ON ELIMINATING INFERIOR REGRESSION MODELS*

by

Deng-Yuan Huang
Institute of Mathematics
Academia Sinica
Taipei, Taiwan

S. Panchapakesan
Department of Mathematics
Southern Illinois University
Carbondale, Illinois

Mimeograph Series #81-25

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #81-25

July 1981

*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 ELIMINATING INFERIOR REGRESSION MODELS*

Deng-Yuan Huang
Institute of Mathematics
Academia Sinica
Taipei, Taiwan

S. Panchapakesan
Department of Mathematics
Southern Illinois University
Carbondale, Illinois

Key Words and Phrases: Linear regression models; eliminating inferior models; multiple correlation; guaranteed probability of correct decision.

AMS 1980 subject classifications: Primary 62F07; secondary 62J05

ABSTRACT

Consider a linear regression model with (p-1) predictor variables which is taken as the "true" model. The goal is to select a subset of all possible reduced models such that all inferior models (to be defined) are excluded with a guaranteed minimum probability. A procedure is proposed for which the exact evaluation of the probability of a correct decision is difficult; however, it is shown that the probability requirement can be met for sufficiently large sample size. Monte Carlo evaluation of the constant associated with the procedure and some ways to reduce the amount of computations involved in the implementation of the procedure are discussed.

1. INTRODUCTION

A problem of great interest to many practitioners of linear regression analysis is that of selecting an appropriate subset of the predictor variables which adequately describe the variance of the response variable. Some of the commonly employed techniques are all possible regressions, forward selection, backward selection and stepwise procedures. These procedures along with some variations and computational methods are given in Draper and Smith (1966). Several criteria for defining the best set of predictor variables and

*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.

various techniques for selecting the best set have been discussed in a nice expository paper by Hocking (1976). A brief review of these methods is also given by Thompson (1978). However, these techniques do not have an accompanying probability guarantee for selecting the best set; moreover, the measures of goodness of models are based on the data. Formulation of this problem in the framework of the multiple decision subset selection procedures of Gupta (1956, 1965) has been recently considered by Arvesen and McCabe (1975), Gupta and Huang (1977), and McCabe and Arvesen (1974). We adopt the same framework here. For details of the general subset selection theory, see Gupta and Panchapakesan (1979).

Formulation of the problem is given in Section 2. In the next section, a procedure is proposed and the infimum of the probability of a correct decision (PCD) is expressed in terms of the models obtained by dropping one predictor variable at a time. Section 3 discusses asymptotic results and establishes the (asymptotic) least favorable configuration for PCD. However, this still does not make the calculation of the necessary constant easy. The next section describes a Monte Carlo method of determining the constant. A few facts which can be effectively used in reducing the amount of computations needed in implementing the procedure are discussed in Section 5.

2. FORMULATION OF THE PROBLEM

Consider the model

$$Y_j = \beta_0 + \beta_1 x_{j1} + \ldots + \beta_{p-1} x_{j,p-1} + \epsilon_j, \quad j = 1,\ldots,N \tag{2.1}$$

where $x_{j,r}$, $j = 1,\ldots,N$, are fixed levels of the predictor variables $x_1,\ldots,x_{p-1}$, the β_i are unknown parameters, and the ϵ_j are independent normal random variables with mean zero and variance σ_0^2. Let $\underline{Y}' = (Y_1,\ldots,Y_N)$, $\underline{\beta}' = (\beta_0,\ldots,\beta_{p-1})$, $\underline{\epsilon}' = (\epsilon_1,\ldots,\epsilon_N)$, $\underline{1}' = (1,\ldots,1)$, and $\underline{x}_i' = (x_{1i},\ldots,x_{Ni})$, $i = 1,\ldots,p-1$. Then the model (2.1) can be written in the familiar matrix form

$$\underline{Y} = X\underline{\beta} + \underline{\epsilon} \tag{2.2}$$

where $X = (\underline{1}\ \underline{x}_1\ \ldots\ \underline{x}_{p-1})$ and the rank of X is assumed to be $p \leq N$. It is assumed that (2.2) represents the "true" model. We wish to compare

with this true model all the models that can be obtained by taking only some of the predictor variables. In order to define inferior models, we need a measure of goodness of a model. For any fixed $\alpha = 0,1,\ldots,p-1$, consider all the $\binom{p-1}{\alpha}$ subsets of the set of predictor variables $\{x_1,\ldots,x_{p-1}\}$ and the corresponding reduced models obtained from (2.2). Associated with these reduced models are the multiple correlation coefficients $R_{i,\alpha}$, $i = 1,2,\ldots,\binom{p-1}{\alpha}$. The indexing of these reduced models can be done in an arbitrary manner. Let $\theta_{i,\alpha} = E(1-R^2_{i,\alpha})$. Then the goodness of a reduced model is defined by comparing $\theta_{i,\alpha}$ for the model with the parameter $\theta_{1,p-1}$ of the true (full) model.

Definition 2.1. A reduced model whose associated parameter $\theta_{i,\alpha}$ is said to be inferior if $\theta_{1,p-1} \leq \delta^* \theta_{i,\alpha}$, where $\delta^* \in (0,1)$ is a specified constant.

It is to be noted that comparison of models based on $\theta_{i,\alpha}$ is equivalent to that based on the expected residual sums of squares in the ANOVA of these models. However, it is more practical to fix δ^* in relation to multiple correlation coefficients as they are unit-free.

The true model is, of course, the best model. While eliminating the inferior models, we do not want to overly reject good models. Formally stated, our goal is: Select a subset of all possible models with preferably a large subset size so that all the inferior models are excluded from the selected subset of models with a guaranteed minimum probability P^* $(0 < P^* < 1)$.

Definition 2.2. Model A is said to be a submodel of Model B if the set of predictor variables of A is a subset of that of B.

Since the multiple correlation coefficient for a model cannot be smaller than the coefficient for any of its submodels, we make the following remark.

Remark 2.1. If any model is inferior, then all its submodels are inferior.

The number of inferior models, t_1, is unknown. Of course, $0 \leq t_1 \leq t-1$, where $t = 2^{p-1}-1$. Let $\Omega(t_1)$ denote the set of all parametric configurations that give rise to exactly t_1 inferior models and let $\Omega = \bigcup_{t_1} \Omega(t_1)$.

We now propose a procedure based on the sample multiple correlation coefficients for the different models. Let $R_{i,\alpha}$ denote the sample coefficient corresponding to the model associated with $\theta_{i,\alpha}$ and set $\hat{\theta}_{i,\alpha} = 1 - R^2_{i,\alpha}$.

3. PROCEDURE $\mathcal{R}$

The proposed procedure $\mathcal{R}$ is: exclude from the selected subset any model for which

$$\hat{\theta}_{i,\alpha} \leq \frac{c}{\delta^*} \hat{\theta}_{1,p-1} \tag{3.1}$$

where the constant $c = c(N,p,P^*) > \delta^*$ is determined such that $P(CD|\mathcal{R})$, the probability of a correct decision using $\mathcal{R}$, satisfies the inequality

$$P(CD|\mathcal{R}) \geq P^*. \tag{3.2}$$

We first note that if any model is excluded by $\mathcal{R}$, then all its submodels are also excluded. Further, we need only to determine the ratio $c/\delta^* = d$ (say).

For a parametric configuration in $\Omega(t_1)$,

$$P(CD|\mathcal{R}) \geq Pr\{\hat{\theta}_{i,p-2} \leq d\hat{\theta}_{1,p-1}, \quad i = 1,\ldots,p-1\}. \tag{3.3}$$

The above inequality is obvious because of Remark 2.1 and the fact that the right-hand side of (3.3) is the probability of a correct decision when $t_1 = t-1$. Consequently,

$$\inf_{\Omega} P(CD|\mathcal{R}) = \inf_{\Omega} Pr\{\hat{\theta}_{i,p-2} \leq d\,\hat{\theta}_{1,p-1}, \quad i = 1,\ldots,p-1\}. \tag{3.4}$$

Let $X_{(i)}$ denote the matrix obtained from X by deleting the column vector $\underline{x}_i$, and $\underline{\beta}_{(i)}$ denote the vector obtained from $\underline{\beta}$ by leaving β_i out. Consider the (p-1) reduced models given by

$$\underline{Y} = X_{(i)}\,\underline{\beta}_{(i)} + \underline{\epsilon}_i, \quad i = 1,\ldots,p-1, \tag{3.5}$$

where $\underline{\epsilon}_i \sim N(\underline{0}, \sigma_i^2 I_n)$. It should be noted that in stating the reduced model

(3.5), we mean that the model is used for prediction purposes using only the (p-1) variables of $X_{(i)}$. However, our comparisons of models are made under the true model assumptions. The expectation of the residual mean square in the corresponding ANOVA evaluated under the true model is σ_i^2 given by the result (e) at the end of this section. The reduced model described in (3.5) reflects this fact.

Now, let SS_i denote the residual sum of squares in the ANOVA corresponding to the model with $X_{(i)}$ and let SS_0 denote the residual sum of squares in the ANOVA of the full model. Then we can summarize our discussion thus far in the following theorem.

Theorem 3.1. For the procedure $\mathcal{R}$ defined in (3.1),

$$\inf_{\Omega} P(CD|\mathcal{R}) \geq \inf_{\Omega} \Pr\{SS_i \geq d\, SS_0,\ i = 1,\ldots,p-1\}. \tag{3.6}$$

Exact evaluation of the infimum on the right-hand side of (3.6) is difficult. We take recourse to asymptotic theory and try to achieve the probability requirement in (3.2) for large N in the next section. We state below a few well-known results in regression theory which we need.

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

(b) $SS_0 = \underline{Y}'\{1 - X(X'X)^{-1}X'\}\underline{Y} = \underline{Y}'Q_0\underline{Y}$, say,

(c) $SS_i/\sigma_0^2 \sim \chi^2(\nu, (X\underline{\beta})'Q_i(X\underline{\beta})/2\sigma_0^2)$ (under the true model)

where $\nu = N - p+1$ and $\chi^2(\nu, \lambda)$ denotes the noncentral chi-square distribution with ν degrees of freedom and noncentrality parameter λ.

(d) $SS_0/\sigma_0^2 \sim \chi^2(\nu_0)$, the (central) chi-square distribution with $\nu_0 = N-p$ degrees of freedom.

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

4. ASYMPTOTIC RESULTS

Since our rule is invariant with respect to $\sigma_0^2 > 0$, we can assume that $\sigma_0^2 = 1$. Following Arvesen and McCabe (1975), we write $\underline{Y}'Q_i\underline{Y} = \underline{U}_i'\underline{U}_i$,

$i = 0,1,\ldots,p-1$, where $\underline{U}_i = B_i\underline{Y}$ with $B_iB_i' = I$ and $B_i'B_i = Q_i$. Here B_i is a $\nu \times n$ matrix for $i = 1,\ldots,p-1$ and B_0 is a $\nu_0 \times n$ matrix. The joint distribution of $\underline{U}' = (\underline{U}_0', \underline{U}_1', \ldots, \underline{U}_{p-1}')$ is multivariate normal in $\nu_0 + (p-1)\nu$ dimensions with mean vector $\underline{\eta}' = (\underline{\eta}_0', \underline{\eta}_1', \ldots, \underline{\eta}_{p-1}')$ with $\underline{\eta}_i = B_i X\underline{\beta}$, $i = 0,1,\ldots,p-1$, and covariance matrix $\Sigma = (\Sigma_{ij})$ where $\Sigma_{ij} = B_iB_j'$. Note that Σ is possibly singular.

Now, letting

$$Z_i = (SS_i - \nu - \underline{\eta}_i'\,\underline{\eta}_i)/\sqrt{2\nu + 4\underline{\eta}_i'\,\underline{\eta}_i}\,, \quad i = 1,\ldots,p-1,$$
$$Z_0 = (SS_0 - \nu_0)/\sqrt{2\nu_0}\,, \tag{4.1}$$

we have

$$\Pr\{SS_i \geq d\,SS_0,\ i = 1,\ldots,p-1\}$$

$$= \Pr\left\{\frac{SS_i}{\sqrt{2\nu + 4\underline{\eta}_i'\,\underline{\eta}_i}} \geq d\,\frac{SS_0}{\sqrt{2\nu_0}}\,\sqrt{\frac{2\nu_0}{2\nu + 4\underline{\eta}_i'\,\underline{\eta}_i}}\,,\quad i = 1,\ldots,p-1\right\}$$

$$\geq \Pr\left\{\frac{SS_i}{\sqrt{2\nu + 4\underline{\eta}_i'\,\underline{\eta}_i}} \geq d\,\frac{SS_0}{\sqrt{2\nu_0}}\,\sqrt{\frac{\nu_0}{\nu}}\,,\quad i = 1,\ldots,p-1\right\}$$

$$= \Pr\left\{Z_i + \frac{\nu + \underline{\eta}_i'\,\underline{\eta}_i}{\sqrt{2\nu + 4\underline{\eta}_i'\,\underline{\eta}_i}} \geq d\,\sqrt{\frac{\nu_0}{\nu}}\,Z_0 + \frac{d\nu_0}{\sqrt{2\nu}}\,,\quad i = 1,\ldots,p-1\right\}$$

$$\geq \Pr\left\{Z_i + \sqrt{\frac{\nu}{2}} \geq d\sqrt{\frac{\nu_0}{\nu}}\;Z_0 + \frac{d\nu_0}{\sqrt{2\nu}}\,,\quad i = 1,\ldots,p-1\right\}.$$

The last inequality follows from the fact that $\underline{\eta}_i'\,\underline{\eta}_i \geq 0$ and $\Phi(t) = (\nu + t)/\sqrt{\nu + 2t}$ is strictly increasing in $t \geq 0$. Thus we have

$$\Pr\,SS_i \geq d\,SS_0,\ i = 1,\ldots,p-1\}$$
$$\geq \Pr\left\{Z_i \geq d\sqrt{\frac{\nu_0}{\nu}}\,Z_0 + \frac{d\nu_0}{\sqrt{2\nu}} - \sqrt{\frac{\nu}{2}}\,,\quad i = 1,\ldots,p-1\right\}. \tag{4.2}$$

It can be easily seen that we have equality in (4.2) when $\underline{\eta}_i' \underline{\eta}_i = 0$. Also, the joint distribution of $Z_0, Z_1,\ldots,Z_{p-1}$ does not depend upon $\underline{\eta}_0, \underline{\eta}_1,\ldots,\underline{\eta}_{p-1}$ for large N. This shows that the worst configuration (asymptotically) for $\Pr\{SS_i \geq d\, SS_0,\ i = 1,\ldots,p-1\}$ is when $\underline{\beta} = \underline{0}$. Thus we can achieve the probability requirement (3.2) for large N if d is determined by

$$\Pr\{Z_i \geq d\sqrt{\frac{\nu_0}{\nu}}\, Z_0 + \frac{d\nu_0}{\sqrt{2\nu}} - \sqrt{\frac{\nu}{2}},\quad i = 1,\ldots,p-1\} = P^* \tag{4.3}$$

where $\underline{Z}' = (Z_0,Z_1,\ldots,Z_{p-1})$ has a multivariate normal distribution with mean zero and covariance matrix $\Gamma = (\rho_{ij})$ with

$$\rho_{ij} = \frac{1}{\nu}\operatorname{tr}(\Sigma_{ij}\Sigma_{ji}) = \frac{1}{\nu}\operatorname{tr}(Q_jQ_i),\quad i \neq j,\ i,j = 1,\ldots,p-1,$$

and

$$\rho_{0j} = \frac{1}{\sqrt{\nu\nu_0}}\operatorname{tr}(\Sigma_{0j}\Sigma_{j0}) = \frac{1}{\sqrt{\nu\nu_0}}\operatorname{tr}(Q_jQ_0),\quad j = 1,\ldots,p-1.$$

5. EVALUATION OF THE CONSTANT

The evaluation of the constant d can be done using an Edgeworth approximation of order $1/\sqrt{N}$ as discussed by Arvesen and McCabe (1975). But, as they have remarked, this may be a formidable problem for $p \geq 4$ or 5. So we resort to Monte Carlo technique. The steps involved are described below.

(1) Generate random observations $Y_1,\ldots,Y_N$ from a standard normal distribution.

(2) Calculate SS_i, $i = 0,1,\ldots,p-1$.

(3) Form the ratio $A = \min_{1\leq i\leq p-1} SS_i/SS_0$.

(4) Repeat steps 1 to 3 m times retaining the values $A_1, A_2,\ldots,A_m$.

(5) Denote the ordered A_i by $A_{[1]} \leq \ldots \leq A_{[m]}$.

Then the estimate of d is $\hat{d} = A_{[r+1]}$, where r is an integer such that $r/m \leq (1-P^*) < (r+1)/m$.

Based on the experiences of McCabe and Arvesen (1974) with their problem, it appears that m = 1000 may give adequate estimates.

6. IMPLEMENTATION OF THE PROCEDURE

The procedure $\mathfrak{R}$ in (3.1) can be restated as follows.

$\mathfrak{R}$: Exclude any model if the corresponding residual sum of squares $SS_j \geq d\ SS_0$. The implementation of the procedure is straightforward since it involves only the evaluations of the residual sums of squares of all the models. However, it may not be necessary to compute the residual sums of squares for all the reduced models. It should be remembered that the rejection of any model implies the rejection of all its submodels. For example, if we have p-1 = 4 predictor variables. We first consider all the one variable models. Suppose that the models $\{x_1\}$ and $\{x_4\}$ are selected and the models $\{x_2\}$ and $\{x_3\}$ are rejected. This automatically means that the models $\{x_1, x_2\}$, $\{x_1, x_3\}$, $\{x_1, x_4\}$, $\{x_2, x_4\}$, $\{x_3, x_4\}$, $\{x_1, x_2, x_3\}$, $\{x_1, x_2, x_4\}$, $\{x_1, x_3, x_4\}$, $\{x_2, x_3, x_4\}$ and $\{x_1, x_2, x_3, x_4\}$ are selected. It leaves only $\{x_2, x_3\}$ to be considered next.

7. ACKNOWLEDGEMENT

This research was supported by the Office of Naval Research contract N00014-75-C-0455 at Purdue University. An earlier version of this paper was written when the second author was visiting at the Institute of Mathematics, Academia Sinica, Taipei with the financial support of National Science Council, Republic of China.

BIBLIOGRAPHY

Arvesen, J. N. and McCabe, G. P. (1975). Subset selection problems for variances with applications to regression analysis. J. Amer. Statist. Assoc. 70, 166-170.

Draper, N. R. and Smith, H. (1966). Applied Regression Analysis. New York: John Wiley.

Gupta, S. S. (1956). On a decision rule for a problem in ranking means. Mimeo. Ser. No. 150, Inst. of Statistics, Univ. of North Carolina, Chapel Hill, NC.

Gupta, S. S. (1965). On some multiple decision (selection and ranking) rules. Technometrics 7, 225-245.

Gupta, S. S. and Huang, D. Y. (1977). On selecting an optimal subset of regression variables. Mimeo. Ser. No. 501, Dept. of Statistics, Purdue Univ., West Lafayette, IN.

Gupta, S. S. and Panchapakesan, S. (1979). Multiple Decision Procedures. New York: John Wiley.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. Biometrics 32, 1-49.

McCabe, G. P. and Arvesen, J. N. (1974). A subset selection procedure for regression variables. J. Statist. Comput. Simul. 3, 137-146.

Thompson, M. L. (1978). Selection of variables in multiple regression: Part 1. A review and evaluation. Int. Statist. Rev. 46, 1-19.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Mimeograph Series #81-25
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): On Eliminating Inferior Regression Models

5. TYPE OF REPORT & PERIOD COVERED: Technical

6. PERFORMING ORG. REPORT NUMBER: Mimeo. Series #81-25

7. AUTHOR(s): Deng-Yuan Huang and S. Panchapakesan

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 1981

13. NUMBER OF PAGES: 9

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)

Linear regression models; eliminating inferior models; multiple correlation; guaranteed probability of correct decision.

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

Consider a linear regression model with (p-1) predictor variables which is taken as the "true" model. The goal is to select a subset of all possible reduced models such that all inferior models (to be defined) are excluded with a guaranteed minimum probability. A procedure is proposed for which the exact evaluation of the probability of a correct decision is difficult; however, it is shown that the probability requirement can be met for sufficiently large sample size. Monte Carlo evaluation of the constant associated with the procedure and some ways to reduce the amount of computations involved in the implementation of the procedure

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

are discussed.