On Selecting an Optimal Subset of Regression Variables*

by

Shanti S. Gupta, Purdue University

and

D. Y. Huang, Academia Sinica, Taipei

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #501

July 1977

*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.

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 these 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.

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.

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

Let $\pi_0, \pi_1, \ldots, \pi_k$ denote k+1 normal populations with variances $\sigma_0^2, \sigma_1^2, \ldots, \sigma_k^2$. Let $\sigma_{[1]}^2 \leq \cdots \leq \sigma_{[k]}^2$ denote the ordered variances. A population π_i is said to be

superior (or good) if $\sigma_0^2 \geq \delta_1^* \sigma_i^2$,

inferior (or bad) if $\sigma_0^2 \leq \delta_2^* \sigma_i^2$,

where δ_1^*, δ_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 that P*, a specified constant.

Let Ω be the parameter space which is the collection of all possible paramater vecotor $\underline{\theta} = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_k^2)$. Let t_1 and t_2 denote, respectively, the unknown number of inferior and superior popualtions 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 δ_1^* and δ_2^*, let

$$\Omega(t_1, t_2) = \{\underline{\theta}: \ \sigma_{[1]}^2 \leq \cdots \leq \sigma_{[t_2]}^2 \leq \frac{\sigma_0^2}{\delta_1^*} < \sigma_{[t_2+1]}^2$$

$$\leq \cdots \leq \sigma_{[k-t_1]}^2 < \frac{\sigma_0^2}{\delta_2^*} \leq \sigma_{[k-t_1+1]}^2 \leq \cdots \leq \sigma_{[k]}^2\}.$$

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) $$\underline{Y} = X\underline{\beta} + \underline{\epsilon}\,,\quad X = (\underline{1},\underline{X}_1,\ldots,\underline{X}_{p-1}),\quad \underline{\beta}' = (\beta_0,\beta_1,\ldots,\beta_{p-1}),$$

where X is an Nxp known matrix of rank $p \leq N$, $\underline{\beta}$ is a px1 parameter vector, and $\underline{\epsilon} \sim N(0,\sigma_0^2 I_N)$, and $\underline{1}' = (1, 1,\ldots,1)$.

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

Consider the models

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

where $X_{(i)} = (\underline{1},\underline{X}_1,\ldots,\underline{X}_{i-1},\underline{X}_{i+1},\ldots,\underline{X}_k)$ and $\underline{\beta}'_{(i)} = (\beta_0,\beta_1,\ldots,\beta_{i-1},\ldots,\beta_{i+1},\ldots,\beta_k)$, and $\underline{\epsilon}_i \sim N(0,\sigma_i^2 I_k)$, i=1,...,k. $X_{(i)}$ associated with model (2) is called population π_i $(1(1 \leq i \leq k)$. The goal is to reject π_i, i.e. to reject X_i, associated with $\sigma^2_{[j]}$, $j = k-t_1+1,\ldots,k$, for any fixed t_1.

Note that

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

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

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

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

(3) $$\sigma_i^2 = \sigma_0^2 + (X\underline{\beta})'Q_i(X\underline{\beta})/\nu.$$

Assume that σ_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 $SS_1,\ldots,SS_k$, we can write

$$\underset{\sim}{Y}'Q_i\underset{\sim}{Y} = U_i'U_i,$$

where

(4) $$U_i = B_iY \text{ and } B_iB_i' = I,\ B_i'B_i = Q_i$$

B_i is an νxN matrix.

The joint distribution of $U' = (U_1',\ldots,U_k')$ is multivariate normal in $k\nu$ dimensions with mean vector $\underset{\sim}{\eta}' = (\underset{\sim}{\eta}_1,\ldots,\underset{\sim}{\eta}_k)$, $\underset{\sim}{\eta}_i = B_iX\underset{\sim}{\beta}$, and covariance matrix $\Sigma = (\Sigma_{ij})$ where $\Sigma = B_iB_j'$. Note that the $k\nu$x$k\nu$ covariance matrix Σ is possibly singular. Let $\Sigma = FF'$ where F is of full column rank r ($r = \text{rank}(\Sigma)$), and let $U = \underset{\sim}{\eta} + FA$ where $A \sim N(\underset{\sim}{0},I_r)$. Thus, the joint characteristic function of $\frac{SS_1}{2},\ldots,\frac{SS_p}{2}$ is (since $SS_i = U_iU_i'$),

$$\varphi(t_1,\ldots,t_k) = E\{\exp(i\sum_{j=1}^{k} t_j(U_j)'U_j/2\}$$

$$= |I - iF'TF|^{-\frac{1}{2}}$$

$$\cdot \exp\tfrac{1}{2}[\underset{\sim}{\eta}'\{iT-TF(I-iF'TF)^{-1}F'T\}\underset{\sim}{\eta}]$$

$$= |I-i\Sigma T|^{-\frac{1}{2}}\exp\tfrac{1}{2}\{\underset{\sim}{\eta}'T(I-i\Sigma T)^{-1}\underset{\sim}{\eta}\},$$

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

We propose the rejection rule of the form:

R: Reject π_i (or reject X_i) is and only if

$$SS_i \geq \frac{\nu c}{\delta_2^*}$$

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

Note that SS_i is associated with U_i or, equivalently, with population π_i and degrees of freedom ν, 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 $\sigma^2_{[i]}$ and degrees of freedom ν. Thus

$$\inf P_{\underline{\theta}}(CD|R) = \inf P_{\underline{\theta}}\{\min_{k-t_1+1\leq i\leq k} SS_{(i)} \geq \frac{\nu c}{\delta_2^*}\}$$

$$= \inf P_{\underline{\theta}}\{\min_{k-t_1+1\leq i\leq k} \frac{SS_{(i)}}{\sigma^2_{[i]}} \geq \frac{\nu c}{\delta_2^*}\frac{1}{\sigma^2_{[i]}}\}$$

$$= \min_{0\leq t_1\leq k} \inf_{\underline{\beta}} P\{\min_{k-t_1+1\leq i\leq k} \frac{SS_{(i)}}{\sigma^2_{[i]}} \geq \nu c\}. \tag{4}$$

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

Thus

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

$$= P\{\min_{1\le i\le k} SS_i \ge \frac{\nu c}{\delta_2^*}\}.$$

Let $Z_j = \frac{1}{2}(SS_j - \nu - \underline{\eta}_j'\underline{\eta}_j)/(\frac{\nu}{2})^{\frac{1}{2}}$ Then

$$P(\frac{SS_i}{2} \ge \frac{\nu c}{2\delta_2^*}, \ 1\le i\le k)$$

$$= P\{Z_j \ge \frac{\nu c}{2\delta_2^*} - (\frac{\nu}{2})^{\frac{1}{2}} - \frac{\underline{\eta}_j'\underline{\eta}_j}{(\frac{\nu}{2})^{\frac{1}{2}}}, \ 1\le j\le k\}$$

$$\ge P\{Z_j \ge \frac{\nu c}{2\delta_2^*} - (\frac{\nu}{2})^{\frac{1}{2}}, \ 1\le j\le k\}.$$

That is, the worst configuration (asymptotically) is when $\underline{\beta} = \underline{0}$. From the multivariate central limit theorem, it follows that for large ν, the joint distribution of $Z_1,\dots,Z_k$ does not depend on $\underline{\eta}_1,\dots,\underline{\eta}_k$ (see [1]). Now the problem is the same as to compute the joint distribution of $SS_1,\dots,SS_k$. Note that here $\Sigma = (\Sigma_{ij})$, $\Sigma_{ij} = \delta_2^{*-1} B_iB_j'$ is $\nu\times\nu$ as given in (4), and $\Sigma_{ii} = \delta_2^{*-1} I$.

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

$$\log|I-i\Sigma T| = \frac{1}{2}\sum_{r=1}^{\infty} i^r tr(\Sigma T)^r/r \tag{5}$$

$$= \frac{1}{2}\sum_{r=1}^{\infty} i^r C_r(t_1,\dots,t_k)/r.$$

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

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

(6)
$$C^2 = \frac{n}{2} \{\sum_{j=1}^{k} t_j^2 + 2 \sum_{i<j} t_i t_j \bar{\delta}_*^2 tr(B_i B_j' B_j B_i')\}$$

and

$$C^3 = \frac{2n}{3} \{ \sum_{j=1}^{k} t_j^3 + 3 \sum_{i \neq j} t_i^2 t_j \bar{\delta}_*^2 tr(B_i B_j' B_j B_i')\}$$

$$+ 6 \sum_{h<i<j} t_h t_i t_j \bar{\delta}_*^3 tr(B_h B_i' B_i B_j B_j' B_h)\}.$$

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

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

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

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

Note that when σ_0^2 is unknown, we can use the same method as above to construct a rule as follows:

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

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 χ^2_{N-p}.

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 ν,

$$E_{\underline{\theta}}(T_1|R) = \sum_{i=k-t_1+1}^{k} P_{\underline{\theta}}\{SS_{(i)} \leq \frac{\nu c}{\delta_2^*}\}$$

$$\leq \sum_{i=k-t_1+1}^{k} P\{\frac{SS_{(i)}}{\sigma^2_{[i]}} < \nu c\}$$

$$\approx \sum_{i=k-t_1+1}^{k} P\{SS_{(i)} < \frac{\nu c}{\delta_2^*}\}$$

$$= \sum_{i=k-t_1+1}^{k} P\{Z_{(i)} < \frac{\nu c}{2\delta_2^*} - (\frac{\nu}{2})^{\frac{1}{2}} - \frac{\underline{\eta}'(j)\underline{\eta}(j)}{(\frac{\nu}{2})^{\frac{1}{2}}}\}$$

$$\leq \sum_{i=k-t_1+1}^{k} P\{Z_{(i)} < \frac{\nu c}{2\delta_2^*} - (\frac{\nu}{2})^{\frac{1}{2}},\ 1 \leq j \leq k\}$$

where $Z_{(i)}$ and $\underline{\eta}_{(i)}$ are associated with $\pi_{(i)}$, $1 \leq i \leq k$. Thus the worst configuration is $\underline{\beta} = \underline{0}$. Hence

$$\sup E_{\underline{\theta}}(T_1|R) = \max_{t_1,t_2} \sup_{\underline{\beta}} \sup_{\underline{\theta}\in\Omega(t_1,t_2)} E_{\underline{\theta}}(T_1|R)$$

$$= \sup_{\underline{\beta}} \sup_{\underline{\theta}\in\Omega(k,1)} E_{\underline{\theta}}(T_1|R)$$

$$= \sum_{i=1}^{k} P\{Z_j < \frac{\nu c}{2\delta_2^*} - (\frac{\nu}{2})^{\frac{1}{2}}\}.$$

<u>Expected number of superior populations that enter the selected subset and its infimum.</u>

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 ν,

$$E_{\underline{\theta}}(T_2|R) = \sum_{i=1}^{t_2} P_{\underline{\theta}}\{SS_{(i)} \leq \frac{\nu c}{\delta_1^*}\}$$

$$= \sum_{i=1}^{t_2} P_{\underline{\theta}}\{\frac{SS_{(i)}}{\sigma^2_{[i]}} \leq \frac{1}{\sigma^2_{[i]}} \cdot \frac{\nu c}{\delta_1^*}\}$$

$$> \sum_{i=1}^{t_2} P\{\frac{SS_{(i)}}{\sigma^2_{[i]}} \leq \nu c\}$$

$$\approx \sum_{i=1}^{t_2} P\{SS_{(i)} \leq \frac{\nu c}{\delta_1^*}\}$$

$$= \sum_{i=1}^{t_2} P\{Z_{(i)} \leq \frac{\nu c}{2\delta_1^*} - (\frac{\nu}{2})^{\frac{1}{2}}\}.$$

Hence

$$\inf E_{\underline{\theta}}(T_2|R) = \min_{t_1,t_2} \inf_{\underline{\beta}} \inf_{\underline{\theta}\in\Omega(t_1,t_2)} E_{\underline{\theta}}(T_2|R)$$

$$= P\{Z_1 \le \frac{\nu c}{2\delta_1^*} - (\frac{\nu}{2})^{\frac{1}{2}}\},$$

where

$$Z_1 = \frac{1}{\sqrt{2\nu}} (SS_1 - \nu),$$ $\delta_1^* SS_1$ has chi-square with ν 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. Mimeo Ser. #384, Department of Statistics, Purdue University, West Lafayette, IN.

[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.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

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, 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.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601 |

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

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.