EVALUATING OF ORDERING RIDGE REGRESSION

Goto, Masashi
Shionogi Kaiseki Center, Shionogi Research Laboratory

Matsubara, Yoshihiro
Shionogi Kaiseki Center, Shionogi Research Laboratory

https://doi.org/10.5109/13127

出版情報：統計数理研究. 18 (3/4), pp.1-35, 1979-03. Research Association of Statistical Sciences
バージョン：
権利関係：

EVALUATION OF ORDINARY RIDGE REGRESSION

By

Masashi GOTÔ*

and

Yoshihiro MATSUBARA*

(Received October 12, 1977)

1. Introduction

We consider the standard model for multiple linear regression,

$$Y=1\mu+X\beta+\varepsilon\,, \tag{1.1}$$

where 1 is an $n\times 1$ vector whose components are all one, X is an $n\times p_0$ matrix of rank p_0 consisting of n observations on p_0 regressors, Y is an $n\times 1$ vector of the corresponding responses and ε is an $n\times 1$ vector whose components are assumed to be normally distributed with mean zero and variance σ^2. Parameters μ, β, and σ^2 are all unknown. β is the $p_0\times 1$ regression coefficient vector. The model (1.1) may have different structures according to combination of assumptions on these variables and parameters, (Y, X, β, σ^2), and therefore it can be interpreted or built at least from five view points (Press (1972)). In this article we shall confine ourselves only to the standard conditional model among these structures, that is, to a case in which p_0 regressors and the corresponding response are jointly distributed (usually (p_0+1) variate normal distribution is assumed) and the response Y is observed under condition that regressors matrix X is given.

Let b be an arbitrary estimate of β. Then the maximum likelihood estimate of μ is obtained by $(\bar{Y}.-\bar{X}'.\, b)$, where $\bar{Y}.=1'Y/n$, $\bar{X}.=1'X/n$.

Thus without loss of generality we can drop the location parameter μ in the model (1.1). Since we can remove mean values from response and regressors, we can rewrite the model (1.1) in such a way that

$$Y=X\beta+\varepsilon\,, \tag{1.2}$$

where $\bar{Y}.=0$, $\bar{X}.=0$. Thus we have

$$E(Y\,|\,X)=X\beta\,,\qquad \mathrm{Var}\,(Y\,|\,X)=\left(I_n-\frac{1}{n}11'\right)\sigma^2\,, \tag{1.3}$$

where I_n denotes the $n\times n$ identity matrix. Hereafter we shall take consideration

* Shionogi Kaiseki Center, Shionogi Research Laboratory, Osaka Japan.

only on the model (1.2).

The least square estimate of β is

$$\hat{\beta}=(X'X)^{-1}X'Y. \tag{1.4}$$

The properties of the estimate $\hat{\beta}$ are well known. It is briefly summarized as follows: it is an unbiased estimate and has minumum variance among unbiased linear estimates and also the residual sum of squares is minimized with it. Since the error vector ε is normal $(0, \sigma^2 I_n)$, $\hat{\beta}$ is the maximum likelihood estimate of β and has as its distribution the p_0-variate normal $(\beta, \sigma^2(X'X)^{-1})$.

A difficulty with the least square estimate, however, is a direct consequence of the expected Euclidean distance between $\hat{\beta}$ and β. In particular, since the total mean square error of $\hat{\beta}$ is the squared distance between $\hat{\beta}$ and β, the following hold:

$$\mathrm{TMSE}(\hat{\beta})=E(\hat{\beta}-\beta)'(\hat{\beta}-\beta)=\sigma^2\sum_{i=1}^{p_0}\frac{1}{\lambda_i} \tag{1.5}$$

and

$$E(\hat{\beta}'\hat{\beta})=\beta'\beta+\sigma^2\sum_{i=1}^{p_0}\frac{1}{\lambda_i}, \tag{1.6}$$

where $\lambda_1\geqq\lambda_2\geqq\cdots\geqq\lambda_{p_0}>0$ are eigenvalues of $X'X$ arranged according to their magnitudes. Thus when the columm vectors of X are getting away from the linear independency, λ_{p_0} becomes smaller, and the norm of $\hat{\beta}$ could be very long in its expectation. It implies that one or more of its coefficients is too large in absolute value.

Ordinary ridge regression, as considered in this article, is an estimation procedure based on

$$\hat{\beta}^*=(X'X+kI_{p_0})^{-1}X'Y, \qquad k\geqq 0. \tag{1.7}$$

Equation (1.7) defines a class of estimates indexed here by a scalar parameter k. Note that $\hat{\beta}^*$ at $k=0$ is the least square estimate and is denoted simply by $\hat{\beta}$, and $\hat{\beta}^*$ is a biased estimate of β if $k>0$. The ordinary ridge estimate minimizes the residual sum of squares under a constraint on the length of the estimate. Thus, $\hat{\beta}^*$ for $k>0$ is shorter than $\hat{\beta}$. In fact, $\hat{\beta}^{*\prime}\hat{\beta}^*$ is a decreasing function of k. The residual sum of squares for $\hat{\beta}^*$ is given by

$$\begin{aligned}\mathrm{RSS}(\hat{\beta}^*)&=(Y-X\hat{\beta}^*)'(Y-X\hat{\beta}^*)\\&=\mathrm{RSS}(\hat{\beta})+(\hat{\beta}^*-\hat{\beta})'X'X(\hat{\beta}^*-\hat{\beta}), \qquad k\geqq 0.\end{aligned} \tag{1.8}$$

Hoerl and Kennard (1970a) have motivated and investigated the properties of the ridge estimate $\hat{\beta}^*$.

In particular, the total mean square error of $\hat{\beta}^*$ is obtained as

$$\begin{aligned}\mathrm{TMSE}(\hat{\beta}^*)&=E(\hat{\beta}^*-\beta)'(\hat{\beta}^*-\beta)=E(\overline{\mathrm{TMSE}})\\&=\sigma^2\sum_{i=1}^{p_0}\frac{\lambda_i}{(\lambda_i+k)^2}+\sum_{i=1}^{p_0}\gamma_i^2\left(\frac{k}{\lambda_i+k}\right)^2,\end{aligned} \tag{1.9}$$

where γ_i is the i-th component of vector, $\gamma=G'\beta$, with a $p_0\times p_0$ matrix of p_0 eigenvectors $G=[g_1, g_2, \cdots, g_{p_0}]$ associated with the eigenvalues, $\lambda_1\geqq\lambda_2\geqq\cdots\geqq\lambda_{p_0}$ of $X'X$.

They have established a type of admissibility condition, namely, the fact there always exists a k_0 such that TMSE $(\hat{\beta}^*)<$TMSE $(\hat{\beta})$. There is, however, no known mathematical method of explicitly determining the value k_0 in a given problem.

In the last six years, research papers too numerous to list here have appeared which either illustrate ridge regression on various data sets, or vigorously criticize ridge techiniques, or propose generalizations and modifications of ridge ideas (Gotô 1976)). From many discussions in those papers, it has been pointed out that the biased (shrunken) estimates such as the ridge estimate may play the important role of adjustment or improvement of the ordinary least square estimate in an ill-conditioned regression problem.

However, since the ridge estimate depends on some unknown parameters, it is necessary in practice to determine the most suitable estimates of the parameters.

In fact it may be no exaggeration to state that the choice of the ridge parameter k decides mostly the performance of the ordinary ridge estimate. As for the optimality criteria for the choice of the ridge parameter, various proposals have been made and, as pointed out by McDonald (1975), there are more than one dozen of formulae and procedures. Here we shall sort out and arrange these procedures, and finally evaluate the performance of the ordinary ridge regression by making use of systematic Monte-Carlo simulation.

We should note that the optimality meant by the "optimal" ridge estimates in various papers has been used in the following two senses: one is that the ridge estimate dominate the ordinary least square estimate in every mean squared error senses, and the other is that it achieves minimum mean squared error in some specific sense. The interpretation of $\hat{\beta}^*$ in the former sense reduces to discuss its admissibility relative to $\hat{\beta}$ on the total mean squared error. One problem in this article is to discuss the possibility of setting up the finite admissible range of $\hat{\beta}^*$ on the ridge parameter k. The problem in the latter sense reduces to choosing the ridge parameter k in such a way that it minimize TMSE $(\hat{\beta}^*)$, or some function of TMSE $(\hat{\beta}^*)$.

In section 2, we shall try to set up the admissible range of $\hat{\beta}^*$ ralative to $\hat{\beta}$ on k in the TMSE sense. We may then evaluate how much $\hat{\beta}$ can be improved by the ridge adjustment on considering the condition under which the finite admissible range of $\hat{\beta}^*$ can be constructed on k. In section 3, we shall sort out and reformulate procedures for choosing the optimal ridge parameter k according to the purposes of the usage of regression. In general, as for goals of regression analysis on observational studies, there are two potentially intended usage: (i) a fit to the data useful for future prediction in the absense of major changes in the system, and (ii) an explanation or interpretation which will link with other studies and be used for prediction under quite different circumstances.

For (i), major interests rest on the predictions of future responses but the inference of β or interpretation of it has only secondary meanings. Thus, for the intended usage in the sense of (i), we adopt as a direct measure of performance the total mean squared error of prediction,

$$\text{TMSEP}(\hat{\beta}^*)=E(\hat{Y}-E(Y))'(\hat{Y}-E(Y))=E(\overline{\text{TMSEP}}), \qquad (1.10a)$$

based on the use of $\hat{\beta}^*$, where $E(Y)=X_0\beta$, and $\hat{Y}=X_0\hat{\beta}^*$ which is the predicted value at a $n_0\times p_0$ matrix X_0 consisting of new n_0 values on p_0 regressors. Thus, we can also write

$$\text{TMSEP}(\hat{\beta}^*)=E(\beta-\hat{\beta}^*)'X_0'X_0(\beta-\hat{\beta}^*)\,. \tag{1.10b}$$

This is often referred to the weighted mean squared error of $\hat{\beta}^*$.

For (ii), the emphasis is put on the estimation of β in order to control responses through manipulation of regressors. It is usually of central importance to find which regressors have important effects. In this object, we evaluate the TMSE$(\hat{\beta}^*)$ given by (1.9) as the measure of performance of $\hat{\beta}^*$.

In section 4, to give some effective suggestions, insights and careful investigations to follow-up studies in a practical or particular regression problems the potential performance of $\hat{\beta}^*$ is evaluated numerically by means of Monte-Carlo simulations.

Particularly, we keep it in mind to construct three phases in the simulation systematically, clarifying the object or view-points in each phase to investigate the performance of $\hat{\beta}^*$. In section 5, for the practice of the ordinary ridge regression, some comments and remarks are given with conclusion. The question of scaling and centering of data (Y, X) are also discussed for calculation of the ordinary ridge regression.

2. Setting up the admissible range

The first term of the right hand side in (1.9), which is the total variance of $\hat{\beta}^*$, is a continuous, monotonically decreasing function of k, and the second term, which is the squared bias of $\hat{\beta}^*$, is a continuous, monotonically increasing function of k. Hoerl & Kennard (1970a) have suggested that TMSE$(\hat{\beta}^*)$ monotonically decreases from $k=0$ to $k=k_{\text{opt}}$, and then monotonically increases from $k=k_{\text{opt}}$ to $k=\infty$. At $k=k_{\text{opt}}$, TMSE$(\hat{\beta}^*)$ attains its minimum.

The ordinary ridge estimate is said to be admissible to the ordinary least square estimate, if TMSE$(\hat{\beta})>$TMSE$(\hat{\beta}^*)$.

In order to find admissible range of $\hat{\beta}^*$ on k, we consider the total excess mean squared error,

$$\begin{aligned}\text{TEXCESS}(\hat{\beta}^*)&=\text{TMSE}(\hat{\beta})-\text{TMSE}(\hat{\beta}^*)\\ &=\sigma^2\sum_{i=1}^{p_0}\frac{1}{\lambda_i}-\sigma^2\sum_{i=1}^{p_0}\frac{\lambda_i}{(\lambda_i+k)^2}-\sum_{i=1}^{p_0}\gamma_i^2\Big(\frac{k}{\lambda_i+k}\Big)^2 \qquad (2.1)\\ &=\sigma^2F(k)\,.\end{aligned}$$

The admissible range of $\hat{\beta}^*$ is definded as a range of k on which the function $F(k)$ takes positive values. Thus, it is necessary to find solutions of equation

$$F(k)=0\,. \tag{2.2}$$

Obviously, $k=0$ is a trivial solution of (2.2) to which the corresponding ridge estimate equals to $\hat{\beta}$. Thus, we must find the non-zero solution of (2.2).

The first derivative of $F(k)$ with respect to k is

$$F'(k)=\frac{dF(k)}{dk}=\sum_{i=1}^{p_0}\frac{2\lambda_i}{(\lambda_i+k)^3}-\sum_{i=1}^{p_0}\left(\frac{\gamma_i^2}{\sigma^2}\right)\frac{2\lambda_i k}{(\lambda_i+k)^3}$$

$$=\sum_{i=1}^{p_0}\frac{2\lambda_i}{(\lambda_i+k)^3}\left(1-\frac{\gamma_i^2}{\sigma^2}k\right). \tag{2.3}$$

Since $F'(k)$ approaches $2\sum_{i=1}^{p_0}\lambda_i^{-2}>0$ as $k\to 0$, the slope of the curve $F=F(k)$ at $k=0$ is positive, so that $F(k)$ increases monotonically from $k=0$ to $k=k_{\text{opt}}$ which gives the maximum of $F(k)$.

Similarly, from above mentioned properties of TMES ($\hat{\beta}^*$), the curve $F=F(k)$ decreases monotonically from $k=k_{\text{opt}}$ to $k=\infty$.

$F'(k)$ approaches 0 as $k\to\infty$ and $F(k)$ approaches flatly

$$F_\infty=\sum_{i=1}^{p_0}\frac{1}{\lambda_i}-\frac{\beta'\beta}{\sigma^2} \tag{2.4}$$

as a limit as $k\to\infty$.

From above results, we can obtain only one non-zero solution of (2.2), if and only if F_∞ is non-positive. Denoting the non-zero solution by k_0, we can construct a finite admissible range of $\hat{\beta}^*$ on k as $(0, k_0)$. The value of k_0 can be determined by use of Newton-Raphson algorithm, that is, by solving the equation

$$k_0^{(s+1)}=k_0^{(s)}-[F'(k_0^{(s)})]^{-1}F(k_0^{(s)}), \qquad s=0, 1, 2, \cdots, \tag{2.5}$$

iteratively, where the initial value $k_0^{(0)}$ may be given arbitrary.

Note that the equation (2.2) depends on the unknown parameters γ_i ($i=1, 2, \cdots, p_0$) and σ^2.

On the other hand, if F_∞ is positive then non-zero solution could not be obtained because the function $F(k)$ takes always positive values for all $k>0$. That is, in this case, $\hat{\beta}^*$ is admissible for all $k>0$. Thus, we can set up the admissible range of $\hat{\beta}^*$ as $(0, \infty)$.

For ill-conditioned data, λ_{p_0}, λ_{p_0-1}, $\cdots$ are usually very small and hence the first term of the right hand side in (2.4) is very large. Hence, for those data we may often have the positive value of F_∞ which may make the admissible range of $\hat{\beta}^*$ to be widely $(0, \infty)$. An appealling point of the ordinary ridge estimate could be said to lie in the fact that the admissible range of $\hat{\beta}^*$ may be wide for ill-conditioned problems.

Since the parameters β and σ^2 are unknown in practice, these usually should be estimated by $\hat{\beta}$ and s^2, respectively, where

$$s^2=\text{RSS}(\hat{\beta})/(n-p_0-1)$$

is the usual unbiased estimate of σ^2. Then $\hat{\gamma}=G'\hat{\beta}$ and s^2 are substituted for γ and σ^2 in (2.2), and hence the sample value of k_0 can be obtained. However, the performance of $\hat{\beta}$ is not good for the ill-conditioned data, so that we consider the ridge estimate $\hat{\beta}^*$ as a countermeasure. Therefore instead of using $\hat{\beta}$ itself we need an

alternative device for obtaining the sample value of k_0.

Now, taking notice of the fact that for given k

$$\max_i \left\{\gamma' \operatorname{Diag}\left[\left(\frac{k}{\lambda_i+k}\right)^2\right]\gamma\right\}=\left(\frac{k}{\lambda_{p_0}+k}\right)^2\gamma'\gamma, \tag{2.6}$$

we can construct more narrow admissible range than $(0, k_0)$, if $F_\infty \leqq 0$.

In fact, we consider the equation

$$F_s(k)=\sum_{i=1}^{p_0}\frac{1}{\lambda_i}-\sum_{i=1}^{p_0}\frac{\lambda_i}{(\lambda_i+k)^2}-\left(\frac{k}{\lambda_{p_0}+k}\right)^2\frac{\beta'\beta}{\sigma^2} \tag{2.7}$$

instead of solving the equation (2.2) directly. Clearly, for almost positive values of k, $F(k)\geqq F_s(k)$ and $|F'(k)|\geqq|F_s'(k)|$, where $F_s'(k)$ is the first derivative of $F_s(k)$ respect to k. Therefore, denoting the non-zero solution of $F_s(k)=0$ by k_s, we have always $k_0>k_s$. If $F_\infty\geqq 0$, we obtain the solution $k_s=\infty$ in the same way as in (2.2).

Noting the form of (2.7), we need not know the vector β itself. In other words, only the squared length $\beta'\beta$ and σ^2 are needed to determine k_s. In practical problem, it is more convenient to estimate k_s than k_0, because the former need only two estimates of $\beta'\beta$ and σ^2 and yields the conservative (small) estimate of k_0, while for the latter we have to know p_0+1 estimates of β itself and σ^2.

The usual estimate of $\beta'\beta$ is $\hat\beta'\hat\beta$. It is well known that

$$E(\hat\beta'\hat\beta)=\beta'\beta+\sigma^2\sum_{i=1}^{p_0}\frac{1}{\lambda_i}. \tag{2.8}$$

For ill-conditioned data, the second term of (2.8) is very large. Therefore $\hat\beta'\hat\beta$ may be larger than $\beta'\beta$. Also, $(n-p_0-1)s^2/\sigma^2$ follows the central chi-square distribution with $(n-p_0-1)$ degrees of freedom. Since the median of chisquare distribution is smaller than mean, the probability that s^2 under-estimates σ^2 is higher than 0.5. Thus, it is expected that $\hat\beta'\hat\beta/s^2$ overestimates $\beta'\beta/\sigma^2$ and that it may yield larger estimate of the squared bias term of TMSE$(\hat\beta^*)$. As we are interested in more narrow range than $(0, k_0)$, we obtain satisfactory one by this estimate.

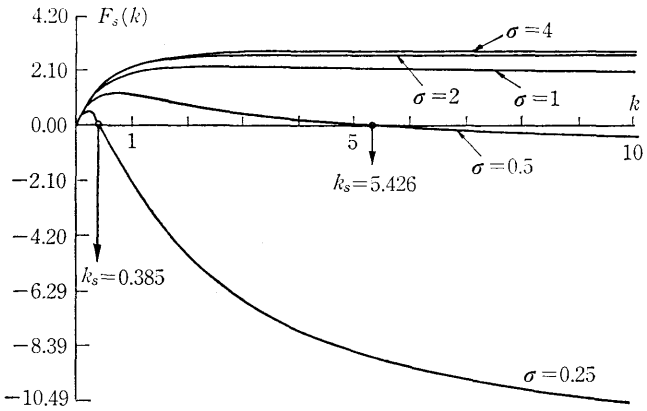
For example, let us assume that the actual regressors matrix X and $\beta'\beta$ as follows: $\beta'\beta=1$ and

$$X'=\begin{pmatrix}X_1'\\X_2'\\X_3'\end{pmatrix}=\begin{bmatrix}-1 & 1 & -1 & 1 & -1 & 1 & -(1-2\alpha) & (1-2\alpha)\\ -1 & 1 & -1 & 1 & (1-2\alpha) & -(1-2\alpha) & 1 & -1\\ -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1\end{bmatrix}. \tag{2.9}$$

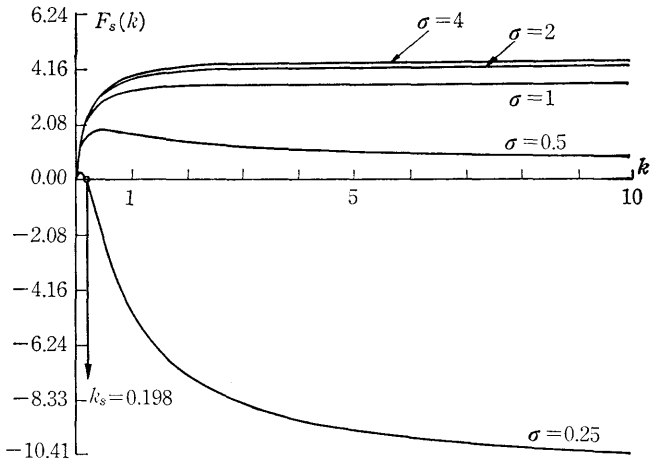
Then X_3 is orthogonal to X_1 and X_2 for all values of α, and the correlation between X_1 and X_2 is $r_{12}=\alpha/(1-\alpha+\alpha^2)$.

Here we consider three cases of $\alpha=0.1$, 0.5, and 0.9, or correspondingly $r_{12}=0.110$, 0.667, and 0.889. Further, the value of σ is assumed to take 0.25, 0.5, 1.0, 2.0, and 4.0.

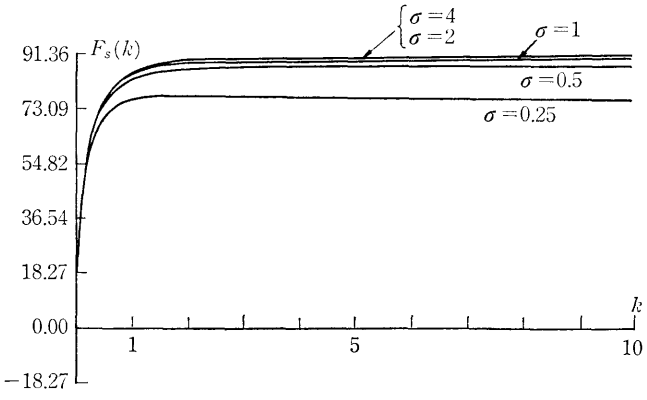
Figure 1 shows five curves of $F_s(k)$ on k corresponding to five values of σ for three cases of α. The cases of $\alpha=0.1$, $\sigma=0.25$; $\alpha=0.1$, $\sigma=0.5$; and $\alpha=0.5$, $\sigma=0.25$ are only three cases when we can obtain the finite solution k_s. Actually, these cases



(a) $\alpha=0.1$



(b) $\alpha=0.5$



(c) $\alpha=0.9$

Fig. 1. Admissible range of $\hat{\beta}^*$: Behaviours of the solution k_s for change of α and σ.

are close to othogonal ones of X.

The values of k_s become infinite as α increases to unity. Therefore, for ill-conditioned data we can see that the admissible range of $\hat{\beta}^*$ could be widely $(0, \infty)$ on k.

3. Choice of the optimal ridge parameter

The choice of any specific k in the admissible range of $\hat{\beta}^*$ is a main purpose of this section. Selection procedures of the k to be described below is relative to some criteria as mentioned in section 1. Here we sort out these procedures from four orientations under which the use of $\hat{\beta}^*$ is intended.

A. Reduction-criteria of $\hat{\beta}'\hat{\beta}$.

(1) Stability condition.

This criterion has been suggested by Hoerl (1962) and formulated by Gotô (1972). Now we denote the residual sum of squares of $\hat{\beta}^*$ by $L_k^2=\mathrm{RSS}(\hat{\beta}^*)$, and its squared length by $l_k^2=(\hat{\beta}^*)'(\hat{\beta}^*)$. Then a value of k which maximizes the second derivative of L_k, that is, the square root of RSS $(\hat{\beta}^*)$, taken with respect to l_k is chosen as an optimal ridge parameter, where the second derivative of L_k with respect to l_k is

$$d^2L_k/dl_k^2=[l_k^2\{\hat{\beta}^{*\prime}(X'X+kI_{p_0})^{-1}\hat{\beta}^*\}^{-1}-k-k^2l_k^2(L_k^2)^{-1}]/L_k\,. \tag{3.1}$$

The value of k may be best in a sense of the stability condition under which it minimizes the k-fold proportion of the decrease of the square root of the squared length relative to the increase of the square root of RSS $(\hat{\beta}^*)$, since the first derivative of L_k with respect to l_k is

$$dL_k/dl_k=-k\frac{l_k}{L_k}\,. \tag{3.2}$$

That is, the value is minimum value of k at which the curve of kl_k/L_k is flat for variable k.

(2) Use of the unbiased estimate of $\beta'\beta$.

From the relation (2.8) unbiased estimate of $\beta'\beta$ is

$$Q_0=\hat{\beta}'\hat{\beta}-\sum_{i=1}^{p_0}\frac{s^2}{\lambda_i}\,. \tag{3.3}$$

McDonald and Galarneau (1975) have proposed that if Q_0 is positive, we should choose a value of k such that it satisfies $Q_0=\hat{\beta}^{*\prime}\hat{\beta}^*$. And otherwise, $k=k_u$, where k_u is zero for $Q_0<0$ and infinity for $Q_0=\infty$. However, the condition associated with the sign of Q_0 corresponds to the possibility of setting up the finite admissible range of $\hat{\beta}^*$ on k as shown in section 2. Actually the case where Q_0 takes positive value is not frequent for ill-conditioned data. Therefore, for most cases where we intend to use the ridge regression, we will take $k_u=0$ or $k_u=\infty$ in the use of this procedure. These choice do not always improve the TMSE performance of $\hat{\beta}^*$ relative to $\hat{\beta}$. In the following simulation studies, we have omitted the procedure by above reasons.

(3) A compromise between the ordinary ridge and the principal regression

The principal regression estimate of β is defined by

$$\hat{\beta}^{+}(p)=G_p\Lambda_p^{-}G_p'X'Y\,,\qquad p\leqq p_0\,. \tag{3.4}$$

The residual sum of squares for it is given by

$$\mathrm{RSS}\,(\hat{\beta}^{+}(p))=\mathrm{RSS}\,(\hat{\beta})+\hat{\beta}'(Z_p-I_p)'X'X(Z_p-I_p)\hat{\beta}\,, \tag{3.5}$$

where $G_p=[g_1, g, \cdots, g_p]$, $\Lambda_p=\mathrm{diag}\,[\lambda_1, \lambda_2, \cdots, \lambda_p]$ are given in section 1, and $Z_p=G_p\Lambda_p^{-1}G_p'X'X$. Thus, from (1.8) and (3.5) for given p, then we obtain a value of k such as satisfying the relation

$$(\hat{\beta}^{*}-\hat{\beta})'X'X(\hat{\beta}^{*}-\hat{\beta})=\hat{\beta}'(Z_p-I_p)'X'X(Z_p-I_p)\hat{\beta} \tag{3.6}$$

within the percent variation accounted for the principal regression estimate.

Similarly, if we can specify an appropriate p, then we can choose an optimal value of k such as satisfying

$$(\hat{\beta}^{*})'(\hat{\beta}^{*})=\hat{\beta}^{+}(p)'\hat{\beta}^{+}(p)\,,\qquad p\leqq p_0\,. \tag{3.7}$$

For the practical use of the procedure, choice of p will be very difficult for us. To take it up in detail is beyond the scope of this article.

B. TMSE oriented criteria.

(4) Harmonic mean of the generalized ridge parameter.

Here, we must come contact with the generalized ridge estimate,

$$\hat{\beta}^{*}(\Delta)=G\Delta\hat{\gamma}=\sum_{i=1}^{p_0}\delta_i\hat{\gamma}_i g_i\,, \tag{3.8}$$

where $\Delta=\mathrm{diag}\,(\delta_1, \delta_2, \cdots, \delta_{p_0})$ is a diagonal matrix whose diagonal elements are shrinkage parameters such that $0\leqq\delta_i\leqq 1$, $i=1, 2, \cdots, p_0$. The TMSE of $\hat{\beta}^{*}(\Delta)$ is given by

$$\mathrm{TMSE}\,(\hat{\beta}^{*}(\Delta))=\sum_{i=1}^{p_0}\left\{\delta_i\frac{\sigma^2}{\lambda_i}+(1-\delta_i)^2\gamma_i^2\right\}. \tag{3.9}$$

The values of δ_i which minimize the $\mathrm{TMSE}\,(\hat{\beta}^{*}(\Delta))$ are obtained as

$$\delta_i^{\mathrm{MSE}}=\frac{\lambda_i}{\lambda_i+\sigma^2/\gamma_i^2}\,,\qquad i=1, 2, \cdots, p_0\,. \tag{3.10}$$

For the general form of ridge estimate $\hat{\beta}_G^{*}=(X'X+K)^{-1}X'Y$, $K=\mathrm{diag}\,(k_1, k_2, \cdots, k_{p_0})$, the values of k_i which minimize the TMSE are equivalently given by

$$k_i^{\mathrm{MSE}}=\sigma^2/\gamma_i^2\,,\qquad i=1, 2, \cdots p_0\,. \tag{3.11}$$

Hoerl et al (1975) and Farebrother (1975) have proposed the harmonic mean of k_i^{MSE} as an optimum value of k, i.e.

$$\frac{1}{k}=\frac{1}{p_0}\sum_{i=1}^{p_0}\frac{1}{k_i^{\mathrm{MSE}}}\qquad\text{or}\qquad k=\frac{p_0\sigma^2}{\beta'\beta}\,. \tag{3.12}$$

In practice, since $\beta'\beta$ and σ^2 are unknown, we substitute $\hat{\beta}'\hat{\beta}$ and s^2 into (3.12) for $\beta'\beta$ and σ^2, and obtain the estimate of k.

The difference between $\mathrm{TMSE}\,(\hat{\beta}^{*})$ in (1.9) and $\mathrm{TMSE}\,(\hat{\beta}^{*}(\Delta^{\mathrm{MSE}}))$ is given by

$$\mathrm{DMSE}\,(k)=\sigma^2\sum_{i=1}^{p_0}\frac{(\lambda_i-k\tau_i)^2}{\lambda_i(\lambda_i+k)^2(1+\tau_i^2)}\,, \tag{3.13}$$

where $\tau_i^2=\lambda_i\gamma_i^2/\sigma^2$. Taking notice of the numerator in (3.13), we obtain a suggested choice of k on which $\hat{\beta}^*$ is most close to $\hat{\beta}^*(\Delta^{\mathrm{MSE}})$ or $\hat{\beta}_G^*$, i.e.

$$k_N=\sum_{i=1}^{p_0}\lambda_i\tau_i^2\Big/\sum_{i=1}^{p_0}\tau_i^4. \tag{3.14}$$

C. Prediction oriented criteria.

(5) Estimate of TMSEP.

An estimate of TMSEP ($\hat{\beta}^*$) in (1.10a) has been given as

$$T_1(k)=(Y-\hat{Y})'(Y-\hat{Y})+2s^2\,\mathrm{trace}\,(I_{p_0}+kA^{-1})^{-2} \tag{3.15}$$

by Allen (1974), where $\hat{Y}=X\hat{\beta}^*$ and this is equivalent to the prediction of the case where $X_0=X$ in (1.10a). A value of k to minimize the function $T_1(k)$ is chosen as a TMSEP-optimal value.

(6) Mallows' C_L statistic.

Mallows (1973) has given an estimate of the scaled total mean squared error of $\hat{\beta}^*$ by

$$C_L(k)=\frac{\mathrm{RSS}\,(\hat{\beta}^*)}{s^2}-n+2\sum_{i=1}^{p_0}\frac{\lambda_i}{\lambda_i+k} \tag{3.16}$$

and recommended the use of k to minimize $C_L(k)$ as an optimallike k of $\hat{\beta}^*$. He has further recommended to plot points of $(V_L(k), C_L(k))$, where

$$V_L(k)=1+\sum_{i=1}^{p_0}\left(\frac{\lambda_i}{\lambda_i+k}\right)^2 \tag{3.17}$$

and to select the point which gives the minimum of $C_L(k)$ on $V_L(k)$ by inspection of the plotting-trace. We should note that the minimum of $C_L(k)$ on k consists with that on $V_L(k)$. This rule suggests that the "to minimize C_L" leads to shrinkage of the ordinary least square estimates towards their average.

(7) PRESS.

Allen (1974) has suggested a statistic of TMSEP in (1.10a) by using cross-validatory method. Now we will denote the l-th row of X by X_l and the corresponding component of Y by Y_l, $l=1, 2, \cdots, n$. Then for $(n-1)\times p_0$ matrix $X_{(l)}$ by exclusion of the l-th row X_l from the $n\times p_0$ matrix X, and $(n-1)\times 1$ vector $Y_{(l)}$ by exclusion of the l-th component from Y, an ordinary ridge estimate of β can be obtained as follows;

$$\hat{\beta}_{(l)}^*=(X_{(l)}'X_{(l)}+kI_{p_0})^{-1}X_{(l)}'Y_{(l)}, \qquad l=1, 2, \cdots, n. \tag{3.18}$$

The l-th response which has been excluded in above estimation of regression coefficients is predicted by making use of this estimate at X_l such as

$$\hat{Y}_l^*=X_l'\hat{\beta}_{(l)}^*, \qquad l=1, 2, \cdots, n. \tag{3.19}$$

The predicted sum of squares of error is defined by

$$\mathrm{PRESS}=\sum_{l=1}^{n}(Y_l-\hat{Y}_l^*)^2. \tag{3.20}$$

The optimal value of k is selected by minimization of PRESS. In other words, each observation is predicted by making use of the other $(n-1)$ observations, and this fact makes PRESS interesting because it simulates prediction.

Dirct calculations of PRESS will require a great deal of labours, computing times and costs. Here we shall give the alternative form of PRESS in order to yield the easy calculation and insight for it. In place of (3.19), we will put the ordinary ridge prediction at X_l;

$$\hat{Y}_l=X_l'\hat{\beta}^*. \tag{3.21}$$

Taking notice of normal equations for $\hat{\beta}^*$ and $\hat{\beta}^*_{(l)}$ which are

$$(X'X+kI_{p_0})\hat{\beta}^*=X'Y, \tag{3.22}$$

$$(X'X-X_lX_l'+kI_{p_0})\hat{\beta}^*_{(l)}=X'Y-X_lY_l, \tag{3.23}$$

we have the following relation,

$$Y_l-\hat{Y}_l^*=\frac{Y_l-\hat{Y}_l}{1-X_l'(X'X+kI_{p_0})^{-1}X_l} \qquad l=1, 2, \cdots, n. \tag{3.24}$$

Thus, PRESS is written by

$$\text{PRESS}=\sum_{l=1}^{n}\frac{(Y_l-\hat{Y}_l)^2}{\{1-X_l'(X'X+kI_{p_0})^{-1}X_l\}^2} \tag{3.25}$$

or

$$\text{PRESS}=(Y-\hat{Y})'D^{-2}(Y-\hat{Y}), \tag{3.26}$$

where $\hat{Y}$ is $n\times 1$ vector whose l-th component is given by (3.21) and D is $n\times n$ diagonal matrix such that

$$D=\text{Diag}\,(I_n-X(X'X+kI_{p_0})^{-1}X'). \tag{3.27}$$

It is obvious at a glance that D is a special weighted matrix in the weighted residual sum of squares.

The form of PRESS in (3.6) suggests another case of weighted matrix. That is, a modified PRESS may be proposed in such a way that

$$\text{MPRESS}=(Y-\hat{Y})'D^{-1}(Y-\hat{Y}) \tag{3.28}$$

by making use of D^{-1} instead of D^{-2} in (3.26).

D. Rule of thumb.

(8) Ridge trace and some criteria on it.

Hoerl and Kennard (1970b) suggest that the best method for achieving a better estimate $\hat{\beta}^*$ on TMSE is to employ a "ridge trace on k", i.e., a plot of the estimates $\hat{\beta}_i^*$, $i=1, 2, \cdots, p_0$, versus k $(0<k\leqq 1)$ to select a single value of k on which the system will stabilize and have the characteristics of an orthogonal system. Regression coefficients for this value of k will not have unreasonable absolute values and some coefficients of these with apparently incorrect signs at $k=0$ will have changed signs. The residual sums of squares for them will not have been inflated to an unreasonable value. By inspecting the trajectory of regression coefficients on the trace, we may

grasp the structure of the factor space and the sensitivity of regression result for special data set.

On the trace, if the system has the characteristics of an orthogonal system, then we have

$$\frac{d\hat{\beta}^*}{dk}=\frac{d\hat{\beta}_0^*}{dk}, \tag{3.29}$$

where $\hat{\beta}_0^*=\frac{1}{1+k}\hat{\beta}$. Thus we obtain

$$(X'X+kI_{p_0})^{-1}X'X(X'X+kI_{p_0})^{-1}-\frac{1}{(1+k)^2}I_{p_0}=0. \tag{3.30}$$

This suggests VIF (Variance Inflation Factor) criterion given by Marquardt (1970). That is, a value of k should be chosen in such a way that

$$\begin{aligned}\mathrm{VIF}\,(k)&=\mathrm{Trace}\,(X'X+kI_{p_0})^{-1}X'X(X'X+kI_{p_0})^{-1} \\ &=\mathrm{Trace}\,A,\ \text{say}\end{aligned} \tag{3.31}$$

equals to $p_0(1+k)^{-2}$ or p_0 (The rule is called SVIF in the latter).

Let us present the ridge trace in a parameter space of $\beta=(\beta_0, \beta_2, \cdots, \beta_{p_0})'$, and take into account of the confidence region of β with $100(1-\alpha)$ % confidence coefficient given by

$$(\hat{\beta}^*-\beta)'X'X(\hat{\beta}^*-\beta)\leqq p_0 s^2 F_{p_0,\,n-p_0-1}(\alpha) \tag{3.32}$$

in which $\hat{\beta}^*$ is used instead of $\hat{\beta}$. Then it may be proposed that a value of k should be selected such as $\hat{\beta}^*$ satisfying the above ellipsidal bound, i.e.

$$(\hat{\beta}-\hat{\beta}^*)'X'X(\hat{\beta}-\hat{\beta}^*)=p_0 s^2 F_{p_0,\,n-p_0-1}(\alpha). \tag{3.33}$$

Obenchain (1976) has called the probability level α the associated probability of $\hat{\beta}^*$ and Gotô (1972) has illustrated the trajectory in case of an ill-conditioned problem of $p=2$.

The associated probability is the estimated probability that the hyperellipsoid (3.33) centered at $\hat{\beta}$ which is just large enough that $\hat{\beta}^*$ lies upon its surface does not cover the true unknown β. Since the associated probability can be interpreted to be the probability that $\hat{\beta}$ is more away from β along some direction than from $\hat{\beta}^*$, it may be appropriate for the performance measure of $\hat{\beta}^*$.

On the other hand, Obenchain (1976) has proposed and recommended "ridge trace on m-scale", where m is called the multicollinearity allowance and defined by

$$m=p_0-\delta_1-\delta_2-\cdots-\delta_{p_0} \tag{3.34}$$

and

$$\delta_i=\frac{\lambda_i}{\lambda_i+k},\qquad i=1, 2, \cdots, p_0. \tag{3.35}$$

He has maintained that there are four distinct reasons for preferring ridge trace plotted against m to those plotted against k; that is, generality and comparability,

linear stability, Bayesian posterior precision interpretation, and rank deficiency interpretation. In our further investigations, we will adopt the ridge trace on m-scale. Note that for plotting of points $(m, \hat{\beta}_i^*)$ we must solve the following iterative equation to find each value of k and the corresponding $\hat{\beta}_i^*$ for each value of m ;

$$k^{(s)}=k^{(s-1)}-(f'(k^{(s-1)}))^{-1}f(k^{(s)}), \tag{3.36}$$

where $k^{(s)}$ is the s-th step solution with initial value $k^{(0)}=0$, and $f'(k)$ is the first derivative respect to k of function,

$$f(k)=m-p_0-\sum_{i=1}^{p_0}\lambda_i(\lambda_i+k)^{-1}. \tag{3.37}$$

(9) SSCBC.

Taking notice of the fact that the correlation matrix of $\hat{\beta}^*$ is given by

$$R(\hat{\beta}^*)=[\text{Diag } A]^{-1/2}A[\text{Diag } A]^{-1/2}, \tag{3.38}$$

we can formulate SSCBC (Sum of Square of Correlations Between Coefficients) criterion suggested by Obenchain (1975) where A is given by (3.31). This measure can be written as

$$\text{SSCBC}=\sum_{i\neq j}^{p_0}\sum^{p_0} R_{ij}^2(\hat{\beta}^*), \tag{3.39}$$

where $R_{ij}(\hat{\beta}^*)$ is the (i, j) element of the matrix $R(\hat{\beta}^*)$. Thus a criterion which we adopt here is to choose a value of k to minimize the SSCBC.

4. Numerical investigation of the ridge estimate

4.1 Preliminary

Here, we consider the way for evaluating a new proposal made such as in the ridge regression method. We want to emphasize that it should consist of four phases. These phases are analogies to four phases for clinical evaluation of "new drug" set-up in "Food and Drug Administration" Principles. We think that a new method or a new program should be evaluated rigorously and systematically such as in a "new drug", since it has not only some effect informations but also "side effects". In general, since any proposer of a new method or a new technique tends to emphasize certan effects of it rather than defects or hidden problems, it will be always necessary for us to evaluate the proposal as rigorously or objectively as possible. This is obvious if we take into account for the fact that there still exist numerous gaps both in the theory and between the theory and its application in practice.

In phase-one study we take aim at illustrating or illuminating any characters or effects of the proposal by making use of some simple example which give the easy interpretation of results, in order to give the insight for follow-up developments and other applications of it. Thus it may be sufficient to grasp the point or intension of the proposal through the illustrative examples. That is, a new drug which has no effect would not be useful, even if it has sufficient safety for general use. In this phase of our proposal, we will employ data used by Hoerl (1962) as an example.

Phase-two study corresponds to a "dose range" study in case of the new drug development. The object in this phase is to clarify the useful range of the proposal, and to suggest the substantial effect of it by comparison with "placebo" which corresponds to the ordinary least square estimate in our problem to evaluate the ridge estimate. Usually, we can obtain a prospect of some performances of the proposal in this stage and see most characteristics of the new method. Thus, the phase-two study is the most important one to the forthcoming development of the proposal.

Phase-three study corresponds to a large scale controlled clinical trial to the new drug over multi-clinics. The object in the phase is to confirm the effects or usefulness of the proposal, suggested in the previous two phases and by executing a large scale Monte-Carlo simulation experiment to examine defects or hidden problems of it. In this stage, it is necessary to construct a design or protocal most carefully in order to test and confirm a hypothesis under which the proposal has some relative effects. We should emphasize to construct the design of the Monte-Carlo simulations most cautiously in this stage. We will propose a design of the phase-three study and execute one case of the simulation following to the design.

Phase-four study is the practical field study of administration of the new drug, for example, through the approval in the "Ministry of Public Welfare". In this phase, it is required to follow-up the performance of the proposal in practice and to explore new problems which have been unexpected until the study in many applications to real data. In our following evaluation for the ridge estimate, we can not put the study of the last phase into practice by means of Monte-Carlo simulations. That is, in what follows, we will give the Monte-Carlo evaluation for the ridge estimate from the phase-one study to the phase-three one.

We begin with some preparations, of detecting and checking the degrees of near-multicollinearity for putting our simulations into practice.

4.2 Degree ot the near-multicollinearity

Thouh the multicollinearity could be considered to result primarily owing to four sources, namely, an incomplete model specification or an over-defined model, sampling technique, physical constraints on the model or in the population, and other apparent constraints on precision of a calculating method or rounding error, the problem for it is not the one of existence but of degree. Hence, we will review some methods for detecting multicollinearity and adopt a few measures in these methods as a check of multicollinearity in our studies.

(1) Farrar and Glauber's statistics.

A number of tests on multicollinearity which are applied very easily have been treated in an article by Farrar and Glauber (1967). The tests have been designed for the detection of multicollinearity, localization of the multicollinearity, and the choice of the multicollinearity profile. Now we assume that the $p_0 \times 1$ vector x_i, whose trasposed vector is the i-th row of X, follows the normal p_0-dimensional distribution and also without loss of generality that the x_i-elements have been normalized such that $X'X$ takes the form of the correlation matrix. It is known that the determinant of $X'X$ becomes zero when there is complete dependency or multicollinearity and

becomes one when the regressors are orthogonal. That is,

$$0 \leqq |X'X| \leqq 1.$$

For the test of detection of multicollinearity, the statistics

$$\chi_0^2 = -\left\{n-1-\frac{2p_0+5}{6}\right\} l_n |X'X| \tag{4.1}$$

is provided from above property. This statistic has the central chi-squared distribution with degrees of freedom $\frac{1}{2}p_0(p_0-1)$ under the hypothesis that the regressors are mutually independent. This result has originally been proved by Bartlett (1950) who used the result obtained by Wilks (1932).

If the test described above indicates the existence of multicollinearity, the second step will be the localization of multicollinearity. The test statistic for this is given by

$$W_i = \left(\frac{n-p_0-1}{p_0-1}\right)(A_{ii}-1), \qquad i=1, 2, \cdots, p_0, \tag{4.2}$$

where A_{ii} is the i-th diagonal element of the inverse matrix of $X'X$. This statistic has the central F-distribution with degrees of freedom (p_0-1) and $(n-p_0-1)$ under the hypothesis that the i-th regressor is independent of other regressors. Here again Wilks's result (1932) has been used. This test should be applied for each variable in order to trace out the multicollinear subset.

Having found the multicollinear subset, it often seems useful to study the pattern of the mutual relations in this subset. The statistic for the test concerning the multicollinearity pattern is given by

$$t_{ij} = \frac{A_{ij}^*(n-p_0-1)^{1/2}}{(1-A_{ij}{}^{*2})^{1/2}}, \qquad i \neq j=1, 2, \cdots, p_0, \tag{4.3}$$

where we denote by A_{ij} the (i, j) non-diagonal element of the inverse matrix of $X'X$ and we put

$$A_{ij}^* = -\frac{A_{ij}}{(A_{ii})^{1/2}(A_{jj})^{1/2}}, \qquad i \neq j=1, 2, \cdots, p_0. \tag{4.4}$$

This statistic has the central t-distribution with degrees of freedom $(n-p_0-1)$, under the hypothesis that the i-th regressor and the j-th regressor are mutually independent. This test should be also applied to each pair of regressors in the multicollinear subset.

(2) The variance inflation factor.

The variance inflation factor given by Marquardt (1970) can be utilized again as a measure of multicollinearity in the case of $k=0$. That is, this measure is the diagonal element A_{ii} of the inverse matrix of $X'X$. This is also represented as follows;

$$A_{ii} = (1-R_i^2)^{-1}, \qquad i=1, 2, \cdots p_0, \tag{4.5}$$

where R_i^2 is the coefficient of determination when the i-th regressor is regressed on the remaining (p_0-1) regressors. When R_i^2 is near unity, A_{ii} is very large. If there

is a single multicollinearity, the large A_{ii} indicate which regressors are linearly related.

(3) The spread and profile of the eigenvalues of regressors matrix.

The variance inflation factor can tell which regressors are involved in a single multicollinearity, but not specify the coefficient in the linear relationship of regressors. A more informative measure is to utilize the smallest eigenvalue λ_{p_0} and the corresponding eigenvector g_{p_0} of $X'X$. The closer λ_{p_0} is to zero the stronger is the linear relationship among the columns of X. Hence, the elements of g_{p_0} moreover give the coefficients in the linear relatioships of p_0 regressors according to the multicollinearity.

4.3 A phase-one study of the ridge estimate

Before carrying out our simulation experiments, we will present some example to illustrate how ridge regression gives helpful results in a wide variety of circumstances. Discussion of data in the example will give some suggestions to us as a substitution for phase-one study.

When it is assumed that a process will behave according to the relationship

$$Y=10+2X_1+3X_2+5X_3 \tag{4.6}$$

on three separate process variables, X_1, X_2, and X_3, two cases of artificial data have been given by Hoerl (1962). The case-two of the data given by Hoerl is slightly modified from the case-one such that the same spread of X values is retained, but the intermediate values are bunched more in the middle range, and the same random errors are added.

Then, two regression equations have been obtained as follows;

$$\hat{Y}=10.24+0.87X_1+0.93X_2+8.15X_3 \tag{4.7}$$

$$\hat{Y}=12.11+8.27X_1-5.52X_2+6.39X_3 \tag{4.8}$$

by means of the least square estimation in the both cases. Which one should be used as a credible relatienship? Hoerl (1962) has claimed that for the case-one the derived relationship (4.7) agrees tolerably well with the original model (4.6), considering the quality and amount of data used, and that for the case-two the estimation (4.8) of the real process becomes intolerably bad under this circumstances. And he has applied the ridge regression only to the case-two, and chosen the optimal ridge estimate based on the stability condition (1) in section 3.

However, in so far as our reexamination being of these results for both two cases, we can not agree with his interpretation by which only the case-one is not intolerably bad and the adjustment by the ridge method is not required. In fact, we can confirm using the methods in section 4.2, that the near-multicollinearity occurs for both cases of the data. Some checks on the near-multicollinearity are given in Table 1 and the near-multicollinearity is highly detected for both cases of the data. That is, it is considered that the inflation of the estimated regression coefficients occurs in both cases, and that the inflation is attributable to the near-multicollinearity

Table 1. Check or detection of near-multicollinrarity: Hoerl's data.

(1) Farrar and Glauber's statistics		
	(case one)	(case two)
	Detection of multicollinearity	
	$\chi_0^2=32.31$ (d.f. $=3$)	$\chi_0^2=39.78$
	Localization of multicollinearity	
	$W_1=50.18$ (d.f. $=(2,7)$)	$W_1=143.24$
	$W_2=29.48$	$W_2=116.33$
	$W_3=30.65$	$W_3=22.84$
	The multicollinearity pattern	
	(d.f. $=7$)	
	$[t_{ij}]=\begin{bmatrix} 0.00 & 2.07 & 2.17 \\ 2.07 & 0.00 & 0.44 \\ 2.17 & 0.44 & 0.00 \end{bmatrix}$	$[t_{ij}]=\begin{bmatrix} 0.00 & 5.66 & 1.26 \\ 5.66 & 0.00 & -0.10 \\ 1.26 & -0.10 & 0.00 \end{bmatrix}$
(2) Variance inflation factor		
	$A_{11}=15.33$	$A_{11}=41.93$
	$A_{22}=9.42$	$A_{22}=34.24$
	$A_{33}=9.76$	$A_{33}=75.25$
(3) Eigenvalues of X		
	$\lambda_1=2.867$	$\lambda_1=2.887$
	$\lambda_2=0.090$	$\lambda_2=0.099$
	$\lambda_3=0.043$	$\lambda_3=0.014$

among regressors as detected above. For the case-one, it will be natural to interpret that the signs of the estimated regression coefficients become the same as in the true relationship only by chance. Unfortunately, Hoerl (1962) has given the miscalculated estimates of the regression coefficients, for which we have given the correct estimates in (4.7) and (4.8).

Now let us apply the ridge regression to both cases of the data. For selection of the optimal ridge estimate, we use the methods of the stability condition, SSCBC, SVIF, as the reduction oriented criteria of $\hat{\beta}'\hat{\beta}$, $k=p_0s^2/\hat{\beta}'\hat{\beta}$ as the minimum TMSE oriented criteria, and $C_L(k)$ and PRESS as the prediction oriented criteria in section 3. In Figure 2(a) and 2(b) are shown the sketch of the ridge trace on the horizontal scale m for the data in the cases-one and two, respectively, together with the optimal value of k in each selection method in the contrast with the result of the least square regression. We evaluate $\overline{\text{TMSE}}$ for the reduction criteria of $\hat{\beta}'\hat{\beta}$ and the minimum TMSE oriented criteria as the performance of $\hat{\beta}^*$ as described in section 1, and $\overline{\text{TMSEP}}$ for the prediction oriented criteria. In Figure 2, $\overline{\text{TMSEP}}\,(i\rightarrow j)$ $(i, j=1, 2)$ means $\overline{\text{TMSEP}}$ for the value of $\hat{Y}$ which has been predicted in the sample of the case j by use of $\hat{\beta}^*$ estimated in one of the case i. In other words, the measure shows the performance of prediction outside the index data by use of $\hat{\beta}^*$.

For both cases, the ridge regression has yielded better performance than the ordinary least square regression in both senses of $\overline{\text{TMSE}}$ and $\overline{\text{TMSEP}}$.

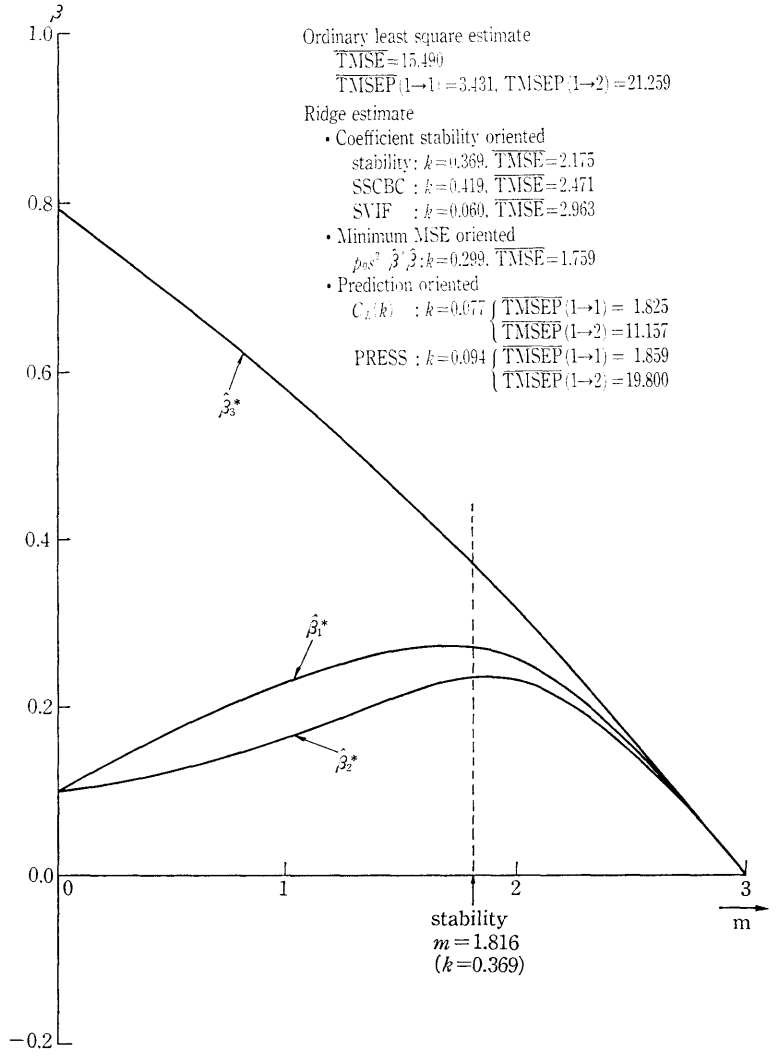


Fig. 2(a). Ridge trace on m-scale: Results of the ridge regression for Hoerl's case-one data.

Clearly from a glance of these figures, it is found that the ridge regression, based on the optimality criteria, gives a good fit to the data for both cases. Further, taking notice of the results related to the stability condition recommended by Hoerl (1962), we can see that the estimated "optimal" ridge parameter is $k=0.37$ for the case-one and $k=0.45$ for the case-two, and the corresponding regression equations are obtained as

$$\hat{Y}=13.06+2.40X_1+2.25X_2+3.79X_3 \tag{4.9}$$

for the former and

$$\hat{Y}=14.00+2.70X_1+2.11X_2+3.23X_3 \tag{4.10}$$

for the latter. These equations are nearly equal and close to the true relationship (4.6) relative to the above least square relationships in (4.7) and (4.8). This fact

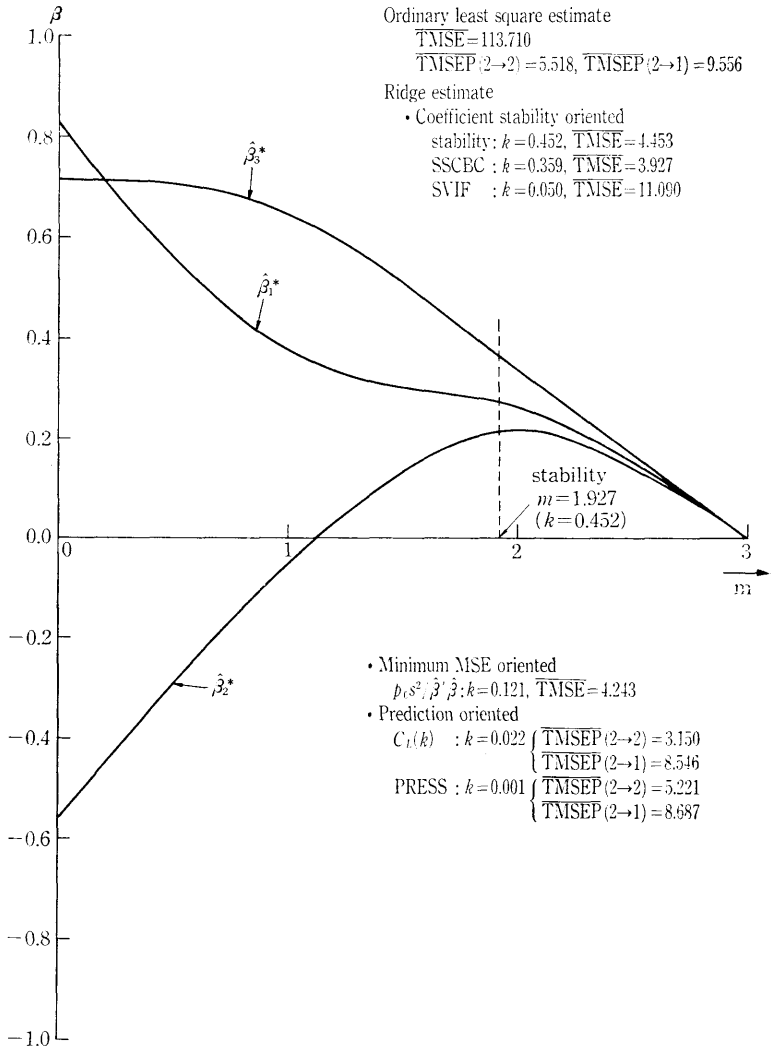


Fig. 2(b). Ridge trace on m-scale: Results of the ridge regression for Hoerl's case-two data.

suggests that the ridge regression may give "stable" common regression coefficients, irrespective of sampling structures of the data. Unfortunately, Hoerl (1962) has not become aware of the above fact, which should be considered to be natural for emphasizing his claim on the ridge regression.

In addition, if we pay attention to the values of k chosen by above methods and the higher degree of near-multicollinearity for the case-two, then we have a suggestion that the ridge estimate based on the prediction oriented criteria have a tendency of getting much closer to $\hat{\beta}$ than on the other criteria, and that k by the stability condition will present a measure of sensitivity which gives the degree of adjustment to each sampling structure of the data since the value of k for the case-two is greater than for the case-one.

That is, the stable ridge estimate as given by the stability condition may be useful for prediction along minor axes of X.

On the other hand, the rules of thumb as SVIF or SSCBC and the stability condition do not always intend to minimize the values of $\overline{\text{TMSE}}$. Thus, we have examined the relation between these criteria functions or statistics and the values of $\overline{\text{TMSE}}$, and shown it for the data in the case-two in Figure 3. Similarly in Figure 4 traces of criteria functions $C_L(k)$ and PRESS are sketched against the values of $\overline{\text{TMSEP}}$

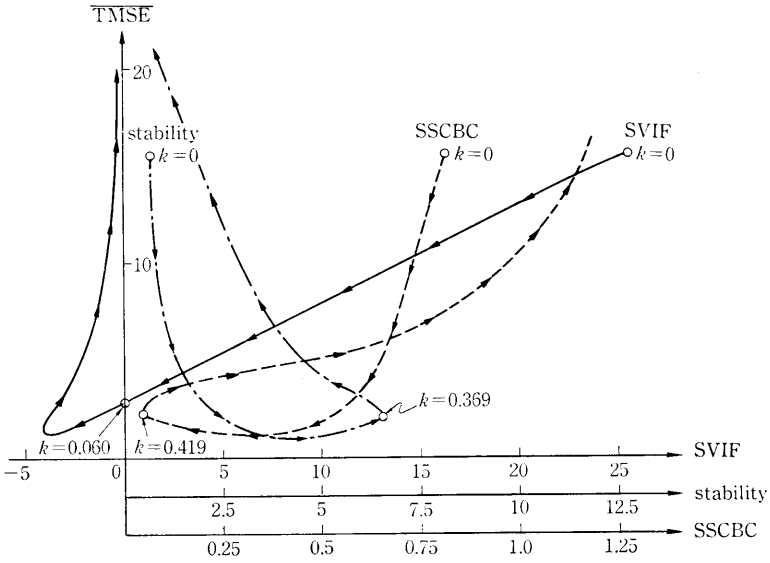


Fig. 3. Relation between the statistics in rule of thumb and $\overline{\text{TMSE}}$: Hoerl's case-two data.

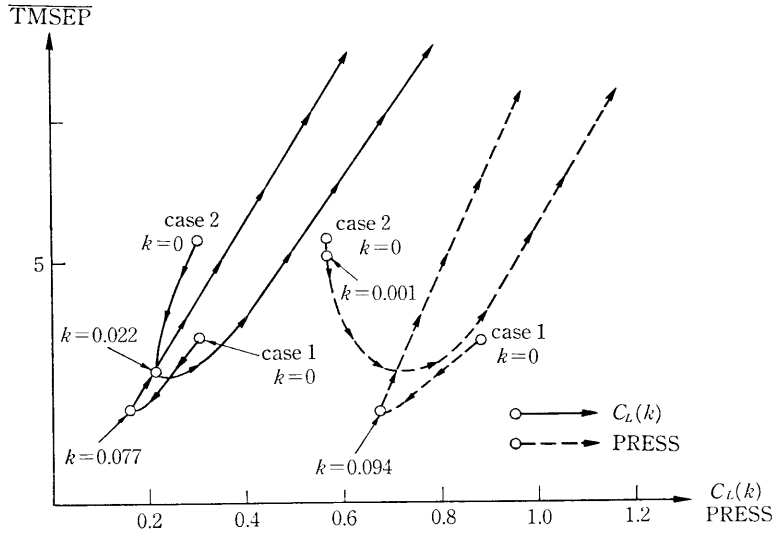


Fig. 4. Relation between $C_L(k)$ and PRESS and $\overline{\text{TMSEP}}$: Hoerl's data.

associated with various k for both cases of the data. From these figures, we can see that the admissible range of $\hat{\beta}^*$, which have been attained by use of the stability condition, SSCBC, SVIF on $\overline{\text{TMSE}}$ may be wide and the one based on $C_L(k)$ and PRESS on $\overline{\text{TMSEP}}$ are relatively narrow. That is, the admissible range of k which gives the predictive stability is narrow, and so the value of k at which minimization of the criterion function can be arrived, may be recommended as the optimal value for the prediction oriented use.

Summing up our reexamination of Hoerl's example, we can obtain a suggestion such that the ridge regression not only fits the data but gives simultaneously a stable estimate of each regression coefficient. This may be also the crux of the matter calimed by Hoerl (1962) who is an original proposer of the ridge regression.

4.4 A design for our phase-two and phase-three studies

The performance of the ridge estimate to be evaluated is dependent on the variance of the random error σ^2, the correlations among the regressors and the unknown regression coefficient vector β. In fact, since the performance of the estimates can be observed as one or two of these factors is changed while the remaining are fixed in the simulation, it is necessary to give the specification of these parameters as carefully and effectively as possible.

(1) Specification of β.

If we have specified each set of regressors whose specifications or generations in the form of regressors matrix will be given in section 4.5, then it is required to choose and specify the true coefficient vector β, such that the improvement of the ridge estimate can be evaluated as effectively as possible and simulation results can give some insights for the following studies.

Now, we shall prepare some properties of the ordinary least square estimate and the ordinary ridge estimate on TMSE which is useful to specify β. Let ϕ denote the cosine of the angle btween the vector β and the eigenvector g_{p_0} corresponding to the smallest eigenvalue λ_{p_0} of $X'X$, i.e.

$$\phi=\frac{\sum_{i=1}^{p_0}\beta_i g_{0i}}{\left(\sum_{i=1}^{p_0} g_{0i}^2\right)^{1/2}\left(\sum_{i=1}^{p_0}\beta_i^2\right)^{1/2}}, \tag{4.11}$$

where g_{0i} and β_i are the i-th components g_{p_0} and β, respectively.

PROPERTY 1: TMSE$(\hat{\beta})$ is independent of the specification of β corresponding to ϕ.

PROPERTY 2: TMSE$(\hat{\beta}^*)$ given in (1.9) is dependent on the specification of β corresponding to ϕ if other parameters σ^2, k and λ_i, $i=1, 2, \cdots p_0$, were known. Hence we obtain two important choice: that is, (i) TMSE$(\hat{\beta}^*)$ is minimized when β is proportional to the eigenvector g_1 corresponding to the largest eigen value λ_1 of $X'X$, i.e. $\phi=0$, subject to the constraint that $\beta'\beta$ is a known constant, and (ii) TMSE$(\hat{\beta}^*)$ is maximized subject to above constraints, when β is proportional to g_{p_0} associated with $\phi=1$.

To obtain a set of regression coefficients in our simulation at first we choose a

value l_0^2 for the squared length of the regression coefficient vector, taking account of the magnitude of the variance σ^2 of the random error to be specified in the next subsection. Secondly, given a value of l_0^2, p_0 candidates of regression coefficients are chosen from the relation of (4.11) or

$$\beta_i'=g_{0i}\cos\theta+g_{1i}\sin\theta\,,\qquad i=1,\,2,\,\cdots\,p_0\,, \tag{4.12}$$

and thus the regression coefficients whose squared length are specified to l_0^2 are calculated by

$$\beta_i=\frac{l_0}{l_1}\beta_i'\,,\qquad i=1,\,2,\,\cdots,\,p_0\,, \tag{4.13}$$

where θ is the angle between the vector β and the vector g_{p_0} corresponding to ϕ, and $l_1^2=\sum_{i=1}^{p_0}(\beta_i')^2$. Let us put $l_0^2=1$ in order for the result to be easy to interpret since the regressors are normalized in such a way that $X'X$ takes the form of correlation matrix. In the study we take only three cases of ϕ, namely $\phi=0$, $\phi=0.5$, and $\phi=1$.

(2) Variance of the random error.

To generate the random error ε_i, $i=1,\,2,\,\cdots,\,n$, at first we have to specify the variance σ^2 of the random error. Usually, three cases are considered to specify σ^2, for evaluating the performance of the ridge regression: that is, (i) specification of $1/\sigma^2$, which is independent of β, (ii) specification of $\beta'\beta/\sigma^2$, which is dependent on the squared length of β, and (iii) specification of $\tau^2=\beta'X'X\beta/\sigma^2$, which is the non-centrality parameter of the non-central chi-squared distribution with degrees of freedom p_0. Especially, using the notation of τ_i^2 in section 3 we can rewrite $\tau^2=\sum_{i=1}^{p_0}\tau_i^2$. In general, since the performance of the ridge estimate is dependent on β, σ^2 and X interactively, it will not be good to observe the performance of it as one of these factors is changed independently while the remaining two are fixed. Thus, from this point of view it may be considered to be more appropriate to evaluate the ridge estimate based on (ii) or (iii) than on (i). However, if the squared length of β is specified to $l_0^2=1$ as in our simulation, the specification of (i) and (ii) are the same. Thus the random error is computed by generation of n random normal variates with mean zero and standard deviation σ based on this specification.

(3) Generation of the regressors matrix and its sampling structure.

Since the conditioning of $X'X$ is known to have an effect on how much bias is necessary to improve an estimate of β, it is the most important point to specify the correlation structure of X, generate the corresponding X, and take up some sample regressors matrices, which have the same correlation structure as X, associated with the index samples and the follow-up samples of our simulations.

Then we could consider two directions in constructing the matrix of X. That is, (1°) X is constructed by a transformation

$$X_{ij}=(1-\rho^2)^{1/2}Z_{ij}+\rho Z_{ip_0}\qquad \begin{matrix} i=1,\,2,\,\cdots,\,n\,,\\ j=1,\,2,\,\cdots,\,p_0-1\,, \end{matrix} \tag{4.14}$$

or a special specification as in (2.9), as a departure from a $n \times p_0$ orthogonal matrix Z, where (i, j) element Z_{ij}'s are usually given by n independent normal random numbers, and where ρ is a certain constant within (0, 1) whose squared value is the correlation between any two regressors. For our phase-two study, we employ conveniently the matrix of (2.9) as X. And (2°) the correlation structure of an ill-conditioned data, for example, referring to a structure of X in literatures, is given and then the sample regressors matrix is obtained by generating n pseudo-random numbers which have the p_0-variate normal distribution with the same correlation structure as X. For (2°), as the ill-conditioned correlation structure among regressors, we may adopt various data such as, for example, the "cement" data published in Hald (1952), p. 647, the ten factors data by Gorman and Toman (1966) which has been analysed by many other authors and by many approaches or proporsals with various different conclusions, the "gas mileage" data by Hocking (1976), the "air pollusion" data by McDonald and Schwing (1973), and so on. In our phase-three study, we adopt a following correlation structure of $p_0=5$, i.e.

$$R=\begin{bmatrix} 1 & 0.3 & 0.9 & 0 & 0 \\ 0.3 & 1 & 0.4 & 0 & 0 \\ 0.9 & 0.4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{4.15}$$

which includes both orthogonal and correlated relations among regressors.

Construction of the index sample X_I and the follow-up sample X_F with the same structure as X are given by use of the so-called "CADEX" algorithm provided by Kennard and Stone (1969) in mixture design problems, and the response vector is then computed by use of the model of (1.2) corresponding to X_I or X_F. CADEX is an abbreviation of "Computer Aided Design of Experiments" algorithm, and the philosophy is as follows: given a set of candidate points which cover the feasible experimental region, a good experimental design is a set of points which uniformly cover the available region. Now if there are p_0 factors $(X_1, X_2, \cdots, X_{p_0})$ that can be controlled, then an experimental design is to be selected from a set of N points in the p_0-dimensional space defined by the factors. We denote these N points as candidates and represent them as

$$X_{(N)}=\begin{bmatrix} x_{11} & x_{12} \cdots x_{1p_0} \\ x_{21} & x_{22} \cdots x_{2p_0} \\ \vdots & \vdots \quad \vdots \\ x_{N1} & x_{N2} \cdots x_{Np_0} \end{bmatrix}. \tag{4.16}$$

A design will be $n \leqq N$ distinct points chosen from the candidates and the design points are chosen sequentially. At each stage in the choice sequence, the aim will be to have the points in the design uniformly spaced over the space defined by the candidates. CADEX begins by picking the two points which are farthest apart and then adds subsequent points which are farthest from the points already in the design.

In fact, let

$$D_{ts}^2=(X_t-X_s)'(X_t-X_s)=\sum_{j=1}^{p_0}(X_{tj}-X_{sj})^2 \tag{4.17}$$

be the squared distance from point t to s and further P_{i*}, $i=1, 2, \cdots, l$, $(l\leqq n)$ be l points which have been already assigned to the design. Then we denote the squared distance from candidate point X_t, not yet in the design, to the nearest design point by

$$\Delta_t^2(l)=\min\{D_{1*t}^2, D_{2*t}^2, \cdots, D_{l*t}^2\}\ . \tag{4.18}$$

We choose the $(l+1)$-st point in the design from among the remaining $(N-l)$ candidates by making use of the criterion

$$\Delta_{l+1}^2=\max_{t\neq l*}\{\Delta_t^2(l)\}\ . \tag{4.19}$$

This implies that among those remaining we choose the point which is farthest from existing design points. If on starting the procedure it is known a priori that certain candidates must be included in the design, we can select the remaining points necessary to make a design of size n, using the relations from (4.16) to (4.18). Otherwise, by choosing the two candidate points X_t and X_s which are farthest apart as mentioned above, we start by calculating the distance,

$$D_{\max}^2=\max\{(X_t-X_s)'(X_t-X_s)\}\ , \qquad t<s\ . \tag{4.20}$$

Here, we note that there is no guarantee of uniqueness of $D_{\max}^2$ or Δ_{l+1}^2 for any value of l and that the procedure is sensitive to the metrices for the various factors because it is based on the calculation of distances. Thus, in the absence of any criterion for the choice of metrices, we may at first standardize the matrix of X such that $X'X$ is the form of correlation matrix and then carry out the procedure. Besides the standardization, it is suggested that in stead of (4.17) we use the distance

$$d_{ts}^2=\sum_{j=1}^{p_0}\left\{\frac{X_{tj}-X_{sj}}{b_j-a_j}\right\}^2, \tag{4.21}$$

where a_j and b_j is the lower and the upper bound for the j-th regressor, respectively.

Thus, by use of the CADEX procedure we can construct various sampling structures of any size n for the matrix of p_0-regressors after the original matrix $X_{(N)}$ has been given for the sufficiently large size N. Then the correlation structure of X may be guaranteed to be same for any size on this CADEX procedure except for the sampling fluctuation. In other words, we can generate the index samples and the follow-up samples of regressors matrix without changing the correlation structure.

4.5 Some results of our phase-two study

A simulation experiment was performed by following the way of design and the specification described in previous section. As the 8×3 regressors matrix X in (2.9) was adopted. The performance of $\hat{\beta}^*$ was evaluated on $\overline{\text{TMSE}}$ for four sources of variation, namely, methods of choosing the optimal k (Factor A), directions of the vector β (Factor B), magnitudes of error deviation σ (Factor C), and degrees of correlation among regressors (Factor D). Levels of each factor are as follows:

Table 2. Analysis of variance on $\overline{\text{TMSE}}$: Result of the phase two study.

Source	S.S.	d.f.	M.S.	F
Method (M)	504.9382	8.	63.1173	1280.270
Direction (ϕ)	7.7776	2.	3.8888	78.880
Error deviant (σ)	1450.5040	4.	362.6257	7355.491
Correlation (α)	54.8086	2.	27.4043	555.866
Interaction ($M\times\phi$)	2.4689	16.	0.1543	3.130
($M\times\sigma$)	1280.1770	32.	40.0055	811.471
($M\times\alpha$)	538.1975	16.	33.6373	682.298
($\phi\times\sigma$)	2.3608	8.	0.2951	5.986
($\phi\times\alpha$)	5.8981	4.	1.4745	29.909
($\sigma\times\alpha$)	91.8654	8.	11.4832	232.925
($M\times\phi\times\sigma$)	4.1257	64.	0.0645	1.308
($M\times\phi\times\alpha$)	3.5908	32.	0.1122	2.276
($M\times\sigma\times\alpha$)	1175.8690	64.	18.3730	372.677
($\phi\times\sigma\times\alpha$)	1.4504	16.	0.0907	1.840
Residual	6.3086	128.	0.0493	
Total	5130.3360	404.	12.6989	

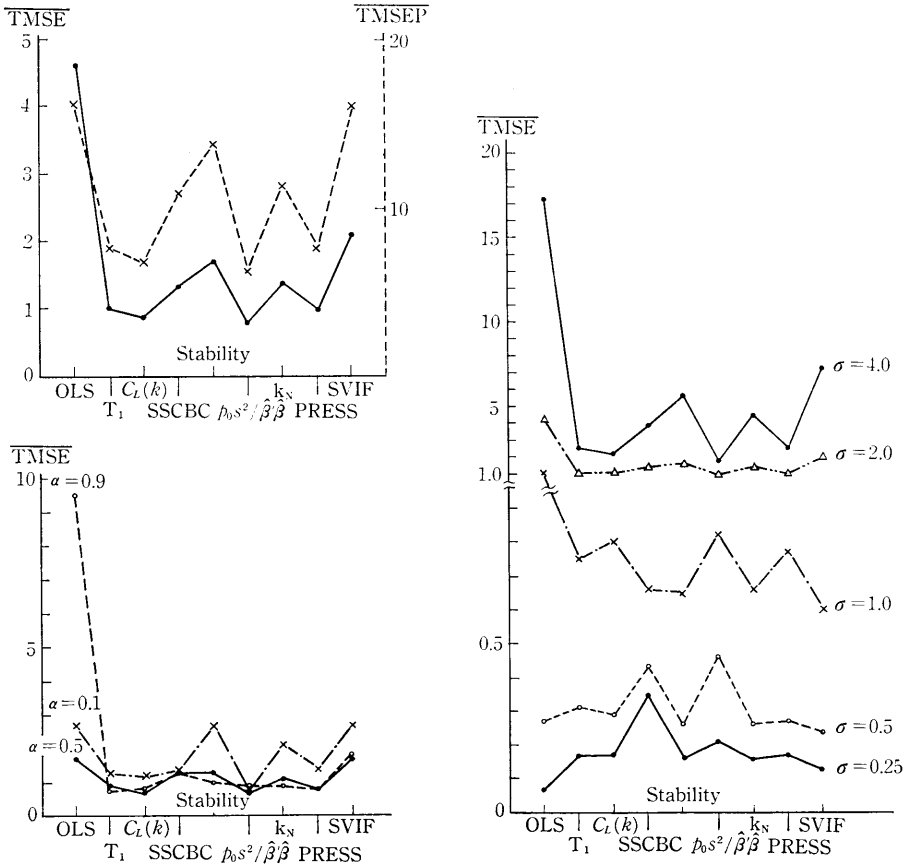


Fig. 5. The estimates of main effect (M) on $\overline{\text{TMSE}}$ and $\overline{\text{TMSEP}}$, and of interaction effects of $M\times\alpha$ and $M\times\sigma$ on $\overline{\text{TMSE}}$.

$A(M)$: OLS (ordinary least square), min $T_1(k)$, min $C_L(k)$, SSCBC, stability condition, $k=p_0s^2/\hat{\beta}'\hat{\beta}$, k_N, PRESS and SVIF.

$B(\phi)$: 0.0, 0.5, and 1.0

$C(\sigma)$: 0.25, 0.5, 1.0, 2.0, and 4.0

$D(\alpha)$: 0.1, 0.5, and 0.9

Table 2 shows the result of analysis of variance on $\overline{\text{TMSE}}$ for the four-way design. All main effects are highly significant and the effct of σ is the largest. We note that the effect of ϕ is small relative to other sources of variation. Figure 5 shows the estimates of the main and interaction effects associated with the method (M) in which we have interested on $\overline{\text{TMSE}}$. The rules of min $T_1(k)$, $C_L(k)$ and $k=p_0s^2/\hat{\beta}\hat{\beta}$

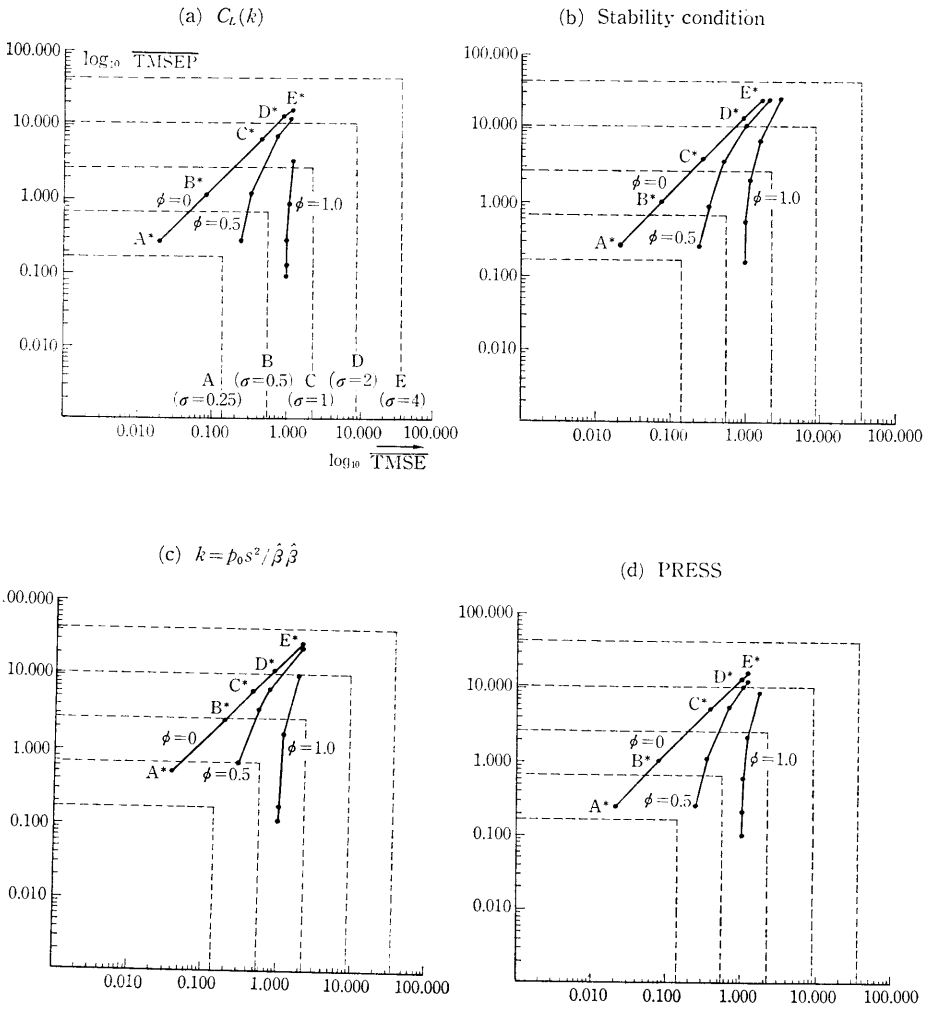


Fig. 6. Trajectory of point ($\log_{10}\overline{\text{TMSE}}$, $\log_{10}\overline{\text{TMSEP}}$) for four methods in the case of $\alpha=0.9$: Real line shows the trace of the posht for $\hat{\beta}^*$ associated with change of ϕ and σ, and dotted line shows the trace for $\hat{\beta}$ associated with change of σ.

give the good performance on $\overline{\text{TMSE}}$. When α is 0.9, $\hat{\beta}$ is improved by the ridge adjustment of any method on $\overline{\text{TMSE}}$. However, as k approches to zero, the stability condition, k_N and SVIF could give larger $\overline{\text{TMSE}}$ than $\hat{\beta}$. Also every method gives the improvement of $\hat{\beta}$ on $\overline{\text{TMSE}}$ when σ exceeds unity. When σ is extremely small, then the least square estimate will be close to true parameter β even if near-multicollinearity occurs among regressors. Thus, the result of $M\times\sigma$ will be natural for our expectation.

Similar analysis of variance for the same sources of variation was carried out on $\overline{\text{TMSEP}}$. Here, the results were omitted to save space, except for main effect (M) which has been shown in Figure 5. We should note that the rule $k=p_0s^2/\hat{\beta}'\hat{\beta}$ which is TMSE oriented criterion also gives the minimum of $\overline{\text{TMSEP}}$. In this stage any rule of thumb does not give better performance among the nine methods. From consideration of the results of the phase-two study, we pick-up $C_L(k)$, stability condition, $k=p_0s^2/\hat{\beta}'\hat{\beta}$, and PRESS as the methods which we should investigate in the following phase-three study.

Figure 6 shows the joint performance on $\overline{\text{TMSE}}$ and $\overline{\text{TMSEP}}$ for these four methods. That is, the horizontal scale is $\log_{10}\overline{\text{TMSE}}$ and the vertical scale is $\log_{10}\overline{\text{TMSEP}}$. The real line from point A^* to E^* shows the trajectory for the ridge estimate in each case of ϕ with change of σ from 0.25 to 4. The dotted line from A to E shows the border-line for the ordinary least square estimate with change of σ from 0.25 to 4. If the point A^* were outside the range enclosed by the border-line associated with A, the corresponding ridge estimate had not given better performance than $\hat{\beta}$ either on $\overline{\text{TMSE}}$ or on $\overline{\text{TMSEP}}$. We note that for $\phi=1.0$, the ridge estimate improve $\hat{\beta}$ along $\overline{\text{TMSEP}}$ rather than along $\overline{\text{TMSE}}$. Even if σ were extremely small ($\sigma\leqq 0.25$), $\hat{\beta}$ would be improved by the ridge adjustment of all methods on $\overline{\text{TMSE}}$ but not on $\overline{\text{TMSEP}}$ when $\phi=0$.

4.6 Some results and summary of our phase-three study

We have performed the simulation experiments for four methods, namely, the stability condition, $k=p_0s^2/\hat{\beta}'\hat{\beta}$, $C_L(k)$ and PRESS which have been selected based on the interpretation of the results of previous phase-two study as the typical one among

Table 3. Analysis of variance on $\overline{\text{TMSE}}$: Result of simulation experiment on CADEX sampling.

Source	S.S.	d.f.	M.S.	F
Method (M)	101.14	4.	25.29	30.72
Direction (ϕ)	91.77	2.	45.88	55.75
Error deviant (σ)	967.50	4.	241.87	293.88
Interaction ($M\times\phi$)	49.32	8.	6.17	7.49
($M\times\sigma$)	520.90	16.	32.56	39.56
($\phi\times\sigma$)	17.70	8.	2.21	2.69
($M\times\phi\times\sigma$)	18.60	32.	0.58	0.71
Residuals	3024.72	3675.	0.82	
Total	4791.66	3749.		

four class of criteria in section 3. In the study, "CADEX sampling" has been applied to generation of regressors matrix for the index sample X_I and for the follow-up sample X_F.

First, 50000×5 matrix with the correlation structure (4.15) was generated as the original regressors matrix $X_{(N)}$ by use of ordinary "multivariate normal random generator method" provided by Box and Müller (1958). Next, 50 replication samples

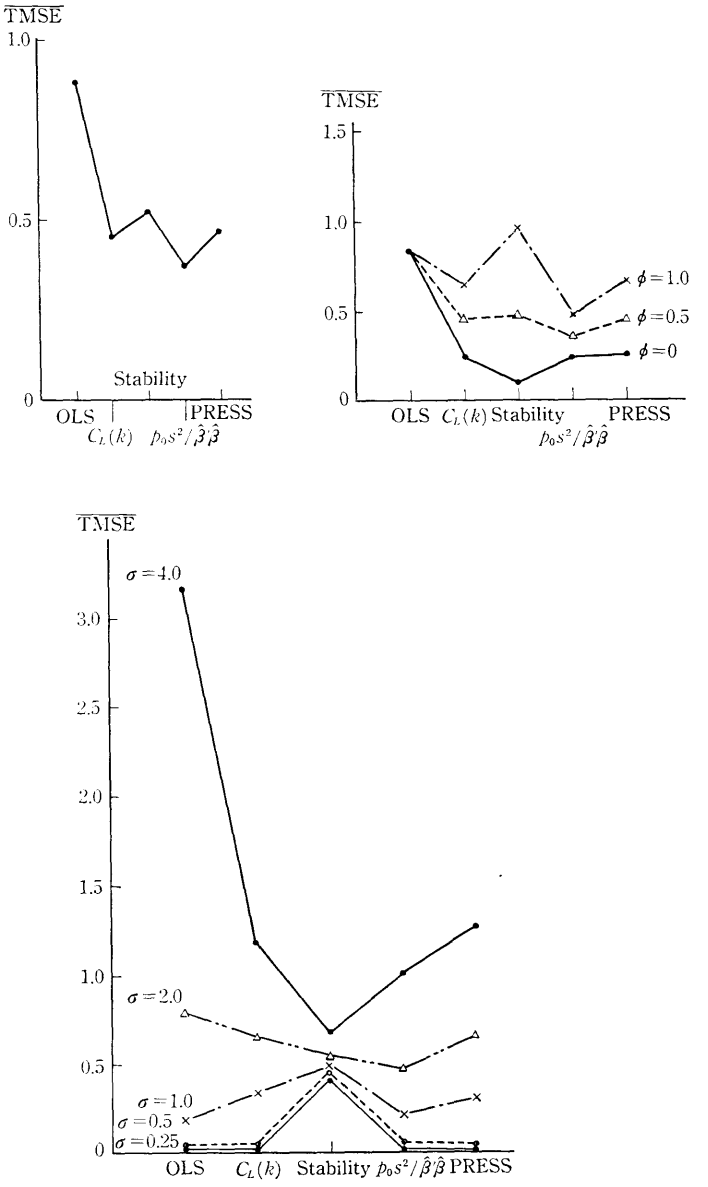


Fig. 7. The estimates of main effect (M), and of interaction effects ($M\times\phi$, $M\times\sigma$) on $\overline{\text{TMSE}}$.

of 50×5 matrix, X_I were constructed with similar correlation structures as (4.15) from 50000×5 matrix $X_{(N)}$ by use of the CADEX sampling. Further, 50 replication samples of set of matrices (X_I, $X_{F(10)}$, $X_{F(20)}$, $X_{F(30)}$, $X_{F(40)}$, $X_{F(50)}$) were sampled by use of CADEX, where $X_{F(J)}$ is $J\times5$ matrix for follow-up sample and is independent of X_I.

And three levels of ϕ were fixed at 0.0, 0.5, and 1.0 as in the phase-two study corresponding to the minimum eigenvalue vector of R in (4.15). Error variances also were fixed at five levels as in the phase-two study, and error terms were generated from normal deviates with mean zero and the error deviations. Finally, 750 simulated models of our phase-three study were constructed and the ridge estimates were calculated and evaluated on these models.

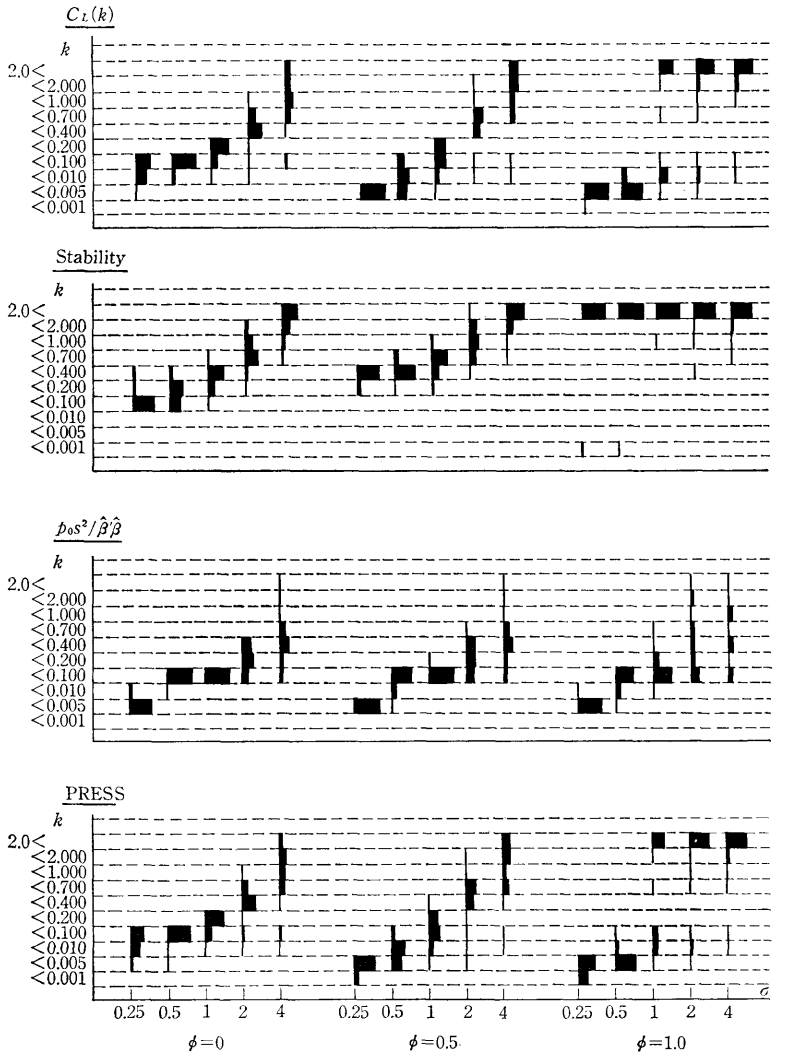


Fig. 8. Sampling distribution of "optimal" k given by $C_L(k)$, Stability condition, $k=p_0s^2/\hat{\beta}'\hat{\beta}$ and PRESS.

(1) Performances of $\hat{\beta}^*$ on TMSE.

Table 3 shows the result of analysis of variance on $\overline{\text{TMSE}}$ for three sources of variation, namely, methods, ϕ_s', and σ_s'. Obviously, all main effects were highly significant and particularly, the effect of σ was the largest. Since we are interested in the main effect due to method, interation effects between method and ϕ, and between method and σ, we have yielded these estimated effects in Figure 7. It will be natural that the rule $k=p_0 s^2/\hat{\beta}'\hat{\beta}$ gives the minimum value of $\overline{\text{TMSE}}$ among four methods since it is only one "minimum TMSE oriented criterion". Except for the stability condition, other three methods give smaller $\overline{\text{TMSE}}$ of $\hat{\beta}^*$ relative to $\hat{\beta}$. Though the stability condition does not give the stable behaviour on $\overline{\text{TMSE}}$ when ϕ changes from zero to unity, it gives the stable narrow range of $\overline{\text{TMSE}}$ for changes of σ. Also the stability condition takes the minimum value of $\overline{\text{TMSE}}$ among four methods when ϕ is zero, namely β is orthgonal to the eigenvector g_{p_0} associated with the minimum eigenvalue λ_{p_0} of $X'X$.

Unfortunately, the performance on $\overline{\text{TMSE}}$ is not good for any method when σ does not exceed unity. Especially, the stability condition gives the worst performance on $\overline{\text{TMSE}}$ for σ no more than unity. However, if the value of σ were extremely small, then ridge adjustment to $\hat{\beta}$ may be not necessary since the value of F_∞ in (2.4) may be negative and so the solution k_s may be very small, that is, the admissible range of $\hat{\beta}^*$ will be extremely narrow. In practice, since it is expected that the value of σ may be moderately large for ill-conditioned data the ridge adjustment to $\hat{\beta}$ will be effective.

Figure 8 shows the sampling distribution of "optimal" k for four methods on 50 replication samples. As ϕ increases from zero to unity and σ increases from 0.25 to 4, the optimal values of k tend to have large ones. Especially, the stability condition tends to give larger values of k than other methods. And the rule $k=p_0 s^2/\hat{\beta}'\hat{\beta}$ gives the conservative (small) values of k under various conditions of ϕ and σ. Thus, the rule will give $\hat{\beta}^*$ close to $\hat{\beta}$, and best $\overline{\text{TMSE}}$ performance.

(2) The predictive performance of $\hat{\beta}^*$ within index samples.

Table 4 shows the result of analysis of variance on $\overline{\text{TMSEP}}$ for the same sources of variation as in Table 3, where $\overline{\text{TMSEP}}$ has been calculated at $X_0=X_I$ in (1.10a). That is, the index sample on which $\hat{\beta}^*$ has been estimated was also used for evalua-

Table 4. Analysis of variance on $\overline{\text{TMSEP}}$ within index samples: Result of simulation experiment on CADEX sampling.

Source	S.S.	d.f.	M.S.	F
Method (M)	35172.41	4.	8793.10	26.82
Direction (ϕ)	16307.55	2.	8153.77	24.87
Error deviant (σ)	1214380.38	4.	303595.10	926.09
Interaction ($M\times\phi$)	8963.30	8.	1120.41	3.42
($M\times\sigma$)	122643.15	16.	7665.20	23.38
($\phi\times\sigma$)	52184.83	8.	6523.10	19.90
($M\times\phi\times\sigma$)	30140.43	32.	941.89	2.87
Residuals	1204752.60	3675.	327.82	
Total	2684544.65	3749.		

tion of its predictive performance. Similarly to the result in Table 3, the effect of σ was highly significant. Figure 9 shows the profile of each method on $\overline{\text{TMSEP}}$ for main effect of method, interaction $M\times\phi$ and $M\times\sigma$ as in Figure 7. The rule $C_L(k)$ and PRESS gave the minimum value of $\overline{\text{TMSEP}}$ as we have expected the result as one of "prediction oriented criterion". Further, for any method, $\overline{\text{TMSEP}}$ of $\hat{\beta}$ was improved by ridge adjustment for every ϕ and σ which is no less than unity. Particularly we note that the variability of $\overline{\text{TMSEP}}$ are smallest in the rule $k=p_0s^2/\hat{\beta}'\hat{\beta}$

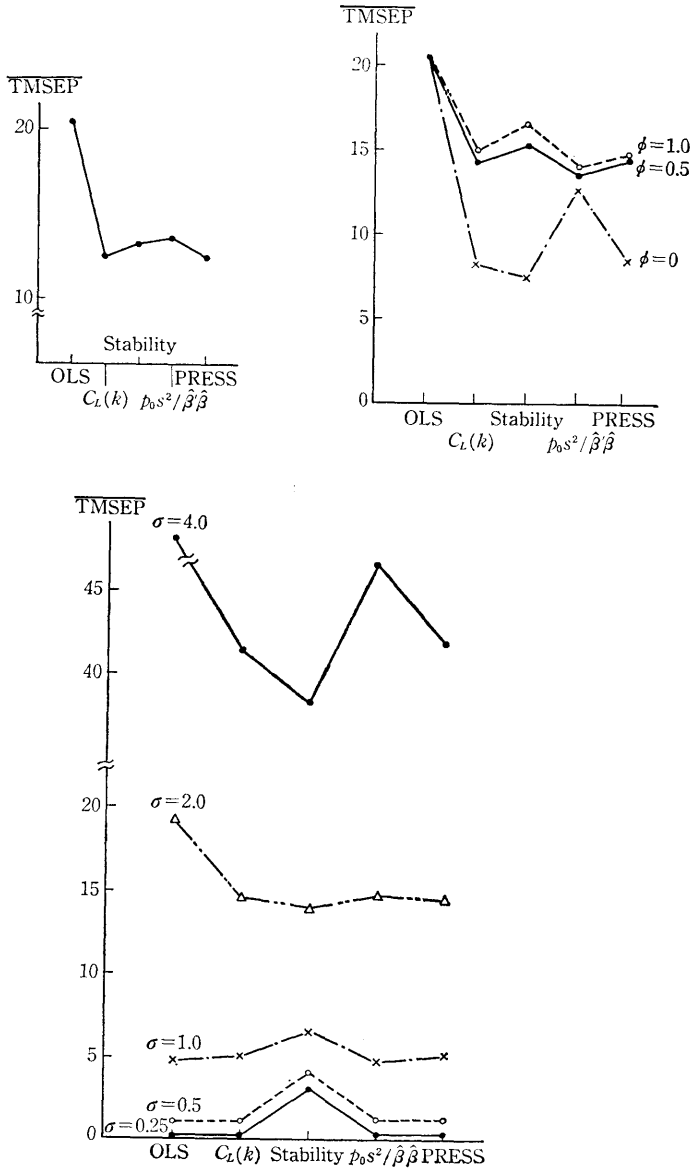


Fig. 9. The estimates of main effect (M), and of interaction effect ($M\times\phi$, $M\times\sigma$) on $\overline{\text{TMSEP}}$.

for the change of ϕ. Also, the improvement of $\overline{\text{TMSEP}}$ of $\hat{\beta}$ becomes small as σ approances large value, and the predictive performance of $\hat{\beta}^*$ is worse than of $\hat{\beta}$ when σ is no more than unity.

(3) The predictive performance of $\hat{\beta}^*$ for follow-up samples.

The values of $\overline{\text{TMSEP}}$ were calculated for $\hat{Y}=X_0\hat{\beta}^*$ at $X_0=X_{F(J)}$, $J=10$, 20, 30, 40 and 50 by use of $\hat{\beta}^*$ which has been estimated on X_I. Table 5 shows number of times such that $\overline{\text{TMSEP}}(\hat{\beta}^*)<\overline{\text{TMSEP}}(\hat{\beta})$ for four methods under the condition imposed on σ, ϕ and J. When σ is relatively small, the stability condition results in the worst predictive performance. For the other three methods, this trend is not ex-

Table 5. Number of times such that $\overline{\text{TMSEP}}$ $(\hat{\beta}^*) < \overline{\text{TMSEP}}$ $(\hat{\beta})$: The predictive performance of $\hat{\beta}^*$ outside index samples.

method	J	10			20			30			40			50		
	ϕ	0.0	0.5	1.0	0.0	0.5	1.0	0.0	0.5	1.0	0.0	0.5	1.0	0.0	0.5	1.0
	σ															
$C_L(k)$	0.25	24	22	24	27	22	21	24	23	24	25	23	23	25	25	25
	0.5	25	20	23	29	22	21	30	21	23	30	22	22	32	23	25
	1.0	31	23	31	36	24	23	36	20	20	37	24	21	37	23	22
	2.0	36	33	47	40	35	44	40	35	44	39	35	40	41	33	40
	4.0	38	40	49	40	40	50	40	40	49	41	42	50	43	42	50
Stability	0.25	9	0	4	7	0	4	6	0	4	10	0	4	8	0	4
	0.5	16	0	4	21	0	1	18	0	1	22	0	1	23	0	1
	1.0	22	9	22	29	12	15	27	8	7	30	9	6	31	11	4
	2.0	24	23	47	34	28	44	32	24	44	36	27	40	37	28	40
	4.0	28	31	49	34	38	50	35	39	49	38	40	50	37	37	50
$p_0s^2/\hat{\beta}'\hat{\beta}$	0.25	27	21	21	28	21	19	28	19	19	27	21	21	26	22	20
	0.5	31	17	19	34	20	17	34	19	16	35	20	17	34	21	18
	1.0	38	23	19	38	23	17	41	19	16	41	21	16	42	22	14
	2.0	42	32	37	43	39	37	42	37	34	43	37	33	43	37	36
	4.0	44	44	47	45	44	47	45	44	48	45	44	48	45	44	47
PRESS	0.25	25	26	33	27	27	32	26	27	35	25	26	33	25	30	34
	0.5	28	21	23	29	20	19	32	21	23	31	21	22	33	23	25
	1.0	32	24	30	36	24	21	36	20	20	38	23	21	37	22	22
	2.0	36	33	47	40	34	44	40	34	44	41	35	40	42	33	40
	4.0	41	40	49	41	41	50	42	40	49	41	42	50	43	43	50

treme, but the improvement of $\hat{\beta}$ becomes more larger as σ and ϕ are greater. For every methods, the trend of improvement of $\hat{\beta}$ was almost invariant for change of sample size J.

5. Concluding remarks

In this article we have evaluated the performance of the ordinary ridge regression on TMSE and TMSEP through consideration of the possibility to construct the finite admissible range of $\hat{\beta}^*$ and some systematic simulation studies. Consequently, we could sum-up following points as our conclusion. That is, (1°) the ridge regression is a data-adaptive method which may be applied for ill-conditioned data, and the adjustment of $\hat{\beta}$ is determined by a biasing parameter k in accordance with the degree of near-multicollinearity in the data, (2°) the admissible range of $\hat{\beta}^*$ on TMSE is wide for ill-conditioned data, and so the ridge regression could be said to be one method such that "it forces the user to recognize that a wide range of point estimates of β can be viable alternatives to $\hat{\beta}$," (3°) the predictive performance of $\hat{\beta}^*$ is better relative to $\hat{\beta}$ for ill-conditioned or large-variable data when the purpose of the usage of regression is in the prediction of future resopnse, (4°) $k=p_0 s^2/\hat{\beta}'\hat{\beta}$ among nine methods described above is recommended as the optimal value from the view point of both (i) prediction oriented use ($\overline{\text{TMSEP}}$ performance) and (ii) control oriented use ($\overline{\text{TMSE}}$ performance) of regression, and the values s^2 and $\hat{\beta}'\hat{\beta}$ in the k are also useful to judge the possibility to construct the finite admissible range of $\hat{\beta}^*$ (see section 2), (5°) the ridge trace on m-scale is a useful method to reflect the profile of relevant ill-conditioned data on the trace and to be compare it with the profile of other data which has been intended to use regression.

Further, since a computer simulation is indeed an experiment, we should give careful attention to experimental design prior to conducting a simulation experment. In the paper, we have particularly noted at the design and phase of simulation. Then we should not neglect the formulation of computer programs and the validation of model and data which could not be mentioned in the paper, and particulary forget the lobours, costs and times for the simulations.

Also, we should note that the standardization of X constitutes the usual instruction to ridge regression users that X be in "correlation form". Of course the ridge regression is invariant on centering of X but not invariant on scaling of it. If we were all nerves to the question of scaling of X, then the following adjustment of k may be useful. That is, by use of a $p_0 \times p_0$ transformation matrix

$$T=\text{diag}\,(1/s_1,\ 1/s_2,\ \cdots 1/s_{p_0}),$$

the ridge estimate for the transformed regressors matrix $Z=XT$ could be obtained as

$$\hat{\beta}_A^*=(Z'Z+k_A I_{p_0})^{-1}Z'Y, \tag{5.1}$$

where s_i^2 is the variance of the i-th regressor variable. $\hat{\beta}_A^*$ can be rewritten as

$$\hat{\beta}_A^*=T^{-1}[X'X+k_A(TT')^{-1}]^{-1}X'Y. \tag{5.2}$$

Thus, we obtain the general form of the ridge estimate $\hat{\beta}_G^*$ in section 3(4) as

$$\hat{\beta}_G^* = T\hat{\beta}_A^* = [X'X + k_A(TT')^{-1}]^{-1}X'Y, \tag{5.3}$$

where from setting at $K = k_A(TT')^{-1}$ we have

$$k_i = k_A/s_i^2, \qquad i = 1, 2, \cdots, p_0. \tag{5.4}$$

6. Acknowledgements

The authors are indebted first and foremost to Professor M. Okamoto (Osaka University) who gave them the valuable suggestions and guidances, and also Professor T. Nagai (Oita University) who gave them helpful comments and discussions and supported them in many ways. They are deeply grateful to Professor C. Asano (Kyushu University) for his advices in the preparation of the final version of a program covering most of the methods in the paper.

References

[1] ALLEN, D.N., *The relationships between variable selection and data augumentation and a method for prediction.*, Technometrics, **16** (1974), 125-127.
[2] BARTLETT, M.S., *Tests of significance in factor analysis*, British Journal of Psychology (Statistical Section), **3** (1950), 77-85.
[3] BOX, G.E.P. and MÜLLER, M.E., *A note on the generation of random normal deviates*, Ann. Math. Statist., **29** (1958), 610-611.
[4] FAREBROTHER, R.W., *The minimum mean square error of linear estimator and ridge regression*, Technometrics, **17** (1975), 127-128.
[5] FARRAR, D.E. and GLAUBER, R.R., *Multicollinearity in regression: the problem revisited*, Review of Economics and Statistics, **49** (1967), 92-107.
[6] GORMAN, W.J. and TOMAN, R.J., *Selection of variables for fitting equation to data*, Technometrics, **8** (1966), 27-51.
[7] GOTÔ, M., *Effects of correlation on multiple linear regression analysis (in Japanese)*. Standardization and Quality Control, **25** (1972), 17-23.
[8] GOTÔ, M., *Adjustment of regression coefficients in regression analysis: some consideration on ridge method* (in Japanese). J. Jap. Statist. Assoc., **6** (1976), 39-61.
[9] HALD, A., Statistical Theory and Engineering Applications. John Wiley and Sons (1952).
[10] HOERL, A.E., *Applications of ridge analysis to regression problems*, Chem. Eng. Prog., **58** (1962), 54-59.
[11] HOERL, A.E. and KENNARD, R.W., *Ridge regression: biased estimation for non-orthogonal problems*, Technometrics, **12** (1970a), 55-67.
[12] HOERL, A.E. and KENNARD, R.W., *Ridge regression: application to non-orthogonal problems*, Technometrics, **12** (1970b), 69-82.
[13] HOERL, A.E. et al., *Ridge regression: some simulations*, Comm. Statist., **4** (1975), 105-123.
[14] HOERL, A.E. and KENNARD, R.W., *Ridge regression: Iterative estimation of the biasing parameter*, Comm. Statist.-Theor. Math., **A5**(1) (1976), 77-88.
[15] HOCKING, R.R., *The analysis and selection of variables in linear regression*, Biometrics, **32** (1976), 1-49.
[16] KENNARD, R.W. and STONE, L.A., *Computer aided design of experiments*, Technometrics, **11** (1969), 137-148.
[17] MALLOWS, C.L., *Some comments on C_p*, Technometrics, **15** (1973), 661-675.
[18] MARQUARDT, D.W., *Generalized inverse, ridge regression, biased linear estimation, and non-linear estimation*, Technometrics, **12** (1970), 591-612.

[19] McDONALD, G.C., *Discussion of "Ridge analysis following a preliminary test of the shrunken hypothesis"*, Technometrics, **17** (1975), 443-445.
[20] McDONALD, G.C. and SCHWING, R.C., *Instabilities of regression estimates relating air pollution to mortality*, Technometrics, **15** (1973), 463-481.
[21] McDONALD, G.C. and GALARNEAU, D.I., *A Monte-Carlo evaluation of some ridge type estimators*, J. Amer. Statist. Assoc., **70** (1975), 407-416.
[22] OBENCHAIN, R.L., *Ridge analysis following a preliminary test of the shrunken hypotheses*, Technometrics, **17** (1975), 431-441.
[23] OBENCHAIN, R.L., Classical F-tests and confidence region for ridge regression. Technometrics, **19** (1977), 429-439.
[24] OBENCHAIN, R.L., Methods of ridge regression, Presentation at Biometric Conference, Boston (1976).
[25] PRESS, S.J., Applied Multivariate Analysis, Holt Rinhart and Winston, Inc., (1972).
[26] WILKS, S.S., *Certain generalizations in the analysis of variance*, Biometrika, **24** (1932), 471-494.