SELECTING IMPORTANT INDEPENDENT VARIABLES IN LINEAR REGRESSION MODELS REVISION(U) PURDUE UNIV LAFAYETTE IN DEPT OF STATISTICS S S GUPTA ET AL. NOV 06
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
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.

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 parameter 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 \( \Delta \) 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

\[
Y = X\beta + \varepsilon
\]  
\[ (2.1) \]

where \( Y' = [Y_1, \ldots, Y_n] \) is an \( 1 \times n \) vector of a random sample, \( X = [1, X_1, \ldots, X_{p-1}] \) is an \( n \times p \) matrix of known constants, \( \beta' = [\beta_0, \beta_1, \ldots, \beta_{p-1}] \) is a \( 1 \times p \) vector of unknown parameters and \( \varepsilon \sim N(0, \sigma^2 I_n) \). 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 \( r \) of the \( p \) 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

\[
E(Y) = X_{ri} \beta_{ri}, \quad i = 1, 2, \ldots, k_r
\]  
\[ (2.2) \]

where \( X_{ri} \) and \( \beta_{ri} \) are obtained from \( X \) and \( \beta \) 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

\[
SS_{ri} = Y'[(I - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri})]Y
= Y'Q_{ri}Y,
\]

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

\[
\frac{SS_{ri}}{\sigma_0^2} \sim \chi^2(\nu_r, \lambda_{ri}),
\]

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

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

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

We have

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

Since \( \sigma_0^2 \) is fixed, it is clear from (2.5) that \( \lambda_{ri} \) should not be large for a good model.

We define any reduced model with associated noncentrality parameter \( \lambda_{ri} \) inferior if \( \lambda_{ri} > \Delta \) 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 \( \sigma_0^2 \), 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

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

and

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

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

$$\lambda_{ri} = \frac{n - p}{2} \frac{SS_{ri} - \nu_r}{SS_{p1} - \nu_r}$$

$$= \frac{n - p_1 - R^2_{ri}}{2 - 1 - R^2} - \frac{\nu_r}{2}.$$  \hfill (2.6)

Proposed Selection Procedure.

We propose the selection rule $S$ as follows:

Exclude (reject) the reduced model $M_{ri}$

$$\text{iff } \lambda_{ri} \geq d_{ri}$$

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

$$\lambda_{ri} \geq d_{ri}$$  \hfill (2.8)

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

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

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

$$\inf_{\lambda_{ri} \geq \Delta} P\{\lambda_{ri} \geq d_{ri}\} = P^*. \quad \hfill (2.11)$$
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:

\[ \inf_{\lambda_{ri} \geq \Delta} P\{V_{ri} \geq D_{ri}\} = P^* \quad (2.12) \]

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

\[ P\{V_{ri} \geq l, \lambda_{ri} = \Delta\} = P^*, \quad (2.13) \]

where \( D_{ri} = \left(\frac{\nu_i}{2} \right)^{\frac{n-p}{p-r}} \left(1 - \frac{n-p}{p-r}\right) \) 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 \( \Delta, n, p, r \) 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.,

\[ \sum_{j=1}^{n}(\nu_j - \eta_{ij})^2 + \sum_{j=1}^{n} \text{var} (\hat{Y}_{ij}) \quad \text{(2.14)} \]
where
\[ \nu_j = \nu(X_{1j}, X_{2j}, \ldots), \] expected value from true equation,
\[ \eta_{ij} = \beta_0 + \sum_{t=1}^{r-1} \beta_t X_{it}, \] expected value from the fitted model \( M_{r_i} \) being used,
\[ (\nu_{ij} - \eta_{ij}) = \text{bias at the } j\text{th data point, and} \]
\[ \hat{y}_i = (\hat{y}_{i1}, \ldots, \hat{y}_{in})' = X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}Y \] is the predicted value under least square estimate in the reduced model \( M_{r_i} \).

For convenience, let \( SSB_{ri} \) stand for \( \sum_{j=1}^{n} (\nu_j - \eta_{ij})^2 \) and define a quantity, \( \Gamma_{ri} \), the stand-

dardized 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
\[ \hat{Y}_i = X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}Y, \]
we have
\[ \text{cov} (\hat{Y}_i) = \sum_{j=1}^{n} (\nu_j - \eta_{ij})^2 = \frac{1}{\sigma_0^2} \sum_{j=1}^{n} \text{var} (\hat{y}_{ij}). \tag{2.16} \]

\[ SSB_{ri} = \sum_{j=1}^{n} (\nu_j - \eta_{ij})^2 = \sum_{j=1}^{n} \left[ E(Y_j) - E(\hat{Y}_{ij}) \right]^2 \]
\[ = \left[ E(Y - \hat{Y}_i) \right]' \left[ E(Y - \hat{Y}_i) \right] \]
\[ = \left\{ E(\left[ I - X_{ri}(X'_{ri}X_{ri})^{-1}X_{ri} \right]Y) \right\}' \left\{ E(\left[ I - X_{ri}(X'_{ri}X_{ri})^{-1}X_{ri} \right]Y) \right\} \]
\[ = (X\beta)'(I - X_{ri}(X'_{ri}X_{ri})^{-1}X_{ri})(X\beta) \]
\[ = 2\sigma_0^2 \lambda_{ri}. \tag{2.17} \]
\[
E(SS_{ri}) = E(Y - \hat{Y}_r)'(Y - \hat{Y}_r)
\]
\[
= E(Y'|I - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}|Y)
\]
\[
= E\{\text{tr} (Y'|I - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}|Y)\}
\]
\[
= \text{tr} \{E[(Y - X\beta)'(I - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri})Y](Y - X\beta)\}
+ \text{tr} \{(X\beta)'[I - X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}](X\beta)\}
\]
\[
= (n - r)\sigma_0^2 + 2\sigma_0^2\lambda_{ri}. \quad (2.18)
\]

From (2.17) and (2.18), we have

\[
E(SS_{ri}) = SS_{Br} + (n - r)\sigma_0^2. \quad (2.19)
\]

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

\[
\Gamma_{ri} = \frac{E(SS_{ri})}{\sigma_0^2} - (n - 2r)
= \nu_r + 2\lambda_{ri} - (n - 2r)
= 2\lambda_{ri} + r. \quad (2.20)
\]

We have an unbiased estimate of \(\Gamma_{ri}\) as follows:

\[
\hat{\Gamma}_{ri} - 2 \cdot \frac{n - p - 2}{n - p} \left[2\lambda_{ri} + (n - r)\right] - (2p - 3r) \quad (2.21)
\]

since

\[
2\lambda_{ri} + (n - r) = (p - r)\nu_{ri}.
\]

Hence, we can show that for \(n - p - 2\), we have

\[
E(\hat{\Gamma}_{ri}) = 2n - \frac{p - 2}{n - p} \cdot (p - r) \cdot \frac{(p - r) + \lambda_{ri}}{(n - p - 2)} \cdot \frac{n}{p - r} \cdot \frac{1}{n - p}
= 2\lambda_{ri} + r \cdot \Gamma_{ri}. \quad (2.22)
\]

From (2.5), we see that \(\lambda_{ri}\) is a measure of the error. That is, \(\lambda_{ri}\) 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

\[
\Gamma_{ri} \approx r.
\]
We require the total squared error to be small for good fit. Hence \( \hat{r}_{ri} \) should be as small as possible and close to \( r \). Hence \( \hat{r}_{ri} \) 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 \( \hat{Y}_{ir} = (\hat{Y}_{i1}, \ldots, \hat{Y}_{in})' = X_{ri}(X'_{ri}X_{ri})^{-1}X'_{ri}Y_i \). Now from (2.21), as \( n \to p + 2 \),

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

is an unbiased estimator of the standardized total squared error \( \Gamma_{ri} \). Also if \( S = R_{ri} = 0 \), then \( \Gamma_{ri} \approx r \).

**The Relation between \( R_{ri}^2 \) and \( \hat{r}_{ri} \).

From (2.7) and (2.21), we have

\[
\Gamma_{ri} = 2(n - p) \cdot \frac{1}{(2n - 3r)} \cdot \hat{r}_{ri}^2.
\]

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

**The Relation between \( F \)-statistic \( V_{ri} \) and \( \hat{r}_{ri} \).

From Theorem 2, we have

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

**The Relation between Mallows’ \( C_{ri} \) and \( \hat{r}_{ri} \).**
Mallows' $C_{ri}$ is defined as follows:

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

Hence

$$\hat{r}_{ri} = \frac{2(n - p - 2)(p - r)}{(n - p)(n - r)} |C_{ri} + (n - 2r)| - (2p - 3r). \quad (2.25)$$

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 $\hat{r}_{ri}$.

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 $\hat{r}_{ri}$ against $r$ with $\hat{r}_{ri}$ as small as possible (see Theorem 2). It follows from the fact that $SSB_{ri}/\sigma^2_{ri} = 2\lambda_{ri}$, that the large values of $\lambda_{ri}$ measure the degree of the departure from the line $\hat{r}_{ri} = r$.

**Computation of Constants $D_{ri}$**

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

$$F(p_1, p_2, \lambda) \approx \frac{(p_1 + 2\lambda)/p_1}{F(p^*, p_2)} \quad (3.1)$$
where
\[ p^* = \frac{(p_1 + 2\lambda)^2}{(p_1 + 4\lambda)}. \]

Hence, we can determine \( D_{ri} \) from the following (approximation) equation (see Theorem 1):
\[
P\{ F(p^*; n - p) \geq \left[ \frac{p - r}{(p - r) + 2\Delta} \right] D_{ri} \} = P^*,
\]
(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
\[
D_{ri} = \frac{(p - r) + 2\Delta}{p - r} F(p^*; n - p; P^*).
\]
(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 \( \Delta \). Assuming the sample size \( n \) to be sufficiently large, we study the asymptotic results for the two-stage selection procedure \( R_s \). Let \( n - p > 4 \).

\[
P^* = P\{ V_{ri} \geq D_{ri} | \lambda_{ri} = \Delta \} = P\{ \frac{V_{ri} - E(V_{ri})}{\sqrt{\text{Var}(V_{ri})}} \geq \frac{D_{ri} - E(V_{ri})}{\sqrt{\text{Var}(V_{ri})}} | \lambda_{ri} = \Delta \} \approx P\{ Z \geq \alpha \} = 1 - \Phi(\alpha),
\]
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

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

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

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

Let

\[
\hat{\Gamma}_{ri}^\alpha = \frac{2(n - p - 2)(p - r)}{n - p} \left\{ \alpha \sqrt{\frac{2(n - p)^2}{(p - r)^2(n - p - 2)}} \sqrt{\frac{(p - r + 2\Delta)^2}{(n - p - 2)(n - p - 4)}} + \frac{p - r + 4\Delta}{n - p - 4} \right\} - (2p - 3r)
- \sqrt{2D\alpha |Ax^2 + Bx + C|^{\frac{1}{2}} - x + 4\Delta + p},
\]

where

\[
A = \frac{(n - p)^2}{(n - p - 2)^2(n - p - 4)};
\]

\[
B = \frac{4\Delta(n - p)^2}{(n - p - 2)^2(n - p - 4)} + \frac{(n - p)^2}{(n - p - 4)(n - p - 2)};
\]

\[
C = \frac{4\Delta^2(n - p)^2}{(n - p - 2)^2(n - p - 4)} + \frac{4\Delta(n - p)^2}{(n - p - 4)(n - p - 2)};
\]

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, \( \hat{\Gamma}_{ri}^\alpha \approx 2\sqrt{2\alpha} \sqrt{x + 4\Delta} - x + p - 4\Delta \).

By letting \( \frac{d\hat{\Gamma}_{ri}^\alpha}{d\Delta} = 0 \), we have \( \Delta \approx 2\alpha^2 - x \), such that \( \frac{d^2\hat{\Gamma}_{ri}^\alpha}{d\Delta^2} > 0 \). Hence, \( \hat{\Gamma}_{ri}^\alpha \) is minimized when \( \Delta \approx 2\alpha^2 - x \). For which, we can find an upper bound of \( \Delta \) such that at least how many variables are excluded for this bound, since \( \Delta \) 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.

<table>
<thead>
<tr>
<th>No.</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>26</td>
<td>6</td>
<td>60</td>
<td>78.5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>29</td>
<td>15</td>
<td>52</td>
<td>74.3</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>56</td>
<td>8</td>
<td>20</td>
<td>104.3</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>31</td>
<td>8</td>
<td>47</td>
<td>87.6</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>52</td>
<td>6</td>
<td>33</td>
<td>95.9</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>55</td>
<td>9</td>
<td>22</td>
<td>109.2</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>71</td>
<td>17</td>
<td>6</td>
<td>102.7</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>31</td>
<td>22</td>
<td>44</td>
<td>72.5</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>54</td>
<td>18</td>
<td>22</td>
<td>93.1</td>
</tr>
<tr>
<td>10</td>
<td>21</td>
<td>47</td>
<td>4</td>
<td>26</td>
<td>115.9</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>40</td>
<td>23</td>
<td>34</td>
<td>83.8</td>
</tr>
<tr>
<td>12</td>
<td>11</td>
<td>66</td>
<td>9</td>
<td>12</td>
<td>113.3</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>68</td>
<td>8</td>
<td>12</td>
<td>109.4</td>
</tr>
</tbody>
</table>

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

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

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

12
Now we apply the procedure $R_\alpha$. At stage 1, we exclude all inferior reduced models. This results in the selection of the models marked *. Thus, we retain the following reduced models:

\{X_1, X_2\}, \{X_1, X_2, X_3\}, \{X_1, X_2, X_4\}, \{X_1, X_3, X_4\}, 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 \( \Delta \) 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


**Technical Report**

**Title:** SELECTING IMPORTANT INDEPENDENT VARIABLES IN LINEAR REGRESSION MODELS

**Authors:** Shanti S. Gupta and Deng-Yuan Huang

**Performing Organization Name and Address:** Purdue University, Department of Statistics, West Lafayette, IN 47907

**Controlling Office Name and Address:** Office of Naval Research, Washington, DC

**Report Date:** November 1986

**Number of Pages:** 15

**Distribution Statement:** Approved for public release, distribution unlimited.

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

**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.