

AD-A044 889

PURDUE UNIV LAFAYETTE IND DEPT OF STATISTICS  
ON SELECTING AN OPTIMAL SUBSET OF REGRESSION VARIABLES.(U)  
JUL 77 S S GUPTA, D Y HUANG  
MIMEOGRAPH SER-501

F/G 12/1

N00014-75-C-0455

UNCLASSIFIED

NL

1 of 1  
ADA044889



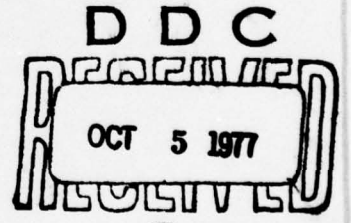
END  
DATE  
FILMED  
11-77  
DDC

P  
B.S.



AD A 044889

# PURDUE UNIVERSITY



*Handwritten initials*  
C

AD No. \_\_\_\_\_  
DDC FILE COPY

## DEPARTMENT OF STATISTICS

## DIVISION OF MATHEMATICAL SCIENCES

DISTRIBUTION STATEMENT A  
Approved for public release;  
Distribution Unlimited

6

On Selecting an Optimal Subset of Regression Variables

10

Shanti S./Gupta by  
Purdue University

and

D. Y./Huang Academia Sinica, Taipei

9

Technical rept.,

Department of Statistics  
Division of Mathematical Sciences

14 Mimeograph Series - 501

11 July 1977

12 16p.

D D C  
OCT 5 1977  
C

15

\*This research was supported by the Office of Naval Research contract  
N00014-75-C-0455 at Purdue University. Reproduction in whole or in  
part is permitted for any purpose of the United States Government.

DISTRIBUTION STATEMENT A  
Approved for public release;  
Distribution Unlimited

1493  
291 730

1B

On Selecting an Optimal Subset of Regression Variables

by

Shanti S. Gupta, Purdue University

and

D. Y. Huang, Academia Sinica, Taipei

ABSTRACT

In the past decade a number of methods have been developed for selecting the "best" or at least a "good" subset of variables in regression analysis. For various reasons, we may be interested in including only a subset say, of size  $r < p$ , the number of independent variables. Various authors have considered this problem and a variety of techniques are presently being used to construct such subsets. Most of them seem to lack justification in terms of statistical theory.

In this paper, we are interested in deriving a selection procedure to select a random size optimal subset such that all inferior independent variables are excluded. Some results on the efficiency of the procedure are also discussed.

ACCESSION for

NTIS

DDC

INMANUIC'D

JUSTIFICATION

BY

DESCRIPTIVE/ANALYTICAL

ABILITY DATES

SPECIAL

White Section

Buff Section

**A**

## On Selecting an Optimal Subset of Regression Variables

by

Shanti S. Gupta, Purdue University  
and

D. Y. Huang, Academia Sinica, Taipei

In the past decade a number of methods have been developed for selecting the "best" or at least a "good" subset of variables in regression analysis. For various reasons, we may be interested in including only a subset say, of size  $r < p$ , the number of independent variables. Various authors have considered this problem and a variety of techniques are presently being used to construct such subsets. They seem to lack justification by statistical theory (see e.g. [2], [6]).

Arvesen and McCabe [1] propose a procedure for selecting a subset within a class of subsets with  $t$  (fixed) independent variables, taking into account the statistical variation of the residual sum of squares. An algorithm for determining the necessary constant  $c$  given the design matrix  $X$  is presented in [4].

In this paper, we are interested in deriving a selection procedure to select a random size subset excluding all inferior independent variables (defined later). Some results on the efficiency of the procedure are also discussed. It should be pointed out that our approach is different

---

\*This research was supported by the Office of Naval Research contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

from Arvesen and McCabe [1] and the approaches used by others.

Let  $\pi_0, \pi_1, \dots, \pi_k$  denote  $k+1$  normal populations with variances  $\sigma_0^2, \sigma_1^2, \dots, \sigma_k^2$ . Let  $\sigma_{[1]}^2 \leq \dots \leq \sigma_{[k]}^2$  denote the ordered variances.

A population  $\pi_i$  is said to be

$$\begin{array}{ll} \text{superior (or good)} & \text{if } \sigma_0^2 \geq \delta_1^* \sigma_i^2, \\ \text{inferior (or bad)} & \text{if } \sigma_0^2 \leq \delta_2^* \sigma_i^2, \end{array}$$

where  $\delta_1^*, \delta_2^*$  are specified constants such that  $0 < \delta_2^* < \delta_1^* < 1$ .

We are interested in devising a procedure which selects a random size subset, that excludes all the inferior populations with a probability not less than  $P^*$ , a specified constant.

Let  $\Omega$  be the parameter space which is the collection of all possible parameter vector  $\underline{\theta} = (\sigma_0^2, \sigma_1^2, \dots, \sigma_k^2)$ . Let  $t_1$  and  $t_2$  denote, respectively, the unknown number of inferior and superior populations in the given collection of  $k+1$  populations. We have  $t_1 \geq 0$ ,  $t_2 \geq 1$  and  $t_1 + t_2 \leq k+1$ . For specified  $\delta_1^*$  and  $\delta_2^*$ , let

$$\begin{aligned} \Omega(t_1, t_2) = \{ \theta : & \sigma_{[1]}^2 \leq \dots \leq \sigma_{[t_2]}^2 \leq \frac{\sigma_0^2}{\delta_1^*} < \sigma_{[t_2+1]}^2 \\ & \leq \dots \leq \sigma_{[k-t_1]}^2 < \frac{\sigma_0^2}{\delta_2^*} \leq \sigma_{[k-t_1+1]}^2 \leq \dots \leq \sigma_{[k]}^2 \}. \end{aligned}$$

Then

$$\Omega = \bigcup_{t_1, t_2} \Omega(t_1, t_2).$$

Let CD stand for a correct decision which is defined to be selection of the subset which excludes all the inferior populations.

Assume the following standard linear model

$$(1) \quad \underline{Y} = X\underline{\beta} + \underline{\epsilon}, \quad X = (1, X_1, \dots, X_{p-1}), \quad \underline{\beta}' = (\beta_0, \beta_1, \dots, \beta_{p-1}),$$

where  $X$  is an  $N \times p$  known matrix of rank  $p \leq N$ ,  $\underline{\beta}$  is a  $p \times 1$  parameter vector, and  $\underline{\epsilon} \sim N(0, \sigma_0^2 I_N)$ , and  $\underline{1}' = (1, 1, \dots, 1)$ .

In what follows, (1) which has  $k (= p-1)$  independent variables, will be viewed as the "true" model.

Consider the models

$$(2) \quad \underline{Y} = X_{(i)}\underline{\beta}_{(i)} + \underline{\epsilon}_i$$

where  $X_{(i)} = (1, X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_k)$  and  $\underline{\beta}_{(i)}' = (\beta_0, \beta_1, \dots, \beta_{i-1}, \dots, \beta_{i+1}, \dots, \beta_k)$ , and  $\underline{\epsilon}_i \sim N(0, \sigma_i^2 I_k)$ ,  $i=1, \dots, k$ .  $X_{(i)}$  associated with model (2) is called population  $\pi_i$  ( $1 \leq i \leq k$ ). The goal is to reject  $\pi_i$ , i.e. to reject  $X_i$ , associated with  $\sigma_{[j]}^2$ ,  $j = k-t_1+1, \dots, k$ , for any fixed  $t_1$ .

Note that

$$SS_i = \underline{Y}' \{I - X_{(i)}(X_{(i)}'X_{(i)})^{-1}X_{(i)}'\} \underline{Y} = \underline{Y}' Q_i \underline{Y},$$

where  $Q_i = [I - X_{(i)}(X_{(i)}'X_{(i)})^{-1}X_{(i)}']$ , then following Searle [5, p. 57],

$$SS_i / \sigma_0^2 \sim \chi^2 \{r(Q_i), (X_{\underline{\beta}})' Q_i (X_{\underline{\beta}}) / (2\sigma_0^2)\},$$

where  $r(Q_i) = N - k = v$ . Note that the noncentrality parameter, in general, is not zero and that

$$(3) \quad \sigma_i^2 = \sigma_0^2 + (X_{\underline{\beta}})' Q_i (X_{\underline{\beta}}) / v.$$

Assume that  $\sigma_0^2$  is known. Since the problem is invariant with respect to the scaling by  $\sigma_0^2 > 0$ , we assume without loss of generality that  $\sigma_0^2 = 1$ .

To obtain the joint distribution of  $SS_1, \dots, SS_k$ , we can write

$$\underline{Y}' Q_i \underline{Y} = U_i' U_i,$$

where

$$(4) \quad U_i = B_i \underline{Y} \text{ and } B_i B_i' = I, \quad B_i' B_i = Q_i$$

$B_i$  is an  $v \times N$  matrix.

The joint distribution of  $U' = (U_1', \dots, U_k')$  is multivariate normal in  $kv$  dimensions with mean vector  $\underline{\eta}' = (\eta_1, \dots, \eta_k)$ ,  $\eta_i = B_i' X \beta$ , and covariance matrix  $\Sigma = (\Sigma_{ij})$  where  $\Sigma = \sigma_i \sigma_j B_i B_j'$ . Note that the  $kv \times kv$  covariance matrix  $\Sigma$  is possibly singular. Let  $\Sigma = FF'$  where  $F$  is of full column rank  $r$  ( $r = \text{rank}(\Sigma)$ ), and let  $U = \underline{\eta} + FA$  where  $A \sim N(\underline{0}, I_r)$ . Thus, the joint characteristic function of

$\frac{SS_1}{2}, \dots, \frac{SS_k}{2}$  is (since  $SS_i = U_i' U_i$ ),

$$\begin{aligned} \varphi(t_1, \dots, t_k) &= E\left\{\exp\left(i \sum_{j=1}^k t_j (U_j)' U_j / 2\right)\right\} \\ &= |I - iF'TF|^{-\frac{1}{2}} \\ &\quad \cdot \exp\left\{\frac{1}{2} [\underline{\eta}' (iT - TF(I - iF'TF)^{-1} F'T) \underline{\eta}]\right\} \\ &= |I - i\Sigma T|^{-\frac{1}{2}} \exp\left\{\frac{1}{2} \underline{\eta}' T (I - i\Sigma T)^{-1} \underline{\eta}\right\}, \end{aligned}$$

where  $T = \text{diag}(t_1, \dots, t_k) \otimes I_\nu$ .

We propose the rejection rule of the form:

R: Reject  $\pi_i$  (or reject  $X_i$ ) is and only if

$$SS_i \geq \frac{vc}{\delta_2^*}$$

where  $\delta_2^* < c < 1$ .

Note that  $SS_i$  is associated with  $U_i$  or, equivalently, with population  $\pi_i$  and degrees of freedom  $\nu$ , whereas  $SS_{[i]}$  is the  $i$ -th smallest sum of squares and  $SS_{(i)}$  is the sum of squares corresponding to the (unknown)  $i$ -th smallest expected sum of squares  $\sigma_{[i]}^2$  and degrees of freedom  $\nu$ . Thus

$$\begin{aligned} \inf_{\theta} P_{\theta}(CD|R) &= \inf_{\theta} P_{\theta} \left\{ \min_{k-t_1+1 \leq i \leq k} SS_{(i)} \geq \frac{vc}{\delta_2^*} \right\} \\ &= \inf_{\theta} P_{\theta} \left\{ \min_{k-t_1+1 \leq i \leq k} \frac{SS_{(i)}}{\sigma_{[i]}^2} \geq \frac{vc}{\delta_2^*} \frac{1}{\sigma_{[i]}^2} \right\} \end{aligned}$$

(4)

$$= \min_{0 \leq t_1 \leq k} \inf_{\beta} P \left\{ \min_{k-t_1+1 \leq i \leq k} \frac{SS_{(i)}}{\sigma_{[i]}^2} \geq vc \right\}.$$

It is clear that the bound in (4) approaches a minimum value as the parameters  $\sigma_{[i]}^2$ ,  $k-t_1+1 \leq i \leq k$  for any  $t_1$ , approach  $\frac{1}{\delta_2^*}$ . Since this limiting probability does not depend on the value of  $\sigma_{[i]}^2$ ,  $k-t_1+1 \leq i \leq k$  for any  $t_1$ , we can assume that they are all equal to  $\frac{1}{\delta_2^*}$ .

Thus

$$\inf P(CD|R)$$

$$= P\left\{ \min_{1 \leq i \leq k} SS_i \geq \frac{vC}{\delta_2^*} \right\}.$$

Let  $Z_j = \frac{1}{2}(SS_j - v - \eta_j' \eta_j) / (\frac{v}{2})^{\frac{1}{2}}$  Then

$$\begin{aligned} & P\left( \frac{SS_i}{2} \geq \frac{vC}{2\delta_2^*}, 1 \leq i \leq k \right) \\ &= P\left\{ Z_j \geq \frac{\sqrt{vC}}{\sqrt{2\delta_2^*}} - \left(\frac{v}{2}\right)^{\frac{1}{2}} - \frac{\eta_j' \eta_j}{(2v)^{\frac{1}{2}}}, 1 \leq j \leq k \right\} \\ &= P\left\{ Z_j \geq \frac{\sqrt{vC}}{\sqrt{2\delta_2^*}} - \left(\frac{v}{2}\right)^{\frac{1}{2}}, 1 \leq j \leq k \right\}. \end{aligned}$$

That is, the worst configuration (asymptotically) is when  $\underline{\beta} = \underline{0}$ . From the multivariate central limit theorem, it follows that for large  $v$ , the joint distribution of  $Z_1, \dots, Z_k$  does not depend on  $\eta_1, \dots, \eta_k$  (see [1]). Now the problem is the same as to compute the joint distribution of  $SS_1, \dots, SS_k$ . Note that here  $\Sigma = (\Sigma_{ij})$ ,  $\Sigma_{ij} = \delta_2^{*-1} B_i B_j'$  is  $v \times v$  as given in (4), and  $\Sigma_{ij} = \delta_2^{*-1} I$ .

Following the discussion in [1], we have the joint cumulant generating function of  $\frac{SS_j}{2}, 1 \leq j \leq k$ , is (see [5]).

$$\begin{aligned} (5) \quad \log |I - i\Sigma T| &= \frac{1}{2} \sum_{r=1}^{\infty} i^r \text{tr}(\Sigma T)^r / r \\ &= \frac{1}{2} \sum_{r=1}^{\infty} i^r C_r(t_1, \dots, t_k) / r. \end{aligned}$$

Thus, the joint cumulant  $K_{r_1, r_2, \dots, r_k}$  of total order  $r = r_1 + r_2 + \dots + r_k$ , can be obtained from the  $r$ th term of (5) by multiplying the coefficient of  $i^r (t_1^{r_1}) \dots (t_k^{r_k})$  by  $r_1! \dots r_k!$  Note that for  $r = 1, 2, 3$ ,

$$C_1 = \frac{n}{2} \sum_{j=1}^k t_j$$

$$(6) \quad C_2 = \frac{n}{2} \left\{ \sum_{j=1}^k t_j^2 + 2 \sum_{i < j} t_i t_j \delta_2^{*-2} \text{tr}(B_i B_j' B_j B_i') \right\}$$

and

$$C_3 = \frac{2n}{3} \left\{ \sum_{j=1}^k t_j^3 + 3 \sum_{i \neq j} t_i^2 t_j \delta_2^{*-2} \text{tr}(B_i B_j' B_j B_i') \right. \\ \left. + 6 \sum_{h < i < j} t_h t_i t_j \delta_2^{*-3} \text{tr}(B_h B_i' B_i B_j' B_j B_h') \right\}.$$

Expression (6) would determine an Edgeworth approximation of order  $v^{-\frac{1}{2}}$  [3]. To compute some constant  $C$  to satisfy

$$(7) \quad \inf P(CD|R) = P\{Z_j \geq \frac{\sqrt{c}}{\sqrt{2\delta_2^*}} - (\frac{v}{2})^{\frac{1}{2}}, 1 \leq j \leq k\} = P^*,$$

where  $Z_j = \frac{1}{\sqrt{2v}} (SS_j - v)$ ,  $1 \leq j \leq k$ , and the covariance matrix of the  $\{Z_j\}$  is given by  $\Gamma = (\rho_{ij})$ ,  $\rho_{ij} = v^{-1} \text{tr}(\Sigma_{ji} \Sigma_{ij})$ ,  $i \neq j$ .

The Fortran program as in [4] can be modified to compute (7).

Note that when  $\sigma_0^2$  is unknown, we can use the same method as above to construct a rule as follows:

$$R': \text{Reject } \pi_i \text{ (or reject } X_i \text{) if and only if } \frac{SS_i}{v} \geq \frac{c}{\delta_2^*} \frac{SS_0}{N-P}$$

where  $\delta_2^* < c < 1$  and

$$SS_0 = \underline{Y}' \{I - X(X'X)^{-1}X'\} \underline{Y} = \underline{Y}' Q_0 \underline{Y}.$$

Here  $SS_0$  is  $\chi_{N-p}^2$ .

Expected number of inferior populations included in the selected subset and its supremum.

For the proposed procedure the number  $T_1$  of inferior populations that enter into the selected subset is a random variable. For fixed values of  $k$  and  $P^*$ , the expected value of  $T_1$  is a function of  $\underline{\theta}$ .

For  $\underline{\theta} \in \Omega(t_1, t_2)$ , and large  $\nu$ ,

$$\begin{aligned} E_{\underline{\theta}}(T_1 | R) &= \sum_{i=k-t_1+1}^k P_{\underline{\theta}} \{SS(i) \leq \frac{\nu c}{2}\} \\ &\leq \sum_{i=k-t_1+1}^k P\left\{\frac{SS(i)}{\sigma^2[i]} < \nu c\right\} \\ &\approx \sum_{i=k-t_1+1}^k P\{SS(i) < \frac{\nu c}{2}\} \\ &= \sum_{i=k-t_1+1}^k P\left\{Z(i) < \frac{\sqrt{\nu c}}{\sqrt{2\delta_2^*}} - \left(\frac{\nu}{2}\right)^{\frac{1}{2}} - \frac{\eta'(j)\eta(j)}{(2\nu)^{\frac{1}{2}}}\right\} \\ &\leq \sum_{i=k-t_1+1}^k P\left\{Z(i) < \frac{\sqrt{\nu c}}{\sqrt{2\delta_2^*}} - \left(\frac{\nu}{2}\right)^{\frac{1}{2}}, 1 \leq j \leq k\right\} \end{aligned}$$

where  $Z(i)$  and  $\eta(j)$  are associated with  $\pi(i)$ ,  $1 \leq i \leq k$ . Thus the worst configuration is  $\beta = 0$ . Hence

$$\begin{aligned}
\sup_{\underline{\theta}} E_{\underline{\theta}}(T_1 | R) &= \max_{t_1, t_2} \sup_{\underline{\theta}} \sup_{\underline{\theta} \in \Omega(t_1, t_2)} E_{\underline{\theta}}(T_1 | R) \\
&= \sup_{\underline{\theta}} \sup_{\underline{\theta} \in \Omega(k, 1)} E_{\underline{\theta}}(T_1 | R) \\
&= \sum_{i=1}^k P\{Z_j < \frac{\sqrt{vc}}{\sqrt{2\delta_1^*}} - (\frac{v}{2})^{\frac{1}{2}}\}.
\end{aligned}$$

Expected number of superior populations that enter the selected subset and its infimum.

Let  $T_2$  denote the random number of superior populations that enter the selected subset. For  $\underline{\theta} \in \Omega(t_1, t_2)$  and for large  $v$ ,

$$\begin{aligned}
E_{\underline{\theta}}(T_2 | R) &= \sum_{i=1}^{t_2} P_{\underline{\theta}}\{SS(i) \leq \frac{vc}{\delta_1^*}\} \\
&= \sum_{i=1}^{t_2} P_{\underline{\theta}}\left\{\frac{SS(i)}{\sigma[i]} \leq \frac{1}{\sigma[i]} \cdot \frac{vc}{\delta_1^*}\right\} \\
&> \sum_{i=1}^{t_2} P\left\{\frac{SS(i)}{\sigma[i]} \leq vc\right\} \\
&\approx \sum_{i=1}^{t_2} P\{SS(i) \leq \frac{vc}{\delta_1^*}\} \\
&= \sum_{i=1}^{t_2} P\{Z(i) \leq \frac{\sqrt{vc}}{\sqrt{2\delta_1^*}} - (\frac{v}{2})^{\frac{1}{2}}\}.
\end{aligned}$$

Hence

$$\begin{aligned} \inf_{\theta} E_{\theta}(T_2|R) &= \min_{t_1, t_2} \inf_{\beta} \inf_{\theta \in \alpha(t_1, t_2)} E_{\theta}(T_2|R) \\ &= P\{Z_1 \leq \frac{\sqrt{vc}}{\sqrt{2\delta_1^*}} - (\frac{v}{2})^{\frac{1}{2}}\}, \end{aligned}$$

where

$Z_1 = \frac{1}{\sqrt{2v}} (SS_1 - v)$ ,  $\delta_1^* SS_1$  has chi-square  
with  $v$  degrees of freedom.

References

- [1] Arvesen, J. N. and McCabe, G. P. Jr. (1975). Subset selection problem for variances with applications to regression analysis. JASA 70, 166-170.
- [2] Arvesen, J. N. and McCabe, G. P. Jr. (1974). Variable selection in regression analysis. Proceedings of the University of Kentucky Conference On Regression with a Large Number of Predictors (Ed. W. O. Thompson and F. B. Cady). Dept. of Statist., Univ. of Kentucky, Lexington, Kentucky.
- [3] Chambers, J. M. (1967). On methods of asymptotic approximation for multivariate distributions. Biometrika 54, 367-383.
- [4] McCabe, G. P. Jr., Arvesen, J. N. and Pohl, R. J. (1973). A computer program for subset selection in regression analysis. Mimeo Series #317, Department of Statistics, Purdue University, West Lafayette, IN.
- [5] Searle, S. R. (1972). Linear Models. New York, John Wiley and Sons, Inc.
- [6] Spjøtvoll, E. (1972). Multiple comparison of regression functions. Ann. Math. Statist. 43, 1076-1088.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Mimeograph Series #501 ✓	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) On Selecting an Optimal Subset of Regression Variables	5. TYPE OF REPORT & PERIOD COVERED Technical	
	6. PERFORMING ORG. REPORT NUMBER Mimeo. Series #501	
7. AUTHOR(s) Gupta, S. S. and Huang, D. Y.	8. CONTRACT OR GRANT NUMBER(s) ONR N00014-75-C-0455 ✓	
9. PERFORMING ORGANIZATION NAME AND ADDRESS Purdue University Department of Statistics ✓ West Lafayette, IN 47907	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS	
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Washington, DC	12. REPORT DATE July 1977	
	13. NUMBER OF PAGES	
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report) Unclassified	
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE	
16. DISTRIBUTION STATEMENT (of this Report)  Approved for public release, distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Subset Selection Procedures, independent variables, noncentrality, Edgeworth approximation, superior and inferior variables.		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) In the past decade a number of methods have been developed for selecting the "best" or at least a "good" subset of variables in regression analysis. For various reasons, <sup>there</sup> we may be interested in including only a subset say, of size $r < p$ , the number of independent variables. Various authors have considered this problem and a variety of techniques are presently being used to construct such subsets. Most of these seem to lack justification in terms of statistical theory.		

→ next page

In this paper, we are interested in deriving a selection procedure to select a random size optimal subset such that all inferior independent variables are excluded. Some results on the efficiency of the procedure are also discussed.

Unclassified