Selecting Important Independent Variables in Linear Regression Models*

by

Shanti S. Gupta
Purdue University

Deng-Yuan Huang
National Taiwan Normal University

Technical Report #86-29

Department of Statistics
Purdue University

Revised November 1986

* This research was supported by the Office of Naval Research Contract N00014-84-C-0167 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Selecting Important Independent Variables in Linear Regression Models*

by

Shanti S. Gupta
Purdue University

Deng-Yuan Huang
National Taiwan Normal University

Abstract

A large body of literature exists on the techniques for selecting the important variables in linear regression analysis. Many of these techniques are ad hoc in nature and have not been studied from a theoretical viewpoint. In this paper we discuss some of the more commonly used techniques and propose a selection procedure based on the statistical selection and ranking approach. This procedure is easy to compute and apply. The procedure depends on the goodness of fit of the model and the total error associated with it.

AMS 1980 Subject Classification: Primary 62F07, Secondary 62J05

Key Words: Selection procedures; Noncentrality parameters; Noncentral F; Total square error; Reduced model; Inferior models; Selection criteria.

* This research was supported by the Office of Naval Research Contract N00014-84-C-0167 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Selecting Important Independent Variables in Linear Regression Models*

by

Shanti S. Gupta
Purdue University

Deng-Yuan Huang
National Taiwan Normal University

Abstract

A large body of literature exists on the techniques for selecting the important variables in linear regression analysis. Many of these techniques are ad hoc in nature and have not been studied from a theoretical viewpoint. In this paper we discuss some of the more commonly used techniques and propose a selection procedure based on the statistical selection and ranking approach. This procedure is easy to compute and apply. The procedure depends on the goodness of fit of the model and the total error associated with it.

Key Words: Selection procedures; Noncentrality parameters; Noncentral F; Total square error; Reduced model; Inferior models; Selection criteria.

* This research was supported by the Office of Naval Research Contract N00014-84-C-0167 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

1. Introduction

The problem of determining the important ("best") subset of independent variables has long been of interest to applied statisticians: primarily, because of the current availability of high-speed computations, this problem has received considerable attention in the recent statistical literature. Hocking (1976) has given an excellent survey of the existing techniques. Several other papers have dealt with various aspects of the problem but it appears that the typical regression user has not benefited appreciably. One reason for the lack of resolution of the problem is the fact that it has not been well defined. For the procedures that we usually discussed in textbooks, the probability of a correct selection is not guaranteed.

The problem of selecting a subset of independent or predict variables is usually described in an idealized setting. That is, it is assumed that

(1) the analyst has data on a large number of potential variables which include all relevant variables and appropriate functions of them plus, possibly, some other extraneous variables and variable functions and

(2) the analyst has available "good" data on which to base the eventual conclusions.

The analysis of residuals (see Draper and Smith (1981)) may reveal different functional forms which might be considered and may even suggest variables which are not initially included. We assume that the process for model building has been completed and the resulting models are true. The problem is to determine an "appropriate" regression model based on a subset of the original set of variables. In this problem there are three ingredients, namely,

(a) the computational technique(s) used to provide the information for the analysis,
(b) the criterion used to analyze the variables and select a subset, if that is appropriate, and
(c) the estimation of the coefficients in the final equation. (cf. Hocking (1976, 1983)).

In this paper, we study this problem from the viewpoint of statistical ranking and selection to investigate some selection criteria. From this approach we can obtain some useful procedures to select important regression variables.

In these studies, we have found that the reduced models are based on noncentrality parameters which provide a measure of goodness of fit for the fitted models. We also propose a statistic to measure the standardized total square error, and study the detection of bias for the fitted model. The statistic we propose is an unbiased estimator which is different from Mallows' Cp statistic. Based on this statistic, a two-stage selection procedure is proposed and studied.

Finally, we mention that we have shown the relation of the noncentrality parameters and the statistic we proposed. We should use both of them to select a good fit and less bias models and at the same time, the total square error is also to be made as small as possible. An asymptotic result is also studied to determine the value Δ of the bias. This asymptotic result enables us to determine at least how many regression variables will be neglected.

2. Some Selection Criteria

Consider the usual linear model

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \tag{2.1}$$

where $\underline{Y}' = [Y_1, \ldots, Y_n]$ is an $1 \times n$ vector of a random sample, $X = [\underline{1}, \underline{X}_1, \ldots, \underline{X}_{p-1}]$ is an $n \times p$ matrix of known constants, $\underline{\beta}' = [\beta_0, \beta_1, \ldots, \beta_{p-1}]$ is a $1 \times p$ vector of unknown parameters and $\underline{\varepsilon} \sim N(\underline{0},\ \sigma_0^2 I_n)$. Here I_n denotes the identity matrix of order $n \times n$. The model (2.1) having $p-1$ independent variables is considered as the true model. Any reduced model whose "X matrix" has r columns is obtained by retaining any $r-1$ of the $p-1$ independent variables $X_1, \ldots X_{p-1}$, where $2 \leq r \leq p$. For each r, $2 \leq r \leq p$, there are $k_r = \binom{p-1}{r-1}$ such models. These k_r reduced models of "size" r are indexed arbitrarily with the indexing variable i going from 1 to k_r. We will refer to a typical model as Model M_{ri}. A reduced model of size r can be written as

$$E(\underline{Y}) = X_{ri}\, \underline{\beta}_{ri}, \quad i = 1, 2, \ldots, k_r \tag{2.2}$$

where X_{ri} and $\underline{\beta}_{ri}$ are obtained from X and $\underline{\beta}$ corresponding to the variables that are retained in the model.

It should be pointed out that all expectations and probabilities are calculated under the true model (2.1).

Usually, we use the residual sum of squares to measure goodness of the fitted model for a random sample. Hence, the expected residual sum of squares is naturally considered as the measurement for the goodness of fit. Large values of this expectation are not desirable. It should be first noted that our comparisons of models are made under the true model assumptions.

For any r, $2 \leq r \leq p$, the residual sum of square SS_{ri} for the reduced model M_{ri}, $1 \leq i \leq k_r$, is as follows

$$\begin{aligned} SS_{ri} &= \underline{Y}'[I - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}]\underline{Y} \\ &= \underline{Y}'Q_{ri}\underline{Y}, \end{aligned} \tag{2.3}$$

where $Q_{ri} = [I - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}]$. Also,

$$\frac{SS_{ri}}{\sigma_0^2} \sim \chi^2\{\nu_r, \lambda_{ri}\}, \tag{2.4}$$

where the degrees of freedom $\nu_r = n - r$, and the noncentrality parameter

$$\lambda_{ri} = (X\underline{\beta})'\ Q_{ri}(X\underline{\beta})/2\sigma_0^2.$$

We note that Q_{ri} is idempotent and symmetric; thus it is positive semi-definite. Hence λ_{ri} is nonnegative, but not zero, in general.

We have

$$E[SS_{ri}] = \nu_r\ \sigma_0^2 + 2\sigma_0^2\ \lambda_{ri}. \tag{2.5}$$

Since σ_0^2 is fixed, it is clear from (2.5) that λ_{ri} should not be large for a good model.

We define any reduced model with associated noncentrality parameter λ_{ri} inferior if $\lambda_{ri} \geq \Delta$ where $\Delta(> 0)$ is a specified constant. Our goal is to eliminate all inferior models from the set of $2^{p-1} - 1$ regression models including the true model.

The residual sum of squares for the full model is denoted by SS_{p1}. Then,

$$E[SS_{p1}/(n-p)] = \sigma_0^2.$$

Hence, we use $SS_{p1}/(n-p)$ to estimate σ_0^2, and denote

$$\hat{\sigma}_0^2 = \frac{SS_{p1}}{n-p}.$$

Now, let R^2 and R^2_{ri} denote the multiple correlation coefficients of the models (2.1) and (2.2), respectively. Hence

$$R^2 = 1 - \frac{SS_{p1}}{(\underline{Y} - \underline{\bar{Y}})'(\underline{Y} - \underline{\bar{Y}})},$$

and

$$R^2_{ri} = 1 - \frac{SS_{ri}}{(\underline{Y} - \underline{\bar{Y}})'(\underline{Y} - \underline{\bar{Y}})},$$

where $\underline{\bar{Y}}' = (\bar{Y}, \ldots, \bar{Y})$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. From (2.5), we propose $\hat{\lambda}_{ri}$ as an estimator of λ_{ri} where

$$\hat{\lambda}_{ri} = \frac{n-p}{2}\frac{SS_{ri}}{SS_{p1}} - \frac{\nu_r}{2} \tag{2.6}$$

$$= \frac{n-p}{2}\frac{1-R^2_{ri}}{1-R^2} - \frac{\nu_r}{2}. \tag{2.7}$$

Proposed Selection Procedure.

We propose the selection rule S as follows:

Exclude (reject) the reduced model M_{ri}

$$\text{iff } \hat{\lambda}_{ri} \geq d_{ri}$$

where d_{ri} is determined by $\inf P\{\hat{\lambda}_{ri} \geq d_{ri}\} = P^*, 0 < P^* < 1$. It can be shown that the following are equivalent forms:

$$\hat{\lambda}_{ri} \geq d_{ri} \tag{2.8}$$

$$\Longleftrightarrow (1 - R^2_{ri}) \geq (d_{ri} + \frac{\nu_r}{2})\frac{2}{n-p}(1 - R^2) \tag{2.9}$$

$$\Longleftrightarrow \frac{(SS_{ri} - SS_{p1})/(p-r)}{SS_{p1}/(n-p)} \geq \left[(d_{ri} + \frac{\nu_r}{2})\frac{2}{n-p} - 1\right]\frac{n-p}{p-r}. \tag{2.10}$$

Hence, the correct decision of excluding all inferior models M_{ri} under the guaranteed probability P^* is equivalent to

$$\inf_{\lambda_{ri} \geq \Delta} P\{\hat{\lambda}_{ri} \geq d_{ri}\} = P^*. \tag{2.11}$$

It is well known that the distribution of the statistic

$$V_{ri} = \frac{(SS_{ri} - SS_{p1})/(p-r)}{SS_{p1}/(n-p)}$$

follows the noncentral F distribution denoted as $F'(p-r, n-p, \lambda_{ri})$ (cf. Graybill (1976)). Thus the critical value d_{ri} in (2.11) can be computed as follows:

$$\inf_{\lambda_{ri} \geq \Delta} P\{V_{ri} \geq D_{ri}\} = P^* \qquad (2.12).$$

From Ghosh (1973), the noncentral F distribution is stochastically decreasing in λ_{ri}. Thus we can compute d_{ri} through the following equation:

$$P\{V_{ri} \geq D_{ri} | \lambda_{ri} = \Delta\} = P^*, \qquad (2.13)$$

where $D_{ri} = \left[(d_{ri} + \frac{\nu_r}{2})\frac{2}{n-p} - 1\right]\frac{n-p}{p-r}$ as in (2.10).

Now, we rewrite the selection procedure S as follows.

Theorem 1. The selection procedure S is equivalent to the following:

Exclude M_{ri} as an inferior model

$$\text{iff } V_{ri} \geq D_{ri}$$

where D_{ri} depends on Δ, n, p, r and P^* and is chosen to satisfy

$$P\{V_{ri} \geq D_{ri} | \lambda_{ri} = \Delta\} = P^*.$$

Total Squared Error as a Criterion for Goodness of Fit.

A measure of "total squared error" was first given by Mallows (1973). He used the statistic, called C_p, to measure the sum of the squared biases plus the squared random errors in Y at all n data points. Daniel and Wood (1980) described the problem as follows.

The total squared error (bias plus random) for n data points, using a fitted model M_{ri} with r terms, is $\sum_{j=1}^{n} E(\hat{Y}_{ij} - \nu_j)^2$, i.e.,

$$\sum_{j=1}^{n}(\nu_j - \eta_{ij})^2 + \sum_{j=1}^{n} \text{var } (\hat{Y}_{ij}) \qquad (2.14)$$

where

$\nu_j = \nu(X_{1j}, X_{2j}, \ldots)$, expected value from true equation,
$\eta_{ij} = \beta_0 + \sum_{\ell=1}^{r-1} \beta_\ell X_{i\ell}$, expected value from the fitted model M_{ri} being used,
$(\nu_{ij} - \eta_{ij})$ = bias at the j*th* data point, and
$\underline{\hat{Y}}_i = (\hat{Y}_{i1}, \ldots, \hat{Y}_{in})' = X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}\underline{Y}$ is the predicted value under least square estimate in the reduced model M_{ri}.

For convenience, let SSB_{ri} stand for $\sum_{j=1}^{n}(\nu_j - \eta_{ij})^2$ and define a quantity, Γ_{ri}, the standardized total squared error, by

$$\Gamma_{ri} = \frac{SSB_{ri}}{\sigma_0^2} + \frac{1}{\sigma_0^2}\sum_{j=1}^{n} \text{var } (\hat{Y}_{ij}). \tag{2.15}$$

Since

$$\underline{\hat{Y}}_i = X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}\underline{Y},$$

we have

$$\begin{aligned}
&\text{cov } (\underline{\hat{Y}}_i) \\
&\quad = E[X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}\underline{Y} - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}X\underline{\beta}] \times \\
&\quad\qquad [X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}\underline{Y} - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}X\underline{\beta}]' \\
&\quad = \sigma_0^2 X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}.
\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
\sum_{j=1}^{n} \text{var } (\hat{Y}_{ij}) &= \sigma_0^2 \text{ tr } X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri} \\
&= r\sigma_0^2;
\end{aligned} \tag{2.16}$$

$$\begin{aligned}
SSB_{ri} &= \sum_{j=1}^{n}(\nu_j - \eta_{ij})^2 = \sum_{j=1}^{n}\left[E(Y_j) - E(\hat{Y}_{ij})\right]^2 \\
&= \left[E(\underline{Y} - \underline{\hat{Y}}_i)\right]'\left[E(\underline{Y} - \underline{\hat{Y}}_i)\right] \\
&= \{E([I - X_{ri}(X'_{ri}X_{ri})^{-1}X_{ri}]\underline{Y})\}'\{E([I - X_{ri}(X'_{ri}X_{ri})^{-1}X_{ri}]\underline{Y})\} \\
&= (X\underline{\beta})'[I - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}](X\underline{\beta}) \\
&= 2\sigma_0^2\lambda_{ri};
\end{aligned} \tag{2.17}$$

and

$$\begin{aligned}
E(SS_{ri}) &= E(\underline{Y} - \hat{\underline{Y}}_i)'(\underline{Y} - \hat{\underline{Y}}_i) \\
&= E\{\underline{Y}'[I - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}]\underline{Y}\} \\
&= E\{\text{ tr } (\underline{Y}'[I - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}]\underline{Y})\} \\
&= \text{ tr } \{E[(\underline{Y} - X\underline{\beta})'(I - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri})(\underline{Y} - X\underline{\beta})] \\
&\qquad + (X\underline{\beta})'[I - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}](X\underline{\beta})\} \\
&= \text{ tr } \{[I - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}]\sigma_0^2\} \\
&\qquad + (X\underline{\beta})'[I - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}](X\underline{\beta})\} \\
&= (n-r)\sigma_0^2 + 2\sigma_0^2\lambda_{ri}. \qquad (2.18)
\end{aligned}$$

From (2.17) and (2.18), we have

$$E(SS_{ri}) = SSB_{ri} + (n-r)\sigma_0^2. \qquad (2.19)$$

From (2.16) and (2.19), we can rewrite (2.15) as

$$\begin{aligned}
\Gamma_{ri} &= \frac{E(SS_{ri})}{\sigma_0^2} - (n-2r) \\
&= \nu_r + 2\lambda_{ri} - (n-2r) \\
&= 2\lambda_{ri} + r. \qquad (2.20)
\end{aligned}$$

We have an unbiased estimate of Γ_{ri} as follows:

$$\hat{\Gamma}_{ri} = 2 \cdot \frac{n-p-2}{n-p}\left[2\hat{\lambda}_{ri} + (n-r)\right] - (2p-3r) \qquad (2.21)$$

since

$$2\hat{\lambda}_{ri} + (n-r) = (p-r)V_{ri}.$$

Hence, we can show that for $n-p > 2$, we have

$$\begin{aligned}
E(\hat{\Gamma}_{ri}) &= 2\frac{n-p-2}{n-p}(p-r)\frac{\{(p-r)+\lambda_{ri}\}\,(n-p)}{(p-r)(n-p-2)} - (2p-3r) \\
&= 2\lambda_{ri} + r = \Gamma_{ri} \qquad (2.22)
\end{aligned}$$

From (2.5), we see that λ_{ri} is a measure of the error. That is, λ_{ri} is used to measure the fitness of the reduced model M_{ri}. If it is a good fit then $\lambda_{ri} \approx 0$. Then from (2.20), we have

$$\Gamma_{ri} \approx r.$$

(Note the notation $\approx$ means "approximately close".) We require the total square error to be small for good fit. Hence $\hat{\Gamma}_{ri}$ should be as small as possible and close to r. Hence λ_{ri} should be as small as possible.

We summarize these results in the following theorem.

Theorem 2. The total squared error (bias plus random) for n data points, using a fitted model M_{ri} with r terms, as defined by Mallows (1973) (see also Daniel and Wood (1980)) is

$$\sum_{j=1}^{n}(\hat{Y}_{ij}-\nu_j)^2,$$

where $\nu_j = E(Y_j)$ and $\underline{\hat{Y}}_i = (\hat{Y}_{i1},\ldots,\hat{Y}_{in})' = X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}\underline{Y}$. Now from (2.21), as $n-p>2$,

$$\hat{\Gamma}_{ri} = 2\cdot\frac{n-p-2}{n-p}[2\hat{\lambda}_{ri}+(n-r)]-(2p-3r)$$

is an unbiased estimator of the standardized total squared error Γ_{ri}. Also if $SSB_{ri}\approx 0$, then $\Gamma_{ri}\approx r$.

The Relation between R^2_{ri} and $\hat{\Gamma}_{ri}$

From (2.7) and (2.21), we have

$$\hat{\Gamma}_{ri} = 2(n-p-2)\frac{1-R^2_{ri}}{1-R^2}-(2n-3r-4). \tag{2.23}$$

Hocking (1976) pointed out that the R_{ri} plot may be quite flat for a given range on r; the coefficient $(n-p-2)$ can magnify small differences causing $\hat{\Gamma}_{ri}$ to increase dramatically as r is decreased.

The Relation between F-statistic V_{ri} and $\hat{\Gamma}_{ri}$

From Theorem 2, we have

$$\hat{\Gamma}_{ri} = \frac{2(n-p-2)(p-r)}{n-p}V_{ri}-(2p-3r). \tag{2.24}$$

The Relation between Mallows' C_{ri} and $\hat{\Gamma}_{ri}$

Mallows' C_{ri} is defined as follows:

$$C_{ri} = (n-r)V_{ri} - (n-2r).$$

Hence

$$\hat{\Gamma}_{ri} = \frac{2(n-p-2)(p-r)}{(n-p)(n-r)}[C_{ri} + (n-2r)] - (2p-3r). \tag{2.25}$$

3. A Two-Stage Selection Procedures R_s

Now we propose the following selection procedure which depends on the procedure S.

R_s: At stage 1, apply the selection procedure S to select some desirable reduced models denoted by the set T. At stage 2, from the set T, we select the reduced model associated with the smallest $\hat{\Gamma}_{ri}$.

From (2.8), (2.9), and (2.10), we see that the following selection rules S_1 and S_2 are all equivalent to S.

S_1: select model M_{ri} if $(1 - R^2_{ri}) \geq d_1(1 - R^2)$;
S_2: select model M_{ri} if $V_{ri} \geq d_2$;

where d_1 and d_2 depend on n, p, r, i and P^*.

Gupta, Huang and Chang (1984) have studied some optimal properties of S_2. Huang and Panchapakesan (1982) have studied some selection procedures related to S_1. S_2 can be used in the stepwise regression analysis. Also S_1 can be used for analysis of all possible regression models.

From the previous discussions, one can use S_2 to compute the critical values d_2 to decide the acceptance or rejection of the reduced models. From the selected models we choose a suitable one by plotting $\hat{\Gamma}_{ri}$ against r with $\hat{\Gamma}_{ri}$ as small as possible (see Theorem 2). It follows from the fact that $SSB_{ri}/\sigma_0^2 = 2\lambda_{ri}$, that the large values of λ_{ri} measure the degree of the departure from the line $\hat{\Gamma}_{ri} = r$.

Computation of Constants D_{ri}

Patnaik (1949) provided an approximation to the noncentral F distribution (cf. Guenther (1979)) by the relation

$$F(p_1, p_2, \lambda) \approx [(p_1 + 2\lambda)/p_1]F(p^*, p_2) \tag{3.1}$$

where

$$p^* = (p_1 + 2\lambda)^2/(p_1 + 4\lambda).$$

Hence, we can determine D_{ri} from the following (approximation) equation (see Theorem 1):

$$P\{F(p^*, n-p) \geq \left[\frac{p-r}{(p-r)+2\Delta}\right] D_{ri}\} = P^*, \tag{3.2}$$

where $p^* = \frac{[(p-r)+2\Delta]^2}{(p-r)+4\Delta}$, and $F(p^*, n-p)$ is the statistic which follows the central F distribution with p^* and $n-p$ degrees of freedom.

Ghosh (1973) has shown that $P\{F(p_1, p_2) \geq c\}$ is monotone decreasing in p_1 and increasing in p_2.

Thus we can use interpolation method to obtain the critical value $c = F(p^*, n-p; P^*)$ by noting the fact that $F(p^*, n-p; P^*) = [F(n-p, p^*; 1-P^*)]^{-1}$. Note that

$$D_{ri} = \frac{(p-r)+2\Delta}{p-r} F(p^*, n-p; P^*). \tag{3.3}$$

Asymptotic Results for R_s

Note that procedure R_s at the first stage satisfies (2.13). Suppose we want to determine the (minimum) number of independent variables to be chosen for a specified value of Δ. Assuming the sample size n to be sufficiently large, we study the asymptotic results for the two-stage selection procedure R_s. Let $n-p > 4$.

$$\begin{aligned} P^* &= P\{V_{ri} \geq D_{ri} | \lambda_{ri} = \Delta\} \\ &= P\{\frac{V_{ri} - E(V_{ri})}{\sqrt{\operatorname{Var}(V_{ri})}} \geq \frac{D_{ri} - E(V_{ri})}{\sqrt{\operatorname{Var}(V_{ri})}} | \lambda_{ri} = \Delta\} \\ &\approx P\{Z \geq \alpha\} \ = \ 1 - \Phi(\alpha), \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal distribution function,

$$E(V_{ri}) = \frac{(p-r+2\Delta)(n-p)}{(n-p-2)(p-r)},$$

and

$$\operatorname{Var}(V_{ri}) = \frac{2(n-p)^2}{(p-r)^2(n-p-2)} \left[\frac{((p-r)+2\Delta)^2}{(n-p-2)(n-p-4)} + \frac{(p-r)+4\Delta}{n-p-4}\right].$$

For a fixed random sample, we have

$$\hat{\Gamma}_{ri} = \frac{2(n-p-2)(p-r)}{n-p} V_{ri} - (2p-3r).$$

Now, we rewrite $\hat{\Gamma}_{ri}$ as follows:

$$\hat{\Gamma}_{ri} = \frac{2(n-p-2)(p-r)}{n-p} \left\{ \sqrt{\mathrm{Var}(V_{ri})} \left[\frac{V_{ri} - E(V_{ri})}{\sqrt{\mathrm{Var}(V_{ri})}} \right] + E(V_{ri}) \right\} - (2p-3r).$$

We are trying to minimize the following function $\hat{\Gamma}^{\alpha}_{ri}$ with $\hat{\Gamma}^{\alpha}_{ri} \leq \hat{\Gamma}_{ri}$ to obtain an upper bound of Δ, for the given value $p - r = x$ and $\alpha < 0$.

Let

$$\begin{aligned}
\hat{\Gamma}^{\alpha}_{ri} &= \frac{2(n-p-2)(p-r)}{n-p} \\
&\qquad \left\{ \alpha \sqrt{\frac{2(n-p)^2}{(p-r)^2(n-p-2)}} \sqrt{\frac{(p-r+2\Delta)^2}{(n-p-2)(n-p-4)} + \frac{p-r+4\Delta}{n-p-4}} \right. \\
&\quad \left. + \frac{(p-r+2\Delta)(n-p)}{(n-p-2)(p-r)} \right\} - (2p-3r) \\
&= \sqrt{2} D \alpha [Ax^2 + Bx + C]^{\frac{1}{2}} - x + 4\Delta + p,
\end{aligned}$$

where

$$\begin{aligned}
A &= \frac{(n-p)^2}{(n-p-2)^2(n-p-4)}; \\
B &= \frac{4\Delta(n-p)^2}{(n-p-2)^2(n-p-4)} + \frac{(n-p)^2}{(n-p-4)(n-p-2)}; \\
C &= \frac{4\Delta^2(n-p)^2}{(n-p-2)^2(n-p-4)} + \frac{4\Delta(n-p)^2}{(n-p-4)(n-p-2)};
\end{aligned}$$

and

$$D = \frac{2(n-p-2)}{n-p}.$$

Since $A \approx 0$, $B \approx 1$, $C \approx 4\Delta$ and $D \approx 2$, hence, $\hat{\Gamma}^{\alpha}_{ri} \approx 2\sqrt{2}\alpha\ \sqrt{x+4\Delta} - x + p + 4\Delta$. By letting $\frac{d\hat{\Gamma}^{\alpha}_{ri}}{d\Delta} = 0$, we have $\Delta \approx 2\alpha^2 - x$, such that $\frac{d^2\hat{\Gamma}^{\alpha}_{ri}}{d\Delta^2} > 0$. Hence, $\hat{\Gamma}^{\alpha}_{ri}$ is minimized when $\Delta \approx 2\alpha^2 - x$. For which, we can find an upper bound of Δ such that at least how many variables are excluded for this bound, since Δ is decreasing in x; see the following example.

Example:

We use the Hald data (Draper and Smith, 1981, Appendix B, page 629) to discuss the procedure as follows.

No.	X_1	X_2	X_3	X_4	Y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

Daniel and Wood (1980, p. 89) have computed C_{ri}'s for all equations. Using their values, we compute $\hat{\Gamma}_{ri}$ as follows:

	Variables in Equation	r	C_{ri}	$\hat{\Gamma}_{ri}$	V_{ri}
i	X_1	2	202.7	82.6	19.25
	X_2	2	142.6	58.0	13.78
	X_1, X_2	3	2.7	1.9	0.97*
	X_3	2	315.3	128.6	29.48
	X_1, X_3	3	198.2	60.6	20.52
	X_2, X_3	3	62.5	19.9	6.95
	X_1, X_2, X_3	4	3.0	3.3	0.89*
	X_4	2	138.8	56.5	13.44
	X_1, X_4	3	5.5	2.8	1.25
	X_2, X_4	3	138.3	42.6	1.45
	X_1, X_2, X_4	4	3.0	3.3	0.89*
	X_3, X_4	3	22.4	7.8	2.94
	X_1, X_3, X_4	4	3.5	3.4	0.94*
	X_2, X_3, X_4	4	7.3	4.1	1.37*
	X_1, X_2, X_3, X_4	5	5.0	5.0	1.00*

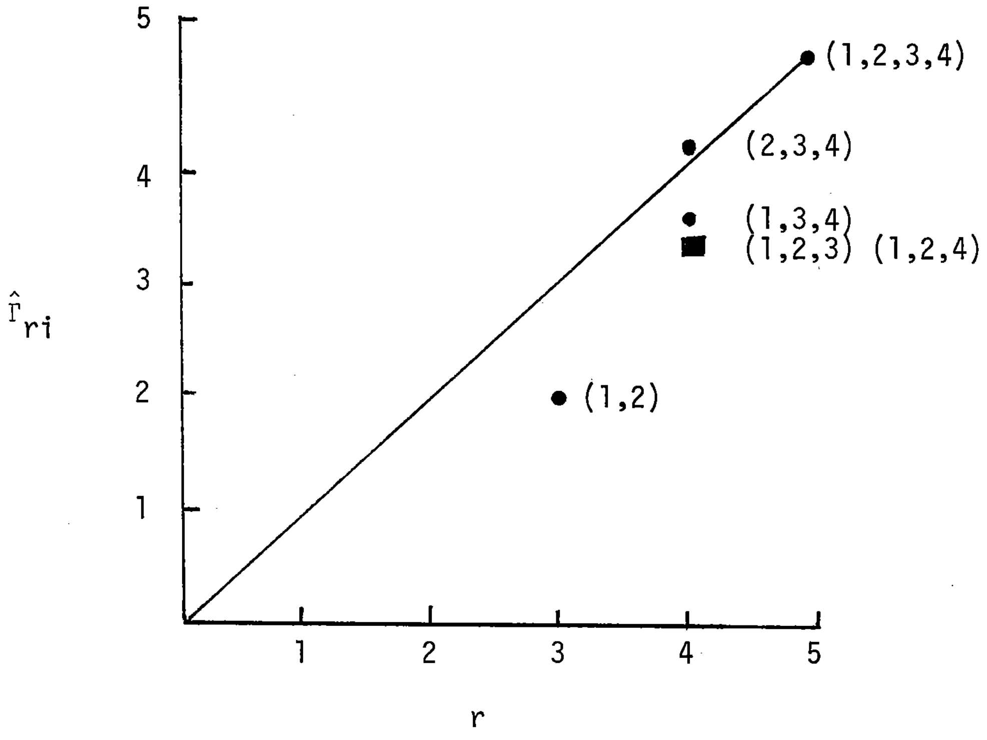
As an illustrative example, we compute some D_{ri}'s in (3.3) for $P^* = 0.90$, and $\Delta = 3$ as follows: $n = 13, p = 5$.

r	2	3	4
D_{ri}	0.939	1.1106	1.647

Now we apply the procedure R_s. At stage 1, we exlude all inferior reduced models. This results in the selection of the models marked $*$. Thus, we retain the following reduced models:

$$\{X_1, X_2\}, \qquad \{X_1, X_2, X_3\}, \qquad \{X_1, X_2, X_4\}, \qquad \{X_1, X_3, X_4\}, \text{ and } \{X_2, X_3, X_4\}.$$

These above are the desired reduced models. Then, we use $\hat{\Gamma}_{ri}$ versus r plot:



● means a single point.

■ means a double point.

From this plot, we see that the reduced model $\{X_1, X_2\}$ is our desired model. Note that after the first stage, we can state with confidence probability $P^* = .90$ that all other models ($\Delta = 3$) — except the 5 reduced models given above — are inferior and have been excluded.

We also note that the largest value of r for the selected models is 4. If we take 4 as an upper bound of r to start with, then an approximate upper bound of Δ can be obtained using the asymptotic relation $\Delta \approx 2\alpha^2 - (p - r)$. For P^*=0.90, we get $\alpha = -1.282$ and therefore $\Delta \approx 2.29$.

References

[1] Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data.* 2nd edition, John Wiley & Sons, New York.

[2] Draper, N. and Smith, H. (1981). Applied Regression Analysis (2nd. ed.) Wiley, New York.

[3] Ghosh, B. K. (1973). Some monotonicity theorems for chi-square, F and t distributions with applications. *Journal of the Royal Statistical Society*, Series B, 35, 480–492.

[4] Graybill, F. A. (1976). *Theory and Application of the Linear Model.* Duxbury Press, Massachusetts.

[5] Guenther, W. C. (1979). The use of noncentral F approximations for calculation of power and sample size. *Amer. Statist.* Vol. 33, No. 4, 209–210.

[6] Gupta, S. S., Huang, D. Y. and Chang, C. L. (1984). Selection procedures for optimal subsets of regression variables. *Design of Experiments* (edited by Santner, T. J. and Tamhane, A. C.), 67–75.

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

[8] Hocking, R. R. (1983). Developments in Linear Regression Methodology: 1959–1982. (with discussions). *Technometrics* 25, 219–249.

[9] Huang, D. Y. and Panchapakesan, S. (1982). On eliminating inferior regression models. *Comm. Statist. A. - Theor. Meth.*, 11(7), 751–759.

[10] Mallows, C. L. (1973). Some comments on *Cp. Technometrics* 15, 661–675.

[11] Patnaik, P. B. (1949). The noncentral chi-squared and F distributions and their applications. *Biometrika* 36, 202–232.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
Technical Report #86-29

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
SELECTING IMPORTANT INDEPENDENT VARIABLES IN LINEAR REGRESSION MODELS

5. TYPE OF REPORT & PERIOD COVERED
Technical

6. PERFORMING ORG. REPORT NUMBER
Technical Report #86-29

7. AUTHOR(s)
Shanti S. Gupta and Deng-Yuan Huang

8. CONTRACT OR GRANT NUMBER(s)
N00014-84-C-0167

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
November 1986

13. NUMBER OF PAGES
15

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)
Selection procedures; Noncentrality parameters; Noncentral F; Total square error; Reduced model; Inferior models; Selection criteria.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A large body of literature exists on the techniques for selecting the important variables in linear regression analysis. Many of these techniques are ad hoc in nature and have not been studied from a theoretical viewpoint. In this paper we discuss some of the more commonly used techniques and propose a selection procedure based on the statistical selection and ranking approach. This procedure is easy to compute and apply. The procedure depends on the goodness of fit of the model and the total error associated with it.