and $\beta_i = 0$ $(i \notin j)$. There are 2^m hypotheses of this type, and our problem is to give a decision rule to select a hypothesis closest to the true hypothesis in some sense. Let S_j be the residual sum of squares under the hypothesis H_j and $\hat\sigma_j^2 = S_j/\{n - \text{card}(j)\}$, where card (j) is the number of elements in the set j.

There is considerable literature on this problem known as selection of variables in a regression model; see review papers by Hocking (1976) and Thompson (1978a, b). More recently, methods have been proposed for the choice of a model by minimizing a criterion function defined on the set of alternative models, i.e. on sets j in our case. Some of these criteria are:

$n \log (n^{-1}S_j) + 2\,\text{card}\,(j)$	Akaike (1973),
$S_j + 2\hat\sigma_j^2\,\text{card}\,(j)$	Akaike (1970),
$S_j + 2\hat\sigma_J^2\,\text{card}\,(j)$	Mallows (1973),
$S_j + \alpha\hat\sigma_J^2\,\text{card}\,(j)$	Shibata (1984),
$n \log (n^{-1}S_j) + \text{card}\,(j) \log n$	Schwartz (1978),
$n \log (n^{-1}S_j) + \text{card}\,(j) c \log \log n$	Hannan & Quinn (1979),
$n \log (n^{-1}S_j) + \text{card}\,(j) C_n$	Bai, Krishnaiah & Zhao (1986),

where J stands for whole set $\{1, \ldots, m\}$, c is a constant and C_n is such that $n^{-1}C_n \to 0$ and $(\log \log n)^{-1}C_n \to \infty$ as $n \to \infty$. The performance of these criteria under the assumption of normality of the error components has been studied by Nishi (1984), Shibata (1984) and others.

In this paper, we introduce a new criterion with a flexible penalty function and prove its strong consistency without making any distributional assumptions. Since this approach admits a wider range of the choice of the penalty function, it may lead to a better performance in small samples by suitably choosing the penalty than those based on fixed penalties.

2. PRELIMINARIES

We need the following lemmas in the sequel.

LEMMA 1. *Denote the eigenvalues of a $k \times k$ symmetric matrix A by $\lambda_1(A) \geq \ldots \geq \lambda_k(A)$. Let $b_1, \ldots, b_m$ be n-vectors and write $G_k = B'_k B_k$, where $B_k = (b_1 : \ldots : b_k)$ $(k = 1, \ldots, m)$. If there exist constants η_1 and η_2 such that*

$$0 < \eta_1 \leq \lambda_m(G_m) \leq \lambda_1(G_m) \leq \eta_2,$$

then

(i) $\eta_1 \leq b'_k b_k \leq \eta_2 \quad (1 \leq k \leq m)$,
(ii) $\eta_1 \leq b'_k Q_{k-1} b_k \leq \eta_2 \quad (1 < k \leq m)$,

where Q_{k-1} is the projection operator onto the orthogonal complement of the space generated by $b_1, \ldots, b_{k-1}$.

LEMMA 2. *Let $X_n = (x_{1n} : \ldots : x_{kn})$, where x_{in} is an n-vector, and E_n be an n-vector variable, for $n = 1, 2, \ldots$, such that $x'_{jn}E_n = O(n \log \log n)^{\frac{1}{2}}$, almost surely, for $1 \leq j \leq k$ and $0 < cn \leq \lambda_k(X'_n X_n)$, where c is a constant. Then $E'_n P_n E_n = O(\log \log n)$, almost surely, where $P_n = X_n(X'_n X_n)^{-1}X'_n$.*

LEMMA 3. *Let $\eta_1, \eta_2, \ldots$ be a sequence of independent and identically distributed random variables such that $E(\eta_1) = 0$, $E(\eta_1^2) = \sigma^2$ and $E(|\eta_1|^3) < \infty$. Further let $a_1, a_2, \ldots$ be a sequence of constants such that*

(i) $B_n^2 = \sum_{i=1}^n a_i^2 \to \infty$, *as* $n \to \infty$;

(ii) $\sum_{i=1}^n |a_i^3| = O\{B_n^3(\log B_n^2)^{-1-\delta}\}$, *for some* $\delta > 0$.

Then, almost surely,

$$T_n = \sum_{i=1}^n a_i \eta_i = O(B_n^2 \log \log B_n^2)^{\frac{1}{2}}.$$

Lemmas 1 and 2 can be proved by elementary calculus and Lemma 3 follows from Theorem 3 of Petrov (1975, p. 111).

3. The discriminant criterion

Consider the regression model (1·1) such that

$$0 < a_1 n \leqslant \lambda_m(X'_n X_n) \leqslant \lambda_1(X'_n X_n) \leqslant a_2 n, \tag{3·1}$$

for some constants a_1 and a_2, and the components $x^1_{jn}, \ldots, x^n_{jn}$ of x_{jn} satisfy the condition

$$\sum_{i=1}^{n} (x^i_{jn})^3 = O\{(x'_{jn} x_{jn})^{3/2} / \log (x'_{jn} x_{jn})\}^{1+\delta}, \tag{3·2}$$

for $1 \leqslant j \leqslant m$ and some $\delta > 0$. Further let the components $\varepsilon_1, \ldots, \varepsilon_n$ of E_n be independent and identically distributed random variables such that

$$E(\varepsilon_i) = 0, \quad E(\varepsilon_i^2) = \sigma^2, \quad E(|\varepsilon_i|^3) < \infty. \tag{3·3}$$

We first consider a simple case, i.e. the models

$$\beta = \beta_{(k)} = (\beta_1, \ldots, \beta_k \neq 0, 0, \ldots, 0) \quad (k = 1, \ldots, m).$$

Let S_k be the residual sum of squares when $\beta_{(k)}$ is fitted and denote $S_m/(n-m)$ by $\hat{\sigma}^2_m$. Define the discriminant criterion $D_n(k) = S_k + k\hat{\sigma}^2_m C_n$ $(k = 1, \ldots, m)$ and the selection rule $k = \hat{k}_n$, where

$$D_n(\hat{k}_n) = \min_{1 \leqslant k \leqslant n} D_n(k).$$

Then we have the following theorem.

Theorem 3·1. *Suppose that the conditions* (3·1), (3·2) *and* (3·3) *hold for* $n = 1, 2, \ldots$ *and* $k = k_0$ *is the true model. Then* $\hat{k}_n \to k_0$, *almost surely, if we choose* C_n *so that*

$$n^{-1} C_n \to 0, \quad (\log \log n)^{-1} C_n \to \infty. \tag{3·4}$$

Proof. By conditions (3·1)–(3·3), applying Lemmas 1–3, one can easily get

$$a_2 n \geqslant x'_{jn} x_{jn} \geqslant a_1 n \to \infty \quad (1 \leqslant j \leqslant m) \tag{3·5}$$

as $n \to \infty$, and

$$a_2 n \geqslant x'_{jn}(I - P_{j-1}) x_{jn} \geqslant a_1 n > 0 \quad (1 \leqslant j \leqslant m), \tag{3·6}$$

where P_i represents the orthogonal projection operator onto the space spanned by $x_{1n}, \ldots, x_{in}$, and almost surely, for $1 \leqslant j \leqslant m$,

$$x'_{jn} E_n = O(n \log \log n)^{\frac{1}{2}}, \tag{3·7}$$

$$E_n P_j E_n = O(\log \log n). \tag{3·8}$$

Now we are in a position to prove the strong consistency of $\hat{k}_n$. First consider the case $k < k_0$. We have, by (3·5)–(3·7) and Cauchy–Schwarz inequality, together with the condition $n^{-1} C_n \to 0$,

$$D_n(k) - D_n(k_0) \geqslant \beta^2_{k_0} x'_{k_0 n}(I - P_{k_0 - 1}) x_{k_0 n} + 2\beta_{k_0} E'_n (I - P_{k_0-1}) x_{k_0 n} - (k_0 - k) C_n \hat{\sigma}^2_m$$
$$\geqslant \beta^2_{k_0} a_1 n + \beta_{k_0} O(n \log \log n)^{\frac{1}{2}} - (k_0 - k) C_n \hat{\sigma}^2_m > 0,$$

almost surely, for n large enough, which implies, almost surely,

$$\liminf \hat{k}_n \geqslant k_0. \tag{3·9}$$

Next, consider the case $k > k_0$. We have

$$D_n(k) - D_n(k_0) = (k - k_0)C_n\hat{\sigma}_m^2 - \sum_{j=k_0+1}^{k} E_n'(P_j - P_{j-1})E_n. \quad (3\cdot 10)$$

Applying (3·5), (3·6), (3·7), (3·8) and Cauchy–Schwarz inequality to (3·10) we have

$$D_n(k) - D_n(k_0) = (k - k_0)C_n\hat{\sigma}_m^2 + O(\log\log n).$$

As $\hat{\sigma}_n^2(m) \to \sigma^2$, almost surely (Gleser, 1966, p. 1053), $D_n(k) - D(k_0) > 0$, almost surely, as $n \to \infty$, implying, almost surely,

$$\limsup \hat{k}_n \leqslant k_0. \quad (3\cdot 11)$$

Then (3·9) and (3·11) prove the theorem. □

THEOREM 3·2. *Under the same conditions as in Theorem* 3·1 *on the model* (1·1), *the choice* $\hat{k}_n$ *such that*

$$D_n(\hat{k}_n) = \min_{1 \leqslant k \leqslant m} D_n(k),$$

where $D_n(k) = n \log \tilde{\sigma}_k^2 + kC_n$, $\tilde{\sigma}_k^2 = S_k/n$, *is strongly consistent for the true value* k_0 *of* k.

It can be proved in the same way as in Theorem 3·1.

4. THE GENERAL CASE

In § 3, we considered the linear model (1·1) and discussed model selection in a class of nested alternative models. Now we consider the 2^m possible models by allowing each component of β to be zero or nonzero. We can approach this problem by using the result of Theorem 3·1 as follows. For each permutation π of the components of β, by a corresponding rearrangement of $x_{1n}, \ldots, x_{mn}$, we have a linear model on which we can apply the method of § 3 and select a model $\hat{k}_\pi$. From among the models $\hat{k}_\pi$, by varying π over all the permutations, we select that which has the smallest number of nonzero components of β. This procedure is equivalent to minimizing $S_j + \text{card}\,(j)\hat{\sigma}_m^2 C_n$ or $n \log (S_j/n) + \text{card}\,(j) C_n$ over j, where j now stands for a subset of the components of β taken as nonzero. Both the procedures provide a strongly consistent estimate of the true model in view of the theorems in § 3. However, they involve heavy computations. In light of this we suggest an alternative which involves only the computation of $m+1$ residual sum of squares.

Let us consider $\beta_{(-i)} = (\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_m)$ and represent the corresponding residual sum of squares by $S_{(-i)}$ $(i = 1, \ldots, m)$. Define $D_n(-i) = S_{(-i)} - S_m - C_n$, where as before S_m is the residual sum of squares without any restriction on the components of β. Then choose the model $\beta_i = 0$ if $D_n(-i) \leqslant 0$, and $\beta_i \neq 0$ if $D_n(-i) > 0$ $(i = 1, \ldots, m)$.

We have the following theorem.

THEOREM 4·1. *Under the conditions of Theorem* 3·1, *the estimated model by the rule given above is strongly consistent for the true model.*

Proof. If in the true model $\beta_i \neq 0$, then, using the second equation above (3·9) with $k_0 = m$ and $k = m - 1$, we have with probability 1, $D_n(-i) > 0$ for all large n; that is β_i is taken to be nonzero in the selected model. Conversely if $\beta_i = 0$, using (3·5)–(3·8) and the Cauchy–Schwarz inequality we get

$$\begin{aligned} D_n(-i) = S_{(-i)} - S_m - C_n &= Y_n'(P_m - P_{(-i)})Y_n - C_n \\ &= E_n'(P_m - P_{(-i)})E_n - C_n \leqslant O(\log\log n) - C_n, \end{aligned}$$

which, together with the condition $(\log\log n)^{-1}C_n \to \infty$ of (3·4), implies that, with probability 1, $D_n(-i)<0$, for all large n; that is β_i is not in the selected model. This completes the proof of Theorem 4·1. □

5. Some comments on the choice of C_n

In Theorems 3·1, 3·2 and 4·1, we proved the strong consistency of our model selection criteria under the conditions (3·4). There are many choices of C_n which ensure (3·4). The actual choice of C_n in any given problem may depend on other considerations such as the consequences of selecting a wrong model. We suggest an ad hoc procedure which appears to be promising.

First, take the full model (1·1) without any restriction on the components of β, and estimate σ^2 by $\hat{\sigma}_m^2 = S_m/(n-m)$ and the residuals by $\hat{E}_n = Y_n - X_n\hat{\beta}$, where $\hat{\beta}$ is the least-squares estimator of β.

Secondly, consider the models,

$$M_k\colon\ Y_n = X_n\gamma_k + E_n \quad (k=1,\ldots,m),$$

where $\gamma_k = (a\hat{\sigma}_m,\ldots,a\hat{\sigma}_m,0,\ldots,0)'$ with the last $(m-k)$ components as zeros and $a<1$ is some chosen constant.

Thirdly, choose C_n of the form αn^{γ} where $\gamma<1$ and construct observations $Y_n = X_n\gamma_k + \hat{E}_n$, where $\hat{E}_n$ is the vector of estimated residuals. For a given combination of α and γ we find which of the models $M_1,\ldots,M_m$ are correctly selected. We call a combination (α, γ) good if all the models are correctly selected. There may be several combinations (α, γ) which are good. We may fix a particular value of γ and look at the set of values of α and choose some representative value. Such a choice of α and γ gives C_n which can be used in the actual selection of a model in a given problem.

In simulation experiments, we chose $a=0{\cdot}6$ to ensure a good performance of the selection rule when the regression coefficients are of order not less than $0{\cdot}6\sigma$. We fixed γ at 0·9 and selected α as $\frac{1}{3}\alpha_{\min}+\frac{2}{3}\alpha_{\max}$ from among the 'good values' of α with $\gamma=0{\cdot}9$. Such a choice of C_n gave good results when the sample was not too small, compared to criteria such as BIC, AIC and jack-knife, cross validation, by leaving one out. Further research is needed for prescribing rules for the choice of C_n.

Acknowledgements

This research was sponsored by the Air Force Office of Scientific Research and the Office of Naval Research.

References

Akaike H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203–17.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pp. 267–81. Budapest: Akademiai Kiado.

Bai, Z. D., Krishnaih, P. R. & Zhao, L. (1986). On the detection of number of signals in the presence of white noise. *J. Mult. Anal.* **20**, 1–25.

Gleser, L. J. (1966). Correction to "on the asymptotic theory of fixed size sequential confidence bounds for linear regression parameters". *Ann. Math. Statist.* **37**, 1053–5.

Hannan, E. J. & Quinn, B. G. (1979). The determination of the order of an autoregression. *J. R. Statist. Soc.* B **41**, 190–5.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32**, 1–49.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661–75.

NISHI, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12**, 758-65.

PETROV, V. V. (1975). *Sum of Independent Random Variables.* Berlin: Springer-Verlag.

SCHWARTZ, G. (1978). Estimating the dimensions of a model. *Ann. Statist.* **6**, 461-4.

SHIBATA, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika* **71**, 43-9.

THOMPSON, M. L. (1978a). Selection of variables in multiple regression. Part I. A review and evaluation. *Int. Statist. Rev.* **46**, 1-19.

THOMPSON, M. L. (1978b). Selection of variables in multiple regression. Part II. Chosen procedures, computations and examples. *Int. Statist. Rev.* **46**, 126-46.

[*Received June* 1988. *Revised November* 1988]