Appl. Statist. (1974),
23, *No.* 1, p. 11.

Predictive Mean Square Error and Stochastic Regressor Variables

By Subhash C. Narula

State University of New York at Buffalo

Summary

When prediction is the main objective, predictive mean square error (p.m.s.e.) seems to be a more reasonable criterion. Here we consider two approaches to improve the p.m.s.e. of the predicted response when predictor variables are stochastic and, in particular, follow a multivariate normal distribution.

The first technique, the subset approach, uses only a subset of the available predictor variables to predict the response. A decision rule to select the subset is given. In the second method, the lambda approach, the regression coefficients are scaled down by a suitable constant. An estimator of the constant is suggested. Both techniques are illustrated by an example.

Keywords: PREDICTIVE MEAN SQUARE ERROR; SUBSET SELECTION; LAMBDA APPROACH; STOCHASTIC VARIABLES; SHRINKAGE

1. Introduction

In many regression problems a large number of regressor variables are available. Walls and Weeks (1969) gave an argument against "overfitting" prediction equations by showing that the variance of the predicted response can never be reduced by the addition of variables to the equation. However, when a variable is left out of the model a biased prediction results. If prediction is the major objective, the mean square error (squared bias + variance) of the predicted response from the observed response would seem to be a reasonable criterion, since it takes into account bias and variability simultaneously (Allen, 1971; Narula, 1971).

Several methods of improving the performance of a regression equation are known. The most commonly used is the rejection of variables, and another is to scale down the regression coefficients by a constant factor. A third is the method known as "ridge regression".

In this paper the first two of these methods are compared. It is easier to make the comparison when a well-defined type of regression problem, in this case prediction, and a well-defined model are considered. Hence we consider the prediction problem when the predictor variables (usually referred to as regressor variables) are assumed stochastic and in particular to follow a multivariate normal distribution. The problem arises naturally in the real world. Ehrenberg (1963) suggested that regression of stochastic regressor variables is useless and Kerridge (1967) suggested that it may be useful if the limitations are understood. We think, if properly understood, it has its uses.

2. Statement of the Problem

Assume that the response variable and the predictor variables follow a joint $(k+1)$-variate normal distribution with unknown mean vector

$$\boldsymbol{\mu}^* = [\mu_0, \mu_1, \ldots, \mu_k]' = [\mu_0, \boldsymbol{\mu}']'$$

and unknown covariance matrix $\boldsymbol{\Sigma}^* = \begin{bmatrix} \sigma_{00} & \boldsymbol{\sigma}' \\ \boldsymbol{\sigma} & \boldsymbol{\Sigma} \end{bmatrix}$. Let $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ be n independent (k-component vector) observations on the predictor variables and $y_1, y_2, \ldots, y_n$ be the corresponding observations on the response variable. Also define the sample mean $\bar{\mathbf{z}} = \sum \mathbf{z}_i/n$, $\bar{y} = \sum y_i/n$ and $\mathbf{x}_i = \mathbf{z}_i - \bar{\mathbf{z}}$, i.e. the values of the predictor variables corrected for the sample means. Let $\mathbf{S}^* = \begin{bmatrix} s_{00} & \mathbf{s}' \\ \mathbf{s} & \mathbf{S} \end{bmatrix}$ be the sample covariance matrix, where $s_{00} = \sum (y_i - \bar{y})^2/(n-1)$, $\mathbf{s} = \sum (y_i - \bar{y})\mathbf{x}_i/(n-1)$ and $\mathbf{S} = \sum \mathbf{x}_i \mathbf{x}_i'/(n-1)$. Since for any observed vector $\mathbf{z}_i$ of the predictor variables, $E(y_i | \mathbf{z}_i) = \mu_0 + \boldsymbol{\sigma}' \boldsymbol{\Sigma}^{-1}(\mathbf{z}_i - \boldsymbol{\mu})$, we can write

$$y_i = \alpha + (\mathbf{z}_i - \boldsymbol{\mu})' \boldsymbol{\beta} + \varepsilon_i,$$

where $\alpha = \mu_0 - \boldsymbol{\sigma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$, $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma}$ and ε_i is random error such that $E(\varepsilon_i) = 0$, and $\operatorname{var}(\varepsilon_i) = \operatorname{var}(y_i | \mathbf{z}_i) = \sigma_{00} - \boldsymbol{\sigma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma} = \sigma_k^2$ (say) and each observation is independently normally distributed. The least squares predicted equation is given by

$$\hat{y}_i = \bar{y} + \mathbf{x}_i' \hat{\boldsymbol{\beta}}, \tag{1}$$

where $\hat{\boldsymbol{\beta}} = \mathbf{S}^{-1}\mathbf{s}$.

We are interested in the problem of predicting the response, y_0, corresponding to an observed vector, $\mathbf{z}_0$, of predictor variables rather than obtaining "better" estimates of the unknown parameters of the model. In real life we generally have the value of $\mathbf{z}_0$ before we can observe y_0 (otherwise why the prediction problem in the first place!). Using (1) the predicted response at $\mathbf{z}_0$ is given by $\hat{y}_0 = \bar{y} + \mathbf{x}_0' \hat{\boldsymbol{\beta}}$ (where $\mathbf{x}_0 = \mathbf{z}_0 - \bar{\mathbf{z}}$) and the conditional (conditioned on $\mathbf{z}_0$) predictive mean square error (p.m.s.e.) by

$$E\{(y_0 - \hat{y}_0)^2 | \mathbf{z}_0\} = \sigma_k^2(1 + 1/n) + \sigma_k^2\{(\mathbf{z}_0 - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{z}_0 - \boldsymbol{\mu}) + k/n\}/(n-k-2).$$

The unconditional p.m.s.e. can be obtained by taking the expectation of the conditional p.m.s.e. over $\mathbf{z}_0$ and is given by

$$E(y_0 - \hat{y}_0)^2 = \sigma_k^2(1 + 1/n)(n-2)/(n-k-2),$$

a result reported by Kerridge (1967).

3. The Subset Approach

The problem of selecting a subset of regressor variables has been widely considered in the literature. Most often, with a few exceptions, Kerridge (1967) and Lindley (1968), the authors have addressed themselves to the problem when the regressor variables are assumed fixed. In this section, we give a decision rule to select a subset of stochastic predictor variables which will have smaller p.m.s.e. than does (1).

Partition the k-component vector of the predictor variables into two parts $\mathbf{z}_i' = [\mathbf{z}_{i1}', \mathbf{z}_{i2}']$ where $\mathbf{z}_{i1}$ (a p-component vector) represents a set of p predictor variables included in the prediction equation and $\mathbf{z}_{i2}$ (a $(k$–$p)$-component vector), those not included. (No loss of generality is involved in this partition since the numbering of the variables is arbitrary.) Accordingly we also partition $\mathbf{x}_i' = [\mathbf{x}_{i1}', \mathbf{x}_{i2}']$, $\bar{\mathbf{z}}' = [\bar{\mathbf{z}}_1', \bar{\mathbf{z}}_2']$,

$\boldsymbol{\mu}' = [\boldsymbol{\mu}_1', \boldsymbol{\mu}_2']$, $\boldsymbol{\sigma}' = [\boldsymbol{\sigma}_1', \boldsymbol{\sigma}_2']$, $\mathbf{s}' = [\mathbf{s}_1', \mathbf{s}_2']$, $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$ and $\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}$ so that the subset prediction equation is given by

$$\tilde{y}_i = \bar{y} + \mathbf{x}_{i1}' \tilde{\boldsymbol{\beta}}_1, \tag{2}$$

where $\tilde{\boldsymbol{\beta}}_1 = \mathbf{S}_{11}^{-1} \mathbf{s}_1$.

Using (2), the predicted response at $\mathbf{z}_0' = [\mathbf{z}_{01}', \mathbf{z}_{02}']$ is given by $\tilde{y}_0 = \bar{y} + \mathbf{x}_{01}' \tilde{\boldsymbol{\beta}}_1$, where $\mathbf{x}_{01} = \mathbf{z}_{01} - \bar{\mathbf{z}}_1$.

3.1. *A Decision Rule*

Since our objective is to improve the p.m.s.e., we would prefer to predict the response at $\mathbf{z}_0$ using the prediction equation that minimizes $E\{(y_0 - \tilde{y}_0)^2 | \mathbf{z}_0\}$ for all possible subsets. (Equation (1) is a special case of the subset equation as it contains all the variables.) But $E\{(y_0 - \tilde{y}_0)^2 | \mathbf{z}_0\}$ depends upon unknown parameters of the model (see Appendix A). An obvious choice is to replace the unknown parameters by their least squares estimators.

Hence our decision rule becomes: Select the subset that minimizes

$$\hat{\sigma}_p^2/n + \hat{\sigma}_p^2(\mathbf{x}_{01}' \mathbf{S}_{11}^{-1} \mathbf{x}_{01} + p/n)/(n-p-2) + (\mathbf{x}_{02}' \hat{\boldsymbol{\beta}}_2 - \mathbf{x}_{02}' \mathbf{S}_{11}^{-1} \mathbf{S}_{12} \hat{\boldsymbol{\beta}}_2)^2, \tag{3}$$

where $\hat{\sigma}_p^2 = s_{00} - \mathbf{s}_1' \mathbf{S}_{11}^{-1} \mathbf{s}_1$.

When we are not sure of the points at which we would like to predict and hence would like to select a subset which has smaller p.m.s.e. on the average, we will select the subset that minimizes $\hat{\sigma}_p^2(1 + 1/n)(n-2)/(n-p-2)$ subject to

$$\hat{\sigma}_p^2/\hat{\sigma}_k^2 \leqslant (n-p-2)/(n-k-2), \quad n > k+2.$$

The above decision rules involve 2^{k-1} matrix inversion—one for each subset. The algorithm given by Garside (1965) and Schatzoff *et al.* (1968) may be used.

4. The Lambda Approach

To improve the mean square error of various estimators, the "shrinkage" or "scale down" technique has been applied by Stein (1960), James and Stein (1961), Sclove (1968) and Thompson (1968). It has never been utilized in this context before. We show that this technique, called lambda approach here, can be usefully applied to the problem at hand.

Hence, instead of using the usual prediction equation, we shall use

$$y_i^* = \bar{y} + \lambda_{\mathbf{z}_i} \mathbf{x}_i' \hat{\boldsymbol{\beta}},$$

where $\hat{\boldsymbol{\beta}} = \mathbf{S}^{-1}\mathbf{s}$ and $\lambda_{\mathbf{z}_i}$ $(0 \leqslant \lambda_{\mathbf{z}_i} \leqslant 1)$ is an unknown constant (the subscripts $\mathbf{z}_i$ of λ is to emphasize that the value of λ depends upon $\mathbf{z}_i$). When $\lambda_{\mathbf{z}_i} = 1$, we have the usual prediction equation and when $\lambda_{\mathbf{z}_i} = 0$, we predict the response by the response mean, $\bar{y}$, alone.

For any observed value of the predictor variables, $\mathbf{z}_0$, the predicted response is given by $y_0^* = \bar{y} + \lambda_{\mathbf{z}_0} \mathbf{x}_0' \hat{\boldsymbol{\beta}}$. Although y_0^* and y_0 are independent, y_0^* is obviously not an unbiased predictor of y_0. The value of $\lambda_{\mathbf{z}_0}$ which minimizes the p.m.s.e. depends upon the unknown parameters of the model (see Appendix B).

However, in a practical situation, an estimate of $\lambda_{\mathbf{z}_0}$, $\hat{\lambda}_{\mathbf{z}_0}$ (say), can be obtained by using the least squares estimates of the unknown parameters as

$$\hat{\lambda}_{\mathbf{z}_0} = \hat{A}/(\hat{A} + \hat{B}), \tag{4}$$

where $\hat{A} = \mathbf{x}_0' \hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}' \mathbf{x}_0 + \hat{\boldsymbol{\beta}}' \mathbf{S} \hat{\boldsymbol{\beta}}/n$ and $\hat{B} = \hat{\sigma}_k^2(\mathbf{x}_0' \mathbf{S}^{-1} \mathbf{x}_0 + k/n)/(n-k-2)$.

It can be observed that $0 \leqslant \hat{\lambda}_{\mathbf{z}_0} \leqslant 1$ and the prediction equation becomes

$$\hat{y}_0^* = \bar{y} + \hat{\lambda}_{\mathbf{z}_0} \mathbf{x}_0' \hat{\boldsymbol{\beta}}.$$

5. Example

On the basis of a student's performance in English, Mathematics, Social Sciences and Natural History tests (here referred to as predictor variables) of the American College Testing service, we want to predict the first year college grade point average, G.P.A., of the student (response variable). The data are assumed to follow a multivariate normal distribution. Out of a total of 83 observations (on the male students only), 25 are selected at random and are used to estimate the parameters of the equation. Based on these estimates the remaining 58 observations are predicted.

It is felt that both the techniques will result in greater improvement, if the sample size used to estimate the unknown parameters is small. To study this, small samples are created artificially in the following manner. From the 25 observations 20 are selected at random. Then from these 20, 15 are selected at random and from these 15, 10 are selected at random. Now each of these sets of 25, 20, 15 and 10 observations are used for estimating the unknown parameters of the model and the same 58 observations are predicted each time.

In this problem, we know the value of the predictor variables (for the 58 observations to be predicted) for each observation. Hence, for the subset approach, the p.m.s.e. is estimated by (3) for each observation and the subset equation with the smallest p.m.s.e. is selected to predict the G.P.A. For the lambda approach similarly, the lambda value is calculated for each observation. In both approaches, residual for each observation is calculated as the observed G.P.A. minus the predicted G.P.A. and the sum of squares of the residuals for the 58 observations is computed.

The results are summarized in Table 1.

Table 1

Residual sum of squares (percentage improvement)

	No. of observations			
Technique	10	15	20	25
Full equation ($\lambda = 1$)	125·48	58·94	55·51	55·14
Lambda approach	91·17 (27·34)	50·86 (13·70)	51·51 (7·21)	51·89 (5·88)
Subset equation	54·03 (56·94)	41·62 (29·38)	45·31 (18·38)	53·91 (2·23)

Note that in the subset approach, a different subset of variables is used for each observation and in the lambda approach, lambda is calculated for each observation. Table 1 clearly indicates the usefulness of both the techniques. In both approaches, the improvement is greater when smaller samples are used for estimating the parameters (at present we have no proof for this phenomenon). The lambda approach is simpler to use but for this problem the subset approach results in larger improvement.

Acknowledgement

I am thankful to Dr John S. Ramberg of the University of Iowa for his encouragement during the research and to Dr Melven R. Novick of the American College

Testing Service, Iowa City, Iowa, for providing the data. I extend my thanks to the referee for comments improving the form of the paper.

This research was supported in part by National Science Foundation Grant No. 30966X.

References

ALLEN, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, **13**, 469–476.

EHRENBERG, A. S. C. (1963). Bivariate regression analysis is useless. *Appl. Statist.*, **12**, 161–179.

GARSIDE, M. J. (1965). The best subset in multiple regression analysis. *Appl. Statist.*, **14**, 196–201.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Sym. Math. Statist. Prob.*, Vol. 1, pp. 361–379.

KERRIDGE, D. (1967). Errors of prediction in multiple regression with stochastic regressor variables. *Technometrics*, **9**, 309–311.

LINDLEY, D. V. (1968). The choice of variables in multiple regression. *J. R. Statist. Soc.* B, **30**, 31–66.

NARULA, S. C. (1971). Least squares regression with mean square error criterion. Ph.D. Dissertation, University of Iowa.

SCHATZOFF, M., TSAO, R. and FIENBERG, S. (1968). Efficient calculations of all possible regressions. *Technometrics*, **10**, 769–780.

SCLOVE, S. L. (1968). Improved estimators for coefficients in linear regression. *J. Amer. Statist. Ass.*, **63**, 596–606.

STEIN, C. (1960). Multiple regression. In *Contributions to Probability and Statistics*, "Essays in Honor of Harold Hotelling", pp. 424–443. Stanford: Stanford University Press.

THOMPSON, J. R. (1968). Some shrinkage techniques for estimating the mean. *J. Amer. Statist. Ass.*, **63**, 113–122.

WALLS, R. E. and WEEKS, D. L. (1969). A note on the variance of a predicted response in regression. *Amer. Statist.*, **23**, 24–26.

Appendix A

A.1. *Some Mathematical Results on the Subset Approach*

If we predict the response using (2), the conditional p.m.s.e. is given by

$$\begin{aligned} E\{(y_0-\tilde{y}_0)^2\,|\,\mathbf{z}_0\} &= E\{(\mathbf{x}_0'\,\boldsymbol{\beta}+\varepsilon_0-\bar{\varepsilon}-\mathbf{x}_{01}'\,\tilde{\boldsymbol{\beta}}_1)^2\,|\,z_0\} \\ &= \sigma_k^2+\sigma_p^2/n \\ &\qquad +\sigma_p^2\{(\mathbf{z}_{01}-\boldsymbol{\mu}_1)'\,\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{z}_{01}-\boldsymbol{\mu}_1)+p/n\}/(n-p-2) \\ &\qquad +\{(\mathbf{z}_{02}-\boldsymbol{\mu}_2)'\,\boldsymbol{\beta}_2-(\mathbf{z}_{01}-\boldsymbol{\mu}_1)'\,\boldsymbol{\Sigma}_{11}^{-1}\,\boldsymbol{\Sigma}_{12}^{-1}\,\boldsymbol{\beta}_2\}^2, \end{aligned} \tag{A.1}$$

where

$$\sigma_p^2 = \sigma_{00}-\boldsymbol{\sigma}_1'\,\boldsymbol{\Sigma}_{11}^{-1}\,\boldsymbol{\sigma}_1.$$

By taking the expectation of the conditional p.m.s.e. over $\mathbf{z}_0$, we obtain the unconditional p.m.s.e. as

$$E(y_0-\tilde{y}_0)^2 = \sigma_p^2(1+1/n)\,(n-2)/(n-p-2). \tag{A.2}$$

Our objective is to minimize the p.m.s.e. If $\mathbf{z}_0$, the vector of predictor variable, is known at which we want to predict, we would prefer to predict the response using the subset equation for which (A.1) is minimum (an equation with all variables is a special case of subset equation).

But if we do not know where we would like to predict, then we would select the equation which has the smallest unconditional p.m.s.e. Hence we would select the subset for which (A.2) is minimum.

A.2. *A Two Variable Problem*

For better understanding of the results in Appendix A.1 we consider the case of two predictor variables. The notation of Section 3 simplifies to

$$\boldsymbol{\sigma} = [\sigma_{01}, \sigma_{02}]', \quad \boldsymbol{\mu} = [\mu_1, \mu_2]' \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}.$$

Also $\mathbf{z}_{i1} = \mathbf{z}_{i1}$, $\mathbf{x}_{i1} = \mathbf{x}_{i1}$, $\boldsymbol{\sigma}_i = \sigma_{0i}$ and $\boldsymbol{\mu}_i = \mu_i$ for $i = 1, 2$, and $\boldsymbol{\Sigma}_{ij} = \sigma_{ij}$ $(i, j = 1, 2)$. We also define $\rho_{ij} = \sigma_{ij}/\sqrt{(\sigma_{ii}\sigma_{jj})}$ $(i = 0, 1, 2; j = 1, 2)$.

Then, given the observation $\mathbf{z}_0$, we would prefer to use the subset equation rather than the full equation whenever the conditional p.m.s.e. of the subset equation is less than or equal to the conditional p.m.s.e. of the full equation, i.e. whenever $E\{(y_0-\tilde{y}_0)^2 | \mathbf{z}_0\} \leqslant E\{(y_0-\hat{y}_0)^2 | \mathbf{z}_0\}$. Hence we would use the equation with x_1 alone whenever $\rho_{02}^2/(1-\rho_{01}^2) \leqslant 1/(n-3)$, for $n > 4$ and $\rho_{12} = 0$. Similarly, we use the equation with x_2 alone when $\rho_{01}^2/(1-\rho_{02}^2) \leqslant 1/(n-3)$, for $n > 4$ and $\rho_{12} = 0$. If both the inequalities are satisfied, we use the equation with smaller p.m.s.e.

When $\rho_{12} = 0$, these results may be graphically represented as in Fig. 1.

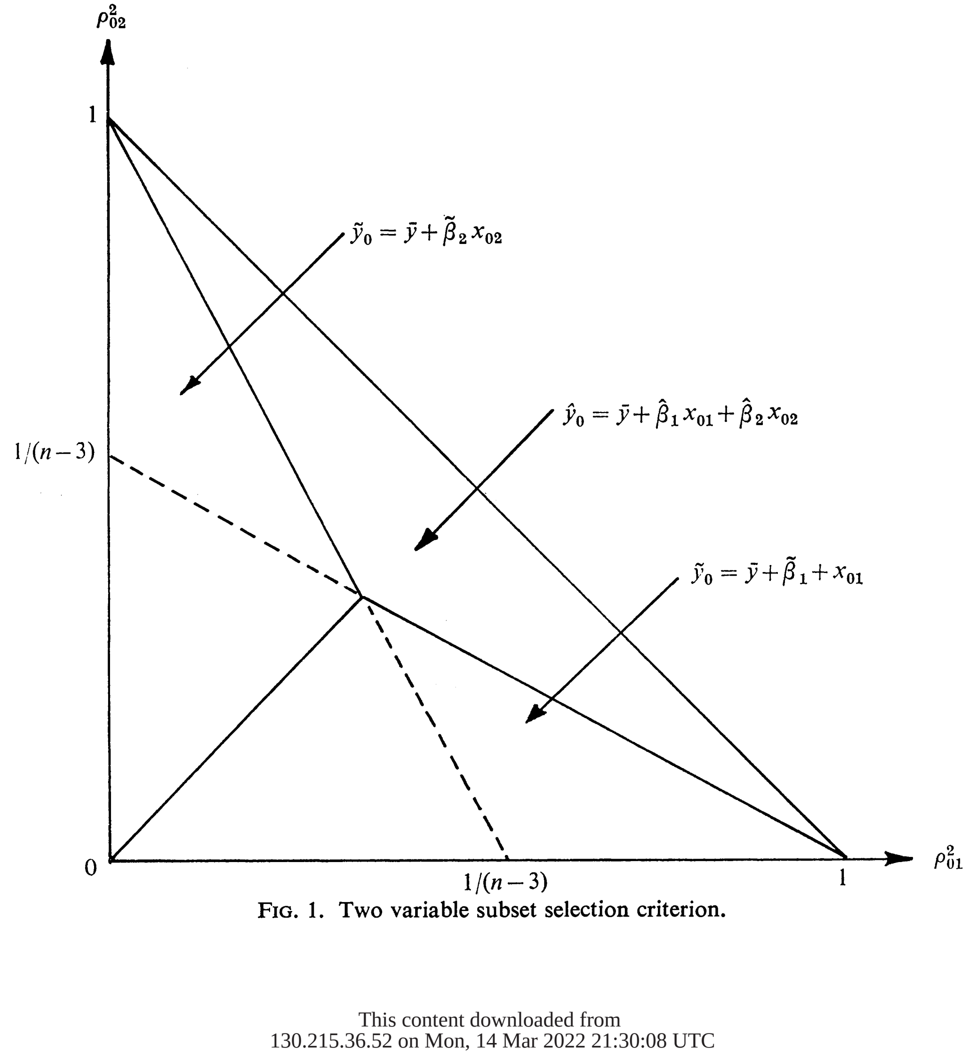
ρ_{02}^2
1
$\tilde{y}_0 = \bar{y} + \tilde{\beta}_2 x_{02}$
$\hat{y}_0 = \bar{y} + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02}$
$1/(n-3)$
$\tilde{y}_0 = \bar{y} + \tilde{\beta}_1 + x_{01}$
0
$1/(n-3)$
1
ρ_{01}^2

FIG. 1. Two variable subset selection criterion.

Appendix B

Some Results on the Lambda Approach

For an observed value of $\mathbf{z}_0$, the p.m.s.e. of the predicted response, $y_0^*(=\bar{y}+\lambda_{\mathbf{z}_0} x_0' \hat{\boldsymbol{\beta}})$ is given by

$$E\{(y_0-y_0^*)^2 \mid \mathbf{z}_0\} = (A+B)\lambda_{\mathbf{z}_0}^2 - 2A\lambda_{\mathbf{z}_0} + A + \sigma_k^2(1+1/n),$$

where

$$A = \boldsymbol{\beta}'(\mathbf{z}_0-\boldsymbol{\mu})(\mathbf{z}_0-\boldsymbol{\mu})'\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}/n$$

and

$$B = \sigma_k^2\{(\mathbf{z}_0-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{z}_0-\boldsymbol{\mu}) + k/n\}/(n-k-2).$$

Since $E\{(y_0-y_0^*)^2 \mid \mathbf{z}_0\}$ is a quadratic in $\lambda_{\mathbf{z}_0}$ and the coefficient of $\lambda_{\mathbf{z}_0}^2$ is always non-negative, the value of $\lambda_{\mathbf{z}_0}$ which minimizes the conditional p.m.s.e. is given by

$$\lambda_{\mathbf{z}_0} = A/(A+B).$$

In case we do not know exactly where we want to predict but want to find a value of λ, which will minimize the p.m.s.e. on the average, we have by taking expectation over $\mathbf{z}_0$

$$E(y_0-y_0^*)^2 = (1+1/n)\,[\{\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} + k\sigma_k^2/(n-k-2)\}\lambda^2 - 2\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}\lambda + \sigma_{00}].$$

Now $E(y_0-y_0^*)^2$ is also a quadratic in λ and the coefficient of λ^2 is non-negative and hence

$$\lambda = \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}/\{\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} + k\sigma_k^2/(n-k-2)\}$$

will minimize the p.m.s.e. A reasonable estimate of λ can be

$$\hat{\lambda} = \hat{\boldsymbol{\beta}}'\mathbf{S}\hat{\boldsymbol{\beta}}/\{\hat{\boldsymbol{\beta}}'\mathbf{S}\hat{\boldsymbol{\beta}} + k\hat{\sigma}_k^2/(n-k-2)\}.$$

Since the calculations for $\hat{\lambda}_{\mathbf{z}_0}$ (an estimator of $\lambda_{\mathbf{z}_0}$) are simple and straightforward, it is not recommended to use $\hat{\lambda}$. These results are given for completeness only.