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 NO0014-75-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 \( \pi_0, \pi_1, \ldots, \pi_k \) 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 \( \pi_i \) is said to be

- superior (or good) if \( \sigma_i^2 > \delta_i^* \sigma_i^2 \),
- inferior (or bad) if \( \sigma_i^2 < \delta_i^* \sigma_i^2 \),

where \( \delta_i^*, \delta_i^* \) are specified constants such that \( 0 < \delta_i^* < \delta_i^* < 1 \).

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 vectors \( \theta = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_k^2) \). Let \( t_1 \) and \( t_2 \) denote, respectively, the unknown number of inferior and superior 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 \( \delta_i^* \) and \( \delta_i^* \), let

\[
\Omega(t_1, t_2) = \{ \theta: \sigma_1^2 \leq \cdots \leq \sigma_{t_2}^2 \leq \delta_i^* \sigma_i^2 < \sigma_{t_2+1}^2 \leq \cdots \leq \sigma_{k-t_1}^2 \leq \sigma_0^2 \leq \sigma_{k-t_1+1}^2 \leq \cdots \leq \sigma_k^2 \}. \]

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
\[ Y = X \beta + \xi, \quad X = (I, X_1, \ldots, X_{p-1}), \quad \beta' = (\beta_0, \beta_1, \ldots, \beta_{p-1}), \]

where \( X \) is an \( N \times p \) known matrix of rank \( p \leq N \), \( \beta \) is a \( p \times 1 \) parameter vector, and \( \xi \sim N(0, \sigma^2 \xi I_N) \), and \( I' = (1, 1, \ldots, 1) \).

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

Consider the models

\[ Y = X(i) \beta(i) + \xi_i \]

where \( X(i) = (I, X_1, \ldots, X_{i-1}, X_i, \ldots, X_k) \) and \( \beta(i) = (\beta_0, \beta_1, \ldots, \beta_{i-1}, \beta_i, \ldots, \beta_k) \), and \( \xi_i \sim N(0, \sigma^2 \xi_i I_k) \), \( i=1, \ldots, k \). \( X(i) \) 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 \( \sigma^2 \xi_j, j = k-t_1+1, \ldots, k \), for any fixed \( t_1 \).

Note that

\[ SS_i = Y' [I - X(i) (X(i)' X(i))^{-1} X(i)'] Y = Y' Q_i Y, \]

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

\[ SS_i / \sigma^2 \sim \chi^2 \{ r(Q_i), (X_i)' Q_i (X_i) / (2\sigma^2) \}, \]

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

\[ \sigma^2_i = \sigma^2 + (X_i)' Q_i (X_i) / \nu. \]
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 $SS_1, \ldots, SS_k$, we can write

$$Y'Q_iY = U_i'U_i,$$

where

$$U_i = B_iY$$

and $B_iB_i = I$, $B_i'B_i = Q_i$.

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

The joint distribution of $U' = (U_1', \ldots, U_k')$ is multivariate normal in $k\omega$ dimensions with mean vector $\eta' = (\eta_1, \ldots, \eta_k)$, $\eta_1 = B_1'X_0$, and covariance matrix $\Sigma = (\Sigma_{ij})$ where $\Sigma = B_iB_j'$. Note that the $k\omega \times k\omega$ covariance matrix $\Sigma$ is possibly singular. Let $\Sigma = FF'$ where $F$ is of full column rank $r$ ($r = \text{rank}(\Sigma)$), and let $U = \eta + FA$ where $A = N(0, I_r)$. Thus, the joint characteristic function of $SS_1, \ldots, SS_k$ is (since $SS_i = U_i'U_i'$),

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

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

$$\cdot \exp\left\{\eta'\{iT-TF(I-iF'TF)^{-1}F'T\eta\}\right\}$$

$$= |I-i\Sigma T|^{-\frac{1}{2}}\exp\left\{\eta'T(I-i\Sigma T)^{-1}\eta\right\},$$
where \( T = \text{diag}(t_1, \ldots, t_k) \otimes I_v. \)

We propose the rejection rule of the form:

\[
R: \text{Reject } \pi_i \text{ (or reject } \chi_i \text{) is and only if } \quad SS_i \geq \frac{\nu c}{\delta^*_2}
\]

where \( \delta^*_2 < c < 1. \)

Note that \( SS_i \) is associated with \( U_i \) or, equivalently, with population \( \pi_i \) and degrees of freedom \( v \), 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 \( v \). Thus

\[
\inf P_{\theta}(CD|R) = \inf P_{\theta}\left\{ \min_{k-t_1+1 \leq i \leq k} SS_{(i)} \geq \frac{\nu c}{\delta^*_2} \right\} 
\]

\[
= \inf P_{\theta}\left\{ \min_{k-t_1+1 \leq i \leq k} \frac{SS_{(i)}}{\sigma^2[i]} \geq \frac{\nu c}{\delta^*_2} \frac{1}{\sigma^*[i]} \right\} 
\]

\[
\geq \min_{0 < t_1 < k} \inf P\left\{ \min_{\beta \leq k-t_1+1 < i \leq k} \frac{SS_{(i)}}{\sigma^2[i]} \geq \nu c \right\}. 
\]

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

\[
\inf P(CD|R) = P\left( \min_{1 \leq i \leq k} SS_i \geq \frac{\nu C}{\delta^2} \right).
\]

Let \( Z_j = \frac{1}{2}(SS_j - \nu - \nu \hat{\pi}_j) / (\frac{\nu}{2})^{\frac{3}{2}} \). Then

\[
P\left( \frac{SS_i}{2} \geq \frac{\nu C}{2\delta^2}, 1 \leq i \leq k \right)
\]

\[
= P\left( Z_j \geq \frac{\nu C}{2\delta^2} - \left(\frac{\nu}{2}\right)^{\frac{3}{2}} - \frac{n_i - \frac{1}{2}}{(\frac{\nu}{2})^2}, 1 \leq j \leq k \right)
\]

\[
\geq P\left( Z_j \geq \frac{\nu C}{2\delta^2} - \left(\frac{\nu}{2}\right)^{\frac{3}{2}}, 1 \leq j \leq k \right).
\]

That is, the worst configuration (asymptotically) is when \( \beta = 0 \).

From the multivariate central limit theorem, it follows that for large \( \nu \), the joint distribution of \( Z_1, \ldots, Z_k \) does not depend on \( \hat{\pi}_1, \ldots, \hat{\pi}_k \) (see [1]). Now the problem is the same as to compute the joint distribution of \( SS_1, \ldots, SS_k \). Note that here \( \Sigma = (\Sigma_{ij}) \),

\( \Sigma_{ij} = \delta \frac{1}{2} B_i B_j^t \) is \( \nu \times \nu \) as given in (4), and \( \Sigma_{ii} = \delta \frac{1}{2} \).

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]).

\[
(5) \quad \log |I - i\Sigma T| = \frac{1}{2} \sum_{r=1}^{\infty} \frac{i^r \text{tr}(\Sigma T)^r}{r}
\]

\[
= \frac{1}{2} \sum_{r=1}^{\infty} \frac{i^r C_r(t_1, \ldots, t_k)}{r}.
\]
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 $r$th term of (5) by multiplying the coefficient of $t_1^{r_1} \ldots (t_k^{r_k})$ 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)

$$C^2 = \frac{n}{2} \left( \sum_{j=1}^{k} t_j^2 + 2 \sum_{i<j} t_i t_j \delta_{ij} \text{tr}(B_i B_j B_i B_j) \right)$$

and

$$C^3 = \frac{2n}{3} \left( \sum_{j=1}^{k} t_j^3 + 3 \sum_{i<j} t_i^2 t_j \delta_{ij} \text{tr}(B_i B_j B_i B_j B_i B_j) \right)$$

$$+ 6 \sum_{h<i<j} t_h t_i t_j \delta_{ij} \text{tr}(B_h B_i B_j B_i B_j B_h).$$

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

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

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

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:

$R'$: Reject $\pi_1$ (or reject $X_1$) if and only if $\frac{SS_1}{\nu} \geq \frac{c}{\delta^2} \frac{SS_0}{N-P}$
where $\frac{\sigma_k}{2} < c < 1$ and

$$SS_0 = Y'(I - X(X'X)^{-1}X')Y = Y'Q_0Y.$$

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

**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 $\theta$.

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

$$E_\theta(T_1|R) = \sum_{i=k-t_1+1}^k P_\theta(SS(i) \leq \frac{\nu c}{\sigma^2})$$

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

$$= \sum_{i=k-t_1+1}^k P(SS(i) < \frac{\nu c}{\sigma^2})$$

$$= \sum_{i=k-t_1+1}^k P\left( Z(i) < \frac{\nu c}{2\sigma^2} - \left( \frac{\nu}{2} \right)^2 - \frac{n_i^j n(j)}{(\nu/2)^2} \right)$$

$$\leq \sum_{i=k-t_1+1}^k P\left( Z(i) < \frac{\nu c}{2\sigma^2} - \left( \frac{\nu}{2} \right)^{\frac{1}{2}}, 1 \leq j \leq k \right)$$

where $Z(i)$ and $n(i)$ are associated with $\pi(i)$, $1 \leq i \leq k$. Thus the worst configuration is $\theta = 0$. Hence
\[ \sup \mathbb{E}_\theta(T_1 | R) = \max_{t_1, t_2} \sup_{\theta \in \Omega(t_1, t_2)} \sup_{\theta \in \Omega(k, 1)} \mathbb{E}_\theta(T_1 | R) \]

\[ = \sup_{\theta \in \Omega(k, 1)} \mathbb{E}_\theta(T_1 | R) \]

\[ = \sum_{i=1}^k P(Z_j < \frac{\nu c}{2\delta^2} - \left(\frac{v}{2}\right)^\frac{1}{2}). \]

**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 \( \theta \in \Omega(t_1, t_2) \) and for large \( \nu \),

\[ \mathbb{E}_\theta(T_2 | R) = \sum_{i=1}^{t_2} \mathbb{P}_\theta \{ SS(i) \leq \frac{\nu c}{\delta^2} \} \]

\[ = \sum_{i=1}^{t_2} \mathbb{P}_\theta \{ \frac{SS(i)}{\sigma[i]} \leq \frac{1}{2} \cdot \frac{\nu c}{\delta^2} \} \]

\[ > \sum_{i=1}^{t_2} \mathbb{P} \left\{ \frac{SS(i)}{\sigma[i]} \leq \nu c \right\} \]

\[ = \sum_{i=1}^{t_2} \mathbb{P} \{ SS(i) \leq \frac{\nu c}{\delta^2} \} \]

\[ = \sum_{i=1}^{t_2} \mathbb{P} \{ Z(i) \leq \frac{\nu c}{2\delta^2} - \left(\frac{v}{2}\right)^\frac{1}{2} \}. \]

Hence
\[
\inf_{\theta} E_{\theta}(T_2 | R) = \min_{t_1, t_2} \inf_{\beta} \inf_{\theta \in \mathcal{A}(t_1, t_2)} E_{\theta}(T_2 | R)
= P\{Z_1 \leq \frac{v \sigma_1}{2\delta_1^2} - (\frac{v}{2})^{\frac{1}{2}}\},
\]

where

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


**REPORT DOCUMENTATION PAGE**

<table>
<thead>
<tr>
<th>1. REPORT NUMBER</th>
<th>2. GOVT ACCESSION NO.</th>
<th>3. RECIPIENT'S CATALOG NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mimeo Series #501</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4. TITLE (and Subtitle)</th>
<th>5. TYPE OF REPORT &amp; PERIOD COVERED</th>
</tr>
</thead>
<tbody>
<tr>
<td>On Selecting an Optimal Subset of Regression Variables</td>
<td>Technical</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7. AUTHOR(s)</th>
<th>8. CONTRACT OR GRANT NUMBER(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gupta, S. S. and Huang, D. Y.</td>
<td>ONR N00014-75-C-0455</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>9. PERFORMING ORGANIZATION NAME AND ADDRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purdue University</td>
</tr>
<tr>
<td>Department of Statistics</td>
</tr>
<tr>
<td>West Lafayette, IN 47907</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>11. CONTROLLING OFFICE NAME AND ADDRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Office of Naval Research</td>
</tr>
<tr>
<td>Washington, DC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>13. NUMBER OF PAGES</th>
</tr>
</thead>
<tbody>
<tr>
<td>July 1977</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>15. SECURITY CLASS. (of this report)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unclassified</td>
</tr>
</tbody>
</table>

**DISTRIBUTION STATEMENT (of this Report)**

Approved for public release, distribution unlimited.

**DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)**

**SUPPLEMENTARY NOTES**

**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.

**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.
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.