Mean Square Error of Prediction as a Criterion for Selecting Variables

DAVID M. ALLEN

University of Kentucky, Lexington, Kentucky

The mean square error of prediction is proposed as a criterion for selecting variables. This criterion utilizes the values of the predictor variables associated with the future observation and the magnitude of the estimated variance. Mean square error is believed to be more meaningful than the commonly used criterion, the residual sum of squares.

1. INTRODUCTION

The problem of selecting variables in multiple regression has received a great deal of attention. Among the more common procedures are the forward selection method, the backward elimination method, and Efroymson's (1960) stepwise regression. These procedures are discussed in Chapter 6 of Draper and Smith (1966). Garside (1965) and Schatzoff, Feinberg, and Tsao (1968) propose efficient methods of enumerating all possible regression equations. Hocking and Leslie (1967), Beale, Kendall, and Mann (1967), LaMotte and Hocking (1970), and Beale (1970a) give procedures for finding the subset regression equation of a specified size having the minimum residual sum of squares. Mantel (1970) supports the backward elimination method. Beale (1970b) compares some of the various methods available.

The objective of all these papers is to minimize the residual sum of squares subject to various conditions such as reasonable computing time or a fixed number of variables in the regression equation. There are at least two objections to the use of the residual sum of squares as a criterion for selecting variables:

(1) If the residual sum of squares were the sole criterion, then one would always use all of the variables. Thus, to delete variables, there must be an additional criterion such as the number of variables used. The degree to which the two criteria are weighted is arbitrary.

(2) The residual sum of squares is not directly related to the commonly used criteria for good prediction and estimation.

This paper presents a criterion and subsequent computing procedure for selecting variables for the purpose of prediction. The selection criterion is based on the mean square error criterion. The use of mean square error takes into account the expected value of the observation being predicted and eliminates the arbitrariness associated with the residual sum of squares. Our procedure is a modification of the method presented by Anderson, Allen, and Cady (1970).

2. The Mean Square Error of Prediction

The classical multiple linear regression model is

$$\underset{(n\times 1)}{\mathbf{Y}} = \underset{(n\times r)}{X}\ \underset{(r\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}} \tag{1}$$

where $\mathbf{Y}$ is a vector of responses; X is a known, full rank matrix of nonstochastic variables; $\boldsymbol{\beta}$ is the unknown weight vector corresponding to X; and $\boldsymbol{\varepsilon}$ is a normally distributed vector of random variables having expected value $\mathbf{0}$ and dispersion matrix $1\sigma^2$. In this model, the Xs may represent different functional forms of the same basic variables, e.g., $X_1 = w$, $X_2 = v$, $X_3 = w^2$, $X_4 = v^2$, $X_5 = wv$. Given data generated by this model, we want to predict the value of a future random variable y having mean $\underset{(1\times r)}{x}\ \boldsymbol{\beta}$ and variance σ^2. The row vector x contains the values of the X variables associated with the future observation.

We would view as ideal, a predictor $\hat{y}$ such that

$$E(\hat{y} - y)^2 \tag{2}$$

is minimum for all values of the unknown parameters. The ideal predictor does not exist, but criterion (2) provides a basis for comparing various predictors. We refer to (2) as the mean square error of prediction (MSEP). The MSEP of a predictor $\hat{y}$ can be expressed as

$$E(\hat{y} - y)^2 = \sigma^2 + \text{Var}\ (\hat{y}) + (E(\hat{y}) - x\boldsymbol{\beta})^2. \tag{3}$$

The last term of (3) is the squared bias of prediction. The last two terms of (3) are the mean square error (MSE) of $\hat{y}$ when viewed as an estimator of $x\boldsymbol{\beta}$. Since the MSEP and MSE differ only by a constant, we are actually considering dual problems: the prediction of y and the estimation of a linear combination of $\boldsymbol{\beta}$.

3. The MSEP Predictor

The least squares predictor is

$$\hat{y}_r = x\mathbf{b} \tag{4}$$

where $\mathbf{b} = (X'X)^{-1}X\mathbf{Y}$. The least squares predictor is unbiased and has variance $x(X'X)^{-1}x'\sigma^2$. Thus its mean square error of prediction is

$$\text{MSEP}_r = \sigma^2 + x(X'X)^{-1}x'\sigma^2. \tag{5}$$

MSEP_r is the basis for the comparison of other predictors. As alternatives to the least squares predictor we consider the class of least squares predictors based on linear functions of the columns of X and x. From this class we select the member having the smallest estimated MSEP.

The linear functions of the columns of X and x are conveniently expressed as

$$XP_a \quad \text{and} \quad xP_a$$

where P_a is an $(r \times a)$ matrix of full rank. As an example, suppose $r = 5$, and we want to examine X_1, $X_2 + X_3$, $X_4 + X_5$ as predictor variables, then

$$P_a = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

If "a" variables are used as predictors, then $k = r - a$ variables are excluded. A convenient representation for the excluded variables is

$$XP_k \quad \text{and} \quad xP_k$$

where P_k is an $(r \times k)$ matrix of full rank such that $P_k'P_a = 0$.

The predictor based XP_a and xP_a is

$$\hat{y}_a = xP_a(P_a'X'XP_a)^{-1}P_a'X'Y. \tag{6}$$

Using the matrix identity

$$P_a(P_a'X'XP_a)^{-1}P_a' = (X'X)^{-1} - (X'X)^{-1}P_k(P_k'(X'X)^{-1}P_k)^{-1}P_k'(X'X)^{-1}$$

we see

$$\hat{y}_a = (x - z)\mathbf{b}$$

where $z = x(X'X)^{-1}P_k(P_k'(X'X)^{-1}P_k)^{-1}P_k'$. The variance of $\hat{y}_a$ is

$$\sigma^2 + x(X'X)^{-1}x'\sigma^2 - z(X'X)^{-1}z'\sigma^2.$$

Since the first two terms comprise the variance of $\hat{y}_r$ (equation (5)) and the third term is non-positive, we see the variance of $\hat{y}_a$ is less than or equal to the variance of $\hat{y}_r$. The bias of $\hat{y}_a$ is $z\boldsymbol{\beta}$, and thus its mean square error of prediction is

$$\text{MSEP}_a = \sigma^2 + x(X'X)^{-1}x'\sigma^2 - z(X'X)^{-1}z'\sigma^2 + (z\boldsymbol{\beta})^2. \tag{7}$$

Since the first two terms of MSEP_a are constant we need to consider only

$$(z\boldsymbol{\beta})^2 - z(X'X)^{-1}z'\sigma^2. \tag{8}$$

(Note that (8) is simply $\text{MSEP}_a - \text{MSEP}_r$. If this quantity is negative, then $\hat{y}_a$ is better than $\hat{y}_r$.) Since (8) depends upon unknown parameters $\boldsymbol{\beta}$ and σ^2 we base our technique on its estimator

$$(z\mathbf{b})^2 - 2z(X'X)^{-1}z'S^2 \tag{9}$$

where $S^2 = (n - r + 2)^{-1}\mathbf{Y}'(1 - X(X'X)^{-1}X')\mathbf{Y}$. This estimator is the minimum mean square error estimator of the class $(z\mathbf{b})^2 - lS^2$. Suppose there are a finite number of "reasonable" values of P_a. Our predictor is $\hat{y}_a$ evaluated at the P_a for which (9) is minimum. When feasible, the appropriate P_a is found by evaluating (9) at each potential value of P_a.

4. Computation

Researchers often like to restrict themselves to subsets of the predictor variables. Subsets are obtained by having P_a matrices such that every column contains one "1", and all other elements equal "0". We will discuss the computation required for considering subsets of the variables. Other linear combinations may be considered in this manner by actual transformation of the variables. Once $(X'X)^{-1}$, $\mathbf{b}$, and S^2 have been obtained it is very easy to compute $\hat{y}_a$ and criterion (9) if z is known. We focus on the computation of z. Begin with $X'X$ and sweep[1] on the "a" pivotal elements of the variables to be included. Denote the resulting matrix by C. Let $i_1, i_2, \cdots, i_a$ be the indices of the variables included in the subset and $j_1, j_2, \cdots, j_k$ be the indices of the excluded variables. The elements of z are given by

$$z_{i_l} = 0, \quad l = 1, 2, \cdots, a$$

and

$$z_{j_m} = x_{j_m} - \sum_{l=1}^{a} x_{i_l} c_{i_l i_m} \qquad m = 1, 2, \cdots, k.$$

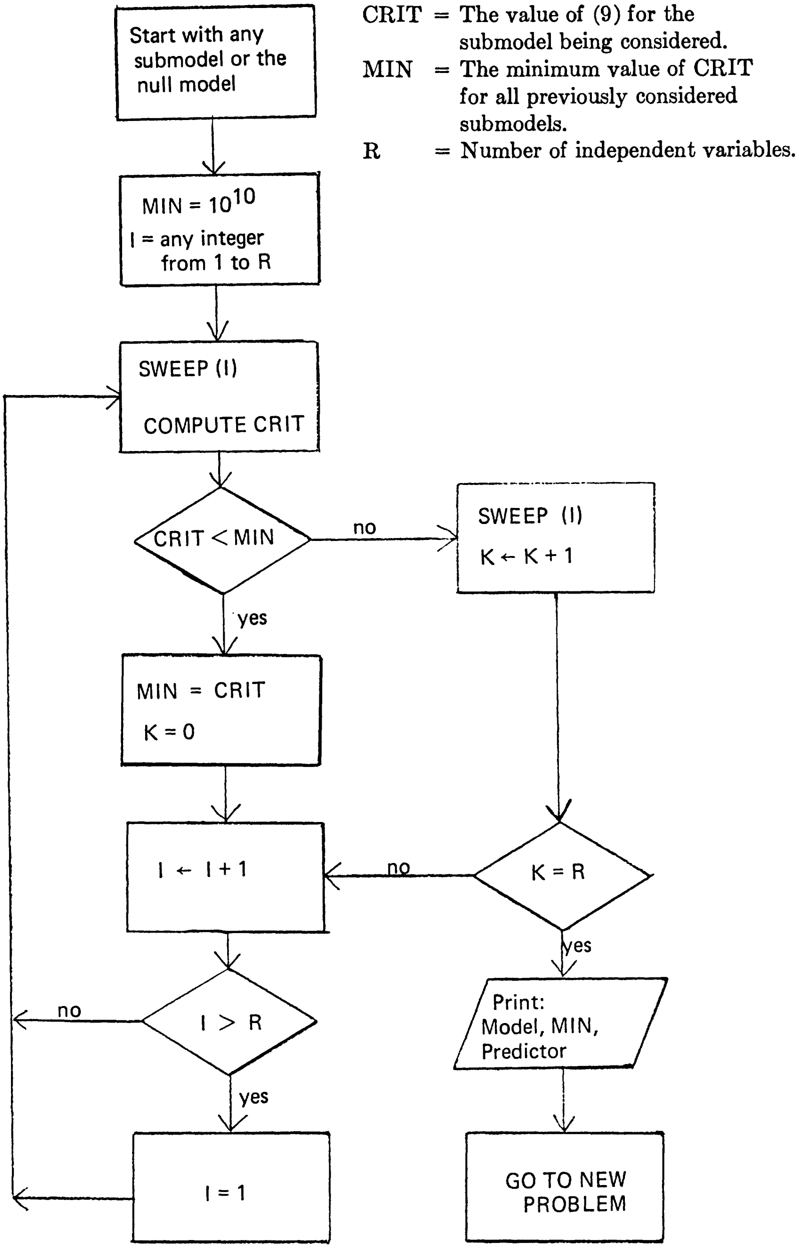
Schatzoff *et al.* present a sequence of sweeps whereby all $2^r - 1$ possible subsets may be considered efficiently.

If r is large it may not be practical to evaluate (9) for each of the possible subsets. If so, we propose the following sequential procedure. Start with any submodel and any variable. Reverse the status of the variable, i.e., if the variable is in the subset, delete it, and if the variable is not in the subset, include it. This is accomplished by sweeping the pivotal element associated with the variable. If the value of (9) is decreased from its previous minimum, consider the next variable. If not, reverse the status of the variable and then consider the next variable. Repeatedly consider all of the variables in turn until a submodel is obtained such that (9) cannot be decreased by either adding or deleting a variable. The logic is illustrated by the flow chart in figure 1. Unlike other sequential algorithms based on changes in the residual sum of squares where the stopping rule is arbitrary, the proposed mean square error procedure has a well-defined stopping point. However, like other sequential procedures, no guarantee exists that an absolute minimum is obtained.

5. Multiple Future Observations

The procedure outlined has been described for one future predicted value. If there are multiple future observations to be predicted, then repeat the process for each future observation. Recall that the predictor variables may be different for each predicted observation. Numerical work has shown that the subset of variables chosen to predict one future observation may be considerably different from the subset chosen to predict another future observation. We offer the following explanation. Suppose $x = (100 \cdots 0)$, then we are predicting an observation having expected value β_1 (the first element of $\boldsymbol{\beta}$) or, equivalently, we are estimating β_1. We have an a priori feeling that X_1 should be in the subset when we are estimating β_1. We have no such feeling about the other variables in regard to estimating β_1. Applying this thought to the other elements

[1] For example, see Schatzoff, Feinberg, and Tsao (1968).



CRIT = The value of (9) for the submodel being considered.
MIN = The minimum value of CRIT for all previously considered submodels.
R = Number of independent variables.

FIGURE 1—The Sequential MSEP Predictor

of $\boldsymbol{\beta}$, it seems quite reasonable that different future values should be based on different subsets of predictor variables.

6. NUMERICAL EXAMPLE

The techniques presented here are illustrated using data from Draper and Smith (1966, p. 351–352). We let $X_1 = 1$, X_2 through X_{10} be as Draper and

Smith define them, and Y be the variable they denote by X_1. We consider four different values of x; these are given in Table 1. The first two of these are the X variables of observations included in the data. The third and fourth are values of x to estimate β_3 and β_8 which are parameters whose least squares estimators have respective variances 52.94 σ^2 and .000848 σ^2. The sequential procedure is used. The process is started by considering X_2 as an addition to a submodel containing only X_1. The results are presented in Table II.

Criterion (9) provides a comparison between submodels for the same future observation. However, one cannot compare the performance of one submodel for two different future observations using (9). We can compare the estimated relative efficiencies of $\hat{y}_r$ to $\hat{y}_a$ for different future observations. Specifically,

$$-\frac{(z\mathbf{b})^2 - 2z(X'X)^{-1}z'S^2}{(1 + x(X'X)^{-1}x')S^2} \tag{10}$$

is an estimator of

$$1 - \frac{\text{MSEP}_a}{\text{MSEP}_r}. \tag{11}$$

The quantity (11) is the proportionate reduction in MSEP due to having used the submodel rather than the full model. While (11) has a maximum value of 1, the maximum value of (10) is $2z(X'X)^{-1}z'/(1 + x(X'X)^{-1}x')$. Thus, if the variance of $\hat{y}_r$ is large and the variance of $\hat{y}_a$ is small, i.e., $z(X'X)^{-1}z'$ approaches $x(X'X)^{-1}x'$ in magnitude, then (10) may be greater than 1. The column in Table II labeled "Estimated percent reduction in MSEP" contains values of (10) converted to percent. Had our major emphasis been estimation, we would have used the estimated MSE_r in the denominator of (10) rather than the estimated MSEP_r.

TABLE I
Values of x' considered

Index	Future observation or linear combination			
	1	2	3	4
1	1.00	1.00	0.	0.
2	6.57	5.87	0.	0.
3	.87	.70	1.	0.
4	4.10	7.50	0.	0.
5	31.00	31.00	0.	0.
6	23.00	22.00	0.	0.
7	0.00	28.00	0.	0.
8	76.70	28.60	0.	1.
9	16.80	56.30	0.	0.
10	5.00	5.00	0.	0.

TABLE II

Results of the sequential selection procedure

Future observation	Predicted value	Estimated percent reduction in MSEP	Submodel
1	8.27	30.5	X_1,X_4,X_7
2	11.52	33.0	X_1,X_6,X_8
3	0.00	170.4	X_1
4	−.080	.0	X_1,X_2,X_8

The first two predicted values were obtained using three variables, but not the same three. Both predictors had a substantial reduction in estimated MSEP when compared to the full model. The full model is very poor for estimating β_3 (recall the large variance). The MSEP predictor appears to be a tremendous improvement. The full model provides a good estimate of β_8 . The MSEP predictor produced the same predicted value with three variables.

ACKNOWLEDGEMENTS

The author is grateful to his colleagues, Drs. Foster B. Cady and R. L. Anderson, for their motivation and helpful suggestions.

REFERENCES

[1] ANDERSON, R. L., ALLEN, D. M., and CADY, F. B. (1970). Selection of Predictor Variables in Linear Multiple Regression. University of Kentucky, Department of Statistics, Technical Report Number 5. (also to appear in a volume honoring George W. Snedecor).

[2] BEALE, E. M. L. (1970a). Selecting an Optimum Subset. *Integer and Nonlinear Programming.* Ed. J. Abadie, North Holland Publishing Co. Amsterdam.

[3] Beale, E. M. L. (1970b). Note on Procedures for Variable Selection in Multiple Regression. *Technometrics 12,* 909–914.

[4] BEALE, E. M. L., KENDALL, M. G., and MANN, D. W. (1967). The Discarding of Variables in Multivariate Analysis. *Biometrika 54,* 357–366.

[5] DRAPER, N. R., and SMITH, H. (1966). *Applied Regression Analysis.* John Wiley Sons, Inc., New York.

[6] EFROYMSON, M. A. (1960). *Multiple Regression Analysis. Mathematical Methods for Digital Computers.* Ed. A. Ralston and H. S. Wilf. John Wiley Sons, Inc., New York.

[7] GARSIDE, M. J. (1965). The Best Sub-set in Multiple Regression Analysis. *Applied Statistics 14,* 196–200.

[8] HOCKING, R. R., and LESLIE, R. B. (1967). Selection of the Best Subset in Regression Analysis. *Technometrics 9,* 531–540.

[9] LAMOTTE, L. R., and HOCKING, R. R. (1970). Computational Efficiency in the Selection of Regression Variables. *Technometrics 12,* 83–93.

[10] MANTEL, N. (1970). Why Stepdown Procedures in Variable Selection. *Technometrics 12,* 621–25.

[11] SCHATZOFF, M., FEINBERG, S. and TSAO, R. (1968). Efficient Calculations of All Possible Regressions. *Technometrics 10,* 769–779.