Estimation of Ridge Regression Under the Integrated Mean Square Error Criterion

Yong B. Lim
Chi H. Choi
Sung H. Park

ABSTRACT

In response surface experiments, a polynomial regression model is often used to fit the respones surface by the method of least squares. However, if the vectors of predictor variables are multicollinear, least squares estimates of the regression parameters have a high probability of being unsatisfactory. Hoerl and Kennard have demonstrated that these undesirable effects of multicollinearity can be reduced by using "ridge" estimates in place of the least squares estimates. Ridge regression theory in literature has been mainly concerned with selection of k for the first order polynomial regression model and the precision of $\hat{\beta}(k)$, the ridge estimator of regression parameters.

The problem considered in this paper is that of selecting k of ridge regression for a given polynomial regression model with an arbitrary order. A criterion is proposed for selection of k in the context of integrated mean square error of fitted responses, and illustrated with an example. Also, a type of admissibility condition is established and proved for the propose criterion.

1. Introduction

This paper considers a problem arising in empirical investigations of the response relationship between a response variable (η) and m independent

*** This research was partially supported by the Korea Science and Engineering Foundation, Seoul, Korea, through 1979-1980. ***

variables($v_1, v_2,\ldots,v_m$) when the relationship is assumed tobe approximated by a polynomial, i.e.,

$$\eta=\alpha^0+\sum_{i=1}^{m}\alpha_i v_i+\sum_{i=1}^{m}\sum_{j=i}^{m}\alpha_{ij}v_i v_j$$

$$+\sum_{i=1}^{m}\sum_{j=i}^{m}\sum_{k=j}^{m}\alpha_{ijk}v_i v_j v_k+\ldots \qquad (1)$$

If the first two terms at the right hand side of (1) is used for regression, the polynomial model is called the first order model, and if the third term is added, it is called the second order model, and soon. When the experimenter decides the order of polynomial model to be used for his experiment, he often faces a large number of polynomial terms (v_i, v_iv_j, etc.) that often result in multicollinearity among the vectors of polynomial terms. Therefore, ridge regression approach for a polynomial in (1) is desirable.

Suppose the second order model is assumed for a response surface experiment. The model contains $p=2m+m(m-1)/2$ polynomial terms besides the intercept term. Since ridge regression theory deals with standardized variables, let the polynomial terms be standardized by the following way. Assume we have n data; $(v_{1k}, v_{2k},\ldots,v_{mk}; w_k)$, $k=1, 2, \ldots, n$, where v_{ik} is the k^{th} level of thei^{th} independent variable v_i and w_k is the k^{th} observation of η. (it is letten not to distinguish a random variable and is observed in this context)

Put $\quad x_i=(v_i-\bar{v}_i)/[\sum_k (v_{ik}-\bar{v}_i)^2]^{1/2}$,

$$y=(w-\bar{w})/[\sum_k (w_k-\bar{w})^2]^{1/2},$$

$$x_i^* x_j^* (v_iv_j-\overline{v_iv_j})/[\sum_k (v_{ik}v_{jk}-\overline{v_iv_j})^2]^{1/2},$$

where $i, j=1, 2, \ldots, m$ and $v_i=\sum_k v_{ik}/n$, $w=\sum_k w_k/n$, and $\overline{v_iv_j}=\sum_k v_{ik}v_{jk}/r$. Then the second order regression modelmay be written as

$$y(x)=\sum_{i=1}^{m}\beta_i x_i+\sum_{i=1}^{m}\sum_{i\geq i}^{m}\beta_{ij}x_i^* x_j^*+\epsilon=x'\beta+\epsilon \qquad (3)$$

where $x'=(x_1, x_2, \ldots x_m, (x_1^*)^2, (x_2^*)^2, (x_m^*)^2, x_{1*}x_2^*, x_2^*x_3^*, \ldots x^*_{m-1}x_m^*)$

is the standardized $(p\times 1)$ input polynomial terms and β is the corresponding $(p\times 1)$ vector of regression coefficients. Note that the model (3) does

not have the intercept term since $\sum_k y_k = 0$.

Suppose X is the $(n \times p)$ matrix of values taken by the polynomial terms in $\underset{\sim}{x}$ at each of the data points and Y is the $(n \times 1)$ vector of observed responses. Then the model (3) may be expressed in matrix notation,

$$Y = X\underset{\sim}{\beta} + \underset{\sim}{e} \tag{4}$$

where $\underset{\sim}{e}$ is an $(n \times 1)$ vector of random errors. We assume, as usual, that $E(\underset{\sim}{e}) = \underset{\sim}{0}$, and $E(\underset{\sim}{e}\underset{\sim}{e}') = \sigma^2 I_n$, where I_n denotes the $(n \times n)$ identity martrix. Since the polynomial terms in X are standardized, $X'X$ is in the form of a correlation matrix;and the vector of correlation coefficients of the response variable with each polynomial term.

The least squares estimate of $\hat{\underset{\sim}{\beta}}$ is given by $\hat{\underset{\sim}{\beta}} = (X'X)^{-1}X'Y$, and it is commonly used in practice. However, when the vectors of polynomial terms are multicollinear, $X'X$ matrix may have one or more small eigenvalues. If we let λ_i be eigenvalues of $X'X$ the expected squared distance between $\hat{\underset{\sim}{\beta}}$ and $\underset{\sim}{\beta}$,

$$E[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)] = e^2 \sum_{i=1}^{p} \lambda_i^{-1}, \tag{5}$$

will be likely to be large. Thus, $\hat{\beta}$ can be expected to be farther from $\underset{\sim}{\beta}$.

To overcome this difficulty, Hoerl and Kennard(1970) have suggested that the least squares estimator be replaced by the ridge estimators, $\hat{\underset{\sim}{\beta}}(k)$, where

$$\hat{\underset{\sim}{\beta}}(k) = (X'X + KI_p)^{-1}X'Y, \quad k > 0. \tag{6}$$

Equation (6) defines a class of estimators indexed by a scalar parameter $k > 0$. Note that $\hat{\beta}(0)$ is the least squares estimator and is denoted simply by $\hat{\beta}$. Hoerl and Kennard(1970) demonstrate that the ridge estimators with the right choice of k(fixed) have smaller total mean square error(TMSE), defined by

$$\text{TMSE}[\hat{\underset{\sim}{\beta}}(k)] = E[(\underset{\sim}{\beta}(k) - \hat{\underset{\sim}{\beta}})'(\underset{\sim}{\beta}(k) - \underset{\sim}{\beta})], \tag{7}$$

than the least squares estimarors. They also establish a type of admissibility condition, namely, there always exists a $k > 0$ such that

$$\text{TMSE}[\hat{\underset{\sim}{\beta}}(k)]<\text{TMSE}[\hat{\underset{\sim}{\beta}}(0)].$$

Unfortunatly, the optimal value of k cannot be determined with certainty because it depends on the unknown parameter vector $\underset{\sim}{\beta}$ and unknown error variance σ^2. In practice, k must be estimated from the data. Hoerl and Kennard(1970) originally suggested a graphical procedure. Since then, several "mechanical" rules for selecting k have been proposed and compared. They are Hoerl and Kennard(1970), Marguardt (1970), Hoerl, Kennard and Baldwin(1975), Guilkey and Murphy(1975), Marquardt and Snee(1975), McDonald and Galarneau(1975), Hocking, Speed and Lynn(1979), Wichern and Churchill(1978), etc. Note that these rules are basically in the context of TMSE.

For response surface experiments, in qeneral, the fitted equation $\hat{y}(\underset{\sim}{x})=\underset{\sim}{x}'\hat{\underset{\sim}{\beta}}(k)$ for the model (3) is intended to be used within some region of $(x_1, x_2,\ldots,x_m)$ space of interest to a researcher. If a criterion for selecting k of ridge estimation is to be developed based on the precision of $\hat{y}(\underset{\sim}{x})$, it seems appropriate to evaluate the performance of $\hat{y}(\underset{\sim}{x})$ over the region of interest. For the concept and use of the region of interest, see Helms(1974), Park (1977, 1978), and Gunst and Mason(1979).

In this paper we introduce the integrated mean square error criterion for selection of k and establish a type of admissibility condition under this criterion. A simple example will be illustrated to show how to select k from a given set of data.

2. Mean Square Error of Estimated Respones in Ridge Regression

Consider the linear model (4). Let Λ be a diagonal matrix of eigenvalues, λ_i, of $X'X$ and P be an orthogonal matrix of corresponding eigenvectors. Then we have

$$P'(X'X)P=\Lambda \text{ and } P'P=I_p.$$

If we write

$$Z=XP,\ \underset{\sim}{\beta}=P\underset{\sim}{\alpha} \tag{8}$$

then the linear model may be written as

$$\begin{aligned} Y&=X\underset{\sim}{\beta}+\underset{\sim}{e}\\ &=Z\underset{\sim}{\alpha}+\underset{\sim}{e}, \end{aligned} \tag{9}$$

where

$$Z'Z=\Lambda.$$

The least squares estimator of α is given by

$$\begin{aligned} \hat{\underset{\sim}{\alpha}}&=(Z'Z)^{-1}Z'Y\\ &=\Lambda^{-1}Z'Y. \end{aligned} \tag{10}$$

which provides the minimum sum of squares of residuals,

$$\begin{aligned} SSE(\hat{\underset{\sim}{\alpha}})&=(Y-Z\hat{\underset{\sim}{\alpha}})'(Y-Z\hat{\underset{\sim}{\alpha}})\\ &=Y'Y-\hat{\underset{\sim}{\alpha}}'Z'Y. \end{aligned} \tag{11}$$

The variance of $\hat{\underset{\sim}{\alpha}}$ is then given by

$$\begin{aligned} \mathrm{Var}(\hat{\underset{\sim}{\alpha}})&=\mathrm{Var}(\Lambda^{-1}Z'Y)\\ &=\mathrm{e}^2\Lambda^{-1} \end{aligned} \tag{12}$$

A ridge estimator of α is defined as

$$\begin{aligned} \hat{\underset{\sim}{\alpha}}(k)&=(Z'Z+KI_p)^{-1}Z'Y\\ &=(\Lambda+KI_p)^{-1}\Lambda\hat{\underset{\sim}{\alpha}} \end{aligned} \tag{13}$$

where k is a positive fixed value, and residual sum of squares of this estimator is as follows.

$$\begin{aligned} SSE[\hat{\underset{\sim}{\alpha}}(k)]&=[Y-Z\hat{\underset{\sim}{\alpha}}(k)]'[Y-Z\hat{\underset{\sim}{\alpha}}(k))\\ &=Y'Y-\hat{\underset{\sim}{\alpha}}(k)'Z'Y-k\hat{\underset{\sim}{\alpha}}(k)'\hat{\underset{\sim}{\alpha}}(k). \end{aligned} \tag{14}$$

The variance-covariance matrix and the bias vector of $\hat{\underset{\sim}{\alpha}}(k)$ are readily shown to be;

$$\begin{aligned} \mathrm{Var}[\hat{\underset{\sim}{\alpha}}(k)]&\mathrm{Var}[(\Lambda+KI_p)^{-1}Z'Y]\\ &=\sigma^2\Lambda(\Lambda+KI_p)^{-1} \end{aligned} \tag{15}$$

and

$$\begin{aligned} \mathrm{Bias}[\hat{\underset{\sim}{\alpha}}(k)]&=E[\hat{\underset{\sim}{\alpha}}(k)-\underset{\sim}{\alpha}]\\ &=[(\Lambda+KI_p)^{-1}\Lambda-I_p]\underset{\sim}{\alpha} \end{aligned}$$

$$=-k(\Lambda+KI_p)^{-1}\underset{\sim}{\alpha}. \tag{16}$$

To look at $\hat{\underset{\sim}{\alpha}}(k)$ from the point of view of mean square error for a ridge estimate of response at $\underset{\sim}{z}$, let

$$\hat{y}_k(\underset{\sim}{z})=\underset{\sim}{z}'\hat{\underset{\sim}{\alpha}}(k)$$

Using (15) and (16), we can easily find the variance andthe bias of $\hat{y}_k(\underset{\sim}{z})$:

$$\mathrm{Var}[\hat{y}_k(\underset{\sim}{z})]=\sigma^2\underset{\sim}{z}'\Lambda\underset{\sim}{z}(\Lambda+KI_p)^{-2}\underset{\sim}{z} \tag{17}$$

and

$$\mathrm{Bias}[\hat{y}_k(\underset{\sim}{z})]=\underset{\sim}{z}'[-k(\Lambda+KI_p)^{-1}\underset{\sim}{\alpha}]. \tag{18}$$

Therefore, the mean squareerror of $\hat{y}_k(\underset{\sim}{z})$ as an estimation of $z'\alpha$

$$\begin{aligned}MSE[\hat{y}_k(\underset{\sim}{z})]&=E[\hat{y}_k(\underset{\sim}{z})-\underset{\sim}{z}'\underset{\sim}{\alpha}]^2\\&=\mathrm{Var}[\hat{y}_k(\underset{\sim}{z})]+\{\mathrm{Bias}[\hat{y}_k(\underset{\sim}{z})]\}^2\\&=e^2\underset{\sim}{z}'\Lambda(\Lambda+KI_p)^{-2}\underset{\sim}{z}\\&\quad+\underset{\sim}{z}'[k^2(\Lambda+KI_p)^{-1}\underset{\sim}{\alpha}\underset{\sim}{\alpha}'(\Lambda+KI_p)^{-1}]\underset{\sim}{z}.\end{aligned} \tag{19}$$

Next, we describe three selection rules of k that appear in literature, so that we can compare these rules with the rule proposed later in this paper. Let

$$\hat{\sigma}^2=SSE(\hat{\underset{\sim}{\alpha}})/(n-p). \tag{20}$$

The first rult rule[Hoerl and Kennard(1970)] is then specified by

$$k=\hat{e}^2/\max_i[\hat{\alpha}_i]^2 \tag{21}$$

where $\hat{\alpha}_i$ is the i^{th}ccmponent of $\hat{\underset{\sim}{\alpha}}$ in (11). The rationale of this rule is that, if $\sigma^2$and $\underset{\sim}{\alpha}$ are knccwn, then the k-value obtained by (21) is sufficient to give ridge estimatcrs having smaller TMSE $[\underset{\sim}{\beta}(\hat{k})]$ than the least squares estimator.

The second rule[Hoerl, Kennard and Baldwin(1975)] proposes that the value of k be choses that the value of k be chosen by

$$k=p\hat{\sigma}^2/\hat{\underset{\sim}{\alpha}}'\hat{\underset{\sim}{\alpha}}. \tag{22}$$

They argue that if $X'X=I_p$, TMSE$[\hat{\underset{\sim}{\beta}}(k)]$ is minimized when $k=p\sigma^2/\underset{\sim}{\alpha}'\underset{\sim}{\alpha}$ and an estimate of this k is given by (22).

The third rule [McDonald and Galarneau(1975)] is specified as follows.

Find $Q=\hat{\alpha}'\hat{\alpha}-\hat{\sigma}^2{}_p\sum_{i=1}^{p}\lambda_i^{-1}$.

If $Q>0$, choose k such that $\hat{\alpha}(k)'\hat{\alpha}(k)=Q$. (23)

If $Q\leq 0$, choose $k=0$.

Note that Q is an unbiased estimator of $\underset{\sim}{\alpha}'\underset{\sim}{\alpha}=\underset{\sim}{\beta}'\underset{\sim}{\beta}$. The rationale of this rule is that, if the estimate of this rule is that, if the estimate of $\underset{\sim}{\alpha}'\underset{\sim}{\alpha}$ is positive, it is desirableto choose k such that $\hat{\underset{\sim}{\alpha}}(k)$has the same expected length as $\underset{\sim}{\alpha}$.

3. The Integrated Mse Criterion and Related Theorems

The integrated mean square error(IMSE) criterion is to determine the value of k which minimizesthe quantity

$$J(k)=\mathrm{IMSE}[\hat{y}_k(\underset{\sim}{z})]=\int_R \mathrm{MSE}[\hat{y}_k(\underset{\sim}{z})]\mathrm{dw}(\underset{\sim}{z}), \tag{24}$$

where R is the region of interest over which the linear model is to be used, and $W(z)$ is a weight function that can be treated as a probability distribution function on R. The $W(\underset{\sim}{z})$ allows for differential importance of $\hat{y}_k(\underset{\sim}{z})$ at different points in the region, and can even be specialized to a discrete set of points if desired.

Using the result in (19), we obtain that

$$J(k)=V(k)+B(k)$$

where $V(k)$ is the integrated variance of $\hat{y}_k(z)$ over R,

$$\begin{aligned} V(k)&=\int_R \mathrm{Var}[\hat{y}_k(\underset{\sim}{z})]\mathrm{dw}(\underset{\sim}{z}) \\ &=\sigma^2{}_R\underset{\sim}{z}'\Lambda(\Lambda+KI_p)^{-2}\underset{\sim}{z}\mathrm{dw}(z) \end{aligned} \tag{26}$$

and $B(k)$ is the integrated squared bias of $\hat{y}_k(\underset{\sim}{z})$ over R,

$$\begin{aligned} B(k)&=_R\{\mathrm{Bias}[\hat{y}_k(\underset{\sim}{z})]\}^2\mathrm{dw}(\underset{\sim}{z}) \\ &=k^2\int_R\underset{\sim}{z}'(\Lambda+KI_p)^{-1}\underset{\sim}{\alpha}\underset{\sim}{\alpha}'(\Lambda+KI_p)^{-1}\underset{\sim}{z}\mathrm{dw}(\underset{\sim}{z}) \end{aligned} \tag{27}$$

We want to prove the following theorems, so that it may be possible to find a $k>0$ whore relnced integrated variance woll more than concel of

the increased integrated squared bias, and thereby improve the integrated mean square error of $\hat{y}_k(z)$.

Theorem1. $V(k)$ is a continuous and strictly dereasing function of k.

Proof: Let $A_1(k)=\Lambda(\Lambda+KI_p)^{-2}$

$$=\begin{pmatrix} \frac{\lambda_1}{(\lambda_1+k)} & 0 & \cdots & 0 \\ 0 & \frac{\lambda_2}{(\lambda_2+k)^2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \frac{\lambda_p}{(\lambda_p+k)^2} \end{pmatrix}$$

Then $V(k)=\sigma^2{}_R z' A_1(k) z \mathrm{dw}(z)$. Since $\lambda_j>0$ for all i and $k>0$, each diagonal element is positive and there is no singularity in $A_1(k)$. Thus $V(k)$ is a continuous function. If we let $A_2(k)=dA_1(k)/dk$, $A_2(k)$ becomes

$$A_2(k)=-2\Lambda(\Lambda+KI_p)^{-3}$$

Since $\Lambda(\Lambda+KI_p)^3$ is positive definite, $A_2(k)$ is negative definite. Thus we have the following result;

$$dv(k)/dk=\sigma^2\int_R z' A_2(k) z \mathrm{dw}(z)<0.$$

Therefore, $V(k)$ is a continuous and strictly decreasing function.

Theorem 2. $B(k)$ is a continuous and monotonically increasing function of k.

Proof: Let $D_1(k)=k^2(\Lambda+KI_p)^{-1}\alpha\alpha'(\Lambda+KI_p)^{-1}$. Then $B(k)=\int_R z' D_1(k) z \mathrm{dw}(z)$, and by the same reason as Theorem 1, $B(k)$ is continuous. Let $D_2(k)=dD_2(k)/dk$ and $d_{ij}(k)$ be the (i,j) element of $D_1(k)$. Then

$$d_{ij}(k)=\frac{k^2\alpha_i\alpha_j}{(\lambda_i+k)(\lambda_j+k)}$$

and the derivative of $d_{ij}(k)$ withrespect to k is given by

$$d'_{ij}(k)=\frac{\alpha_i\alpha_j[k^2(\lambda_i+\lambda_j)+2k\lambda_i\lambda_j]}{(\lambda_i+k)^2(\lambda_j+k)^2.}$$

Therefore, it can be shown that

$$D_2(k)=k(\Lambda+KI_p)^{-2}[k(\Lambda\alpha\alpha'+\alpha\alpha'\Lambda)+2\Lambda\alpha\alpha'\Lambda](\Lambda+KI_p)^{-2}$$
$$=K(\Lambda+KI_p)^{-2}[\Lambda\alpha\alpha'(\Lambda+KI_p)+(\Lambda+KI_p)\alpha\alpha'^{-2}](\Lambda+KI_p)^{-2}$$

Now, we want to show that $D_2(K)$ is positive semi-definite. First, note that $D_2(K)$ can bə written as follows.

$$D_2(K)=K(\Lambda+KI_p)^{-\frac{3}{2}}\Lambda^{\frac{1}{2}}[D_3(K)+D_4(K)]\Lambda^{\frac{1}{2}}(\Lambda+KI_p)^{-\frac{3}{2}}$$

where

$$D_3(K)=(\Lambda+KI_p)^{-\frac{1}{2}}\Lambda^{\frac{1}{2}}\underset{\sim}{\alpha}\underset{\sim}{\alpha}'\Lambda^{-\frac{1}{2}}(\Lambda+KI_p)^{\frac{1}{2}}$$

and

$$D_4(k)=(\Lambda+KI_p)_1{}^2\Lambda^{-}{}_1{}^2\underset{\sim}{\alpha}\underset{\sim}{\alpha}'\Lambda_1{}^2(\Lambda+KI_p)^{-1/2}.$$

If $A=Q^{-1}BQ$ for a nonsingular matrix Q, A and B are said to be similar, and the eigenvalues of A are equal to those of B. Since $D_3(k)$ and $D_4(k)$ are similar to $\underset{\sim}{\alpha}\underset{\sim}{\alpha}$. $D^{\varepsilon}(k)$ and $D_4(k)$ have the same eigenvalue as $\underset{\sim}{\alpha}\underset{\sim}{\alpha}'$. Also, since for any column vector $\underset{\sim}{\alpha}$, $\underset{\sim}{\alpha}\underset{\sim}{\alpha}'$ is positive semi-definite, $D_3(k)$ are positive semi-definite. Hence, $D_2(k)$ is positive semi-definite and $dB(k)/d\,k$ is nonnegative. Therefore, $B(k)$ is a continuous and monotonically increasing function of k.

Theorem 3. There always exists a $k>0$ such that the integrated mean square error of a response ridege estimator is less than that of the least squares estimator.

Proof: First, note that the least squares estimator is $\alpha(0)$. To prove the theorem, it is only necessary to show that there always exists a small positive value of k such that $d[\mathrm{IMSE}(y_k(\underset{\sim}{z})]/dk<0$. From Theorems 1 and 2 above, we know that

$$\begin{aligned}d[J(k)]/dk&=dv(k)/dt+dB(k)dk\\&=-2\sigma^2\int_R\underset{\sim}{z}'\Lambda(\Lambda+KI_p)^{-3}\underset{\sim}{z}dW(\underset{\sim}{z})\\&+K\int_R\underset{\sim}{z}'(\Lambda+KI_p)^{-2}[K(\Lambda\underset{\sim}{\alpha}\underset{\sim}{\alpha}'+\alpha'\alpha\Lambda)\\&+2\Lambda\underset{\sim}{\alpha}\underset{\sim}{\alpha}'\Lambda](\Lambda+KI_p)^{-2}\underset{\sim}{z}\mathrm{dw}(z).\end{aligned}$$

But $dv(k)/dk$ approaches $-2\ \sigma^2\int_R\underset{\sim}{z}'\Lambda^{-2}\underset{\sim}{z}\mathrm{dw}(z)$ as k goes to $0+$, and $dB(k)/k$approaches 0 as k goes to $0+$. Hence,

$$\lim_{k\to 0+}\{d[(k)]/dk\}=-2\sigma^2\int_R\underset{\sim}{z}'\Lambda^{-2}\underset{\sim}{z}dw(z)<0.$$

4. A Selection Rule

In this section, we show how to obtain the value of k under the proposed

criterion. Since we do not know σ^2 and, we cannot use $J(k)=\mathrm{IMSE}[\hat{y}_k(\underset{\sim}{z})]$ directly. Instead, we may find unbiased estimates of σ^2 and $\underset{\sim}{\alpha}$, and then use them in place of the unknown parameters. From (12), we can obtain that

$$E\hat{\underset{\sim}{\alpha}}\hat{\underset{\sim}{z}}'=\underset{\sim}{\alpha}\underset{\sim}{\alpha}'+\sigma^2\Lambda^{-1} \tag{28}$$

Also, from general linear model theory we know that the expected value of $\hat{\sigma}^2$ in (20) is

$$E(\hat{\sigma}^2)=\sigma^2.$$

Next, consider the following matrix of region moments,

$$M(\underset{\sim}{z})=\int_R \underset{\sim}{z}\underset{\sim}{z}'dw(\underset{\sim}{z})$$

$$=\begin{pmatrix} m_{11} & m_{12} & \cdots & m_{22} \\ & m_{22} & \cdots & \\ & \vdots & & \vdots \\ & (\text{sym.}) & & m_{pp} \end{pmatrix} \tag{29}$$

Note from (8) that

$$\underset{\sim}{z}=p'\underset{\sim}{x} \tag{30}$$

where

$$\underset{\sim}{x}=(x_1,x_2,\ldots,x_m,(x_1^*)^2,(x_1^*x_2^*)\ldots)$$

as defined in (3).

If we substitute (30) into (29), and use that P is an orthogonal matrix, the following identity is established.

$$\begin{aligned} M(z)&=\int_R P'\underset{\sim}{x}\underset{\sim}{x}'Pdw(\underset{\sim}{x})\,|P'| \\ &=P'\left[\int_R \underset{\sim}{x}\underset{\sim}{x}'dw(\underset{\sim}{x})\right]P \\ &=P'M(\underset{\sim}{x})P, \end{aligned} \tag{31}$$

where R' is the region of interest in the $s=x(x_1,x_2,\ldots,x_m)$ space, and $W(x)$ is, in fact, a multivariate probability distribution function of $m$independent variables, x_i, $i=1,2,\ldots,m$, since the components $(x_i^*)^2$ and $x_i^*x_j^*$ in x are functions of $x_1,x_2,\ldots,x_n$.

The use of $M(\underset{\sim}{z})$ and trace operator gives us the following expression for $J(k)$.

$$J(k)=\sigma^2 T_r[\Lambda(\Lambda+KI_p)^{-2}M(\underset{\sim}{z})]$$
$$+T_r[k^2(\Lambda+KI_p)^{-1}\underset{\sim}{\alpha}\underset{\sim}{\alpha}'(\Lambda+KI_p)^{-1}M(\underset{\sim}{z})],$$

where T_r denotes trace. If we use the o.k. estimates $\hat{\sigma}^2$ and for σ^2 and $\underset{\sim}{\alpha}\underset{\sim}{\alpha}'$, we can obtain an o.k. estimate of $J(k)$,

$$\hat{J}(k)=\hat{\sigma}^2 T_r[\Lambda(\Lambda+KI_p)^{-2}M(\underset{\sim}{z})] \tag{33}$$
$$+T_r[k^2(\Lambda+KI_p)^{-1}(\hat{\underset{\sim}{\alpha}}\hat{\underset{\sim}{\alpha}}'-\hat{\sigma}^2\Lambda^{-1})(\Lambda+KI_p)^{-1}M(\underset{\sim}{z})]$$

Therefore, for a given set of data, the proposed selection rule is to choose the value of k that minimizes $J(k)$, which may be written as

$$J(k)=\hat{\sigma}^2\sum_{i=1}^{p}\frac{\lambda_i m_{ij}}{(\lambda_i+k)}^2+2k^2\sum_{i=1}^{p}\sum_{j=i+1}^{p}\frac{\hat{\alpha}_i\tilde{\alpha}_j m_{ij}}{(\lambda_i+k)(\lambda_j+k)}$$
$$+k^2\sum_{i=1}^{p}\frac{(\hat{\alpha}_i{}^2-\hat{\sigma}^2/\lambda_i)m_{ii}}{(\lambda_i+k)^2} \tag{34}$$

One of the simplest forms of $M(\underset{\sim}{x})$, which is often used in practice, occurs when R' is a symmetric region about the origin,

$$R'=\{(x_1,x_2,\ldots,x_m);-a\leq x_i\leq a,\ i=1,2,\ldots,m\} \tag{35}$$

where a is a positive constant. Since each x_i is a standardized variable, a natural choice of a is 1. Suppose every point $\underset{\sim}{x}_s$ on R' has an equal importance so that $W(\underset{\sim}{x})$is the uniform distribution on R' and ${}_R W(\underset{\sim}{x})d\underset{\sim}{x}_s=1$. Then

$$M(\underset{\sim}{x})=(2a)^{-m}\int_R \underset{\sim}{x}\underset{\sim}{x}'d\underset{\sim}{x}_s. \tag{36}$$

Another simple form of $M(\underset{\sim}{x})$ occurs when

$$W(\underset{\sim}{x})=1/\mathrm{n},\ \text{design points}$$
$$=0,\ \text{elsewhere.}$$

Then

$$M(\underset{\sim}{x})=\int_R \underset{\sim}{x}\underset{\sim}{x}'dW(\underset{\sim}{x})=X'X/\mathrm{n} \tag{37}$$

and

$$M(\underset{\sim}{z})=P'M(\underset{\sim}{x})P$$
$$=P'X'XP/\mathrm{n}$$
$$=\Lambda/\mathrm{n} \tag{38}$$

For this case, the $J(k)$ in (24) is, since $\hat{y}_k(\hat{\underset{\sim}{z}})=\underset{\sim}{z}'\hat{\underset{\sim}{\alpha}}(k)$,

$$J(k) = {}_R\underset{\sim}{z}'E[\hat{\underset{\sim}{\alpha}}(k)-\underset{\sim}{\alpha}][\hat{\underset{\sim}{\alpha}}(k)-\underset{\sim}{\alpha}]'z_c W(\underset{\sim}{z})$$
$$= Tr\{E[\hat{\underset{\sim}{\alpha}}(k)-\underset{\sim}{\alpha}][\hat{\underset{\sim}{\alpha}}(k)-\underset{\sim}{\alpha}]'M(\underset{\sim}{z})\}$$
$$= \frac{1}{n}E[\hat{\underset{\sim}{\alpha}}(k)-\underset{\sim}{\beta}]'\Lambda[\hat{\underset{\sim}{\alpha}}(k)-\underset{\sim}{\alpha}]$$
$$= \frac{1}{n}E(L_2{}^2), \qquad (39)$$

where $E(L_2{}^2)=E[\hat{\underset{\sim}{\alpha}}(k)-\underset{\sim}{\alpha}]'\Lambda[\hat{\underset{\sim}{\alpha}}(k)-\underset{\sim}{\alpha}]$ was used by Stein(1960), Hocking, Speed and Lynn (1976) and others as a "design dependent" criterion for estimation fo regression parameters. When we ure r/m for $M(z)$ in (33) we have

$$\hat{J}(k) = \frac{\hat{\sigma}^2}{n}\sum_{i=1}^{p}\frac{\lambda_i{}^2}{(\lambda_i+k)^2} + \frac{k^2}{n}\sum_{i=1}^{p}\frac{\lambda_i(\hat{\alpha}_i{}^2-\hat{\sigma}/\lambda_i)}{(\lambda_i+k)^2}$$

5. An example

In this section, an example is shown to obtain the value of k for a polynomial regression model under the proposed IMSE criterion. An experiment was conducted to determine the second order polynomial regression which relates the amount (parts per million, ppm) of water soluble in the soil(w) with the concentration (weight percent, wt %) of clay (v_1) and the soil P.H. (λ_2). Twenty observations were taken and readings recorded as shown in Table 1. These data were presented and described by Myers(1971).

$$X'X = \begin{pmatrix} 1 & -0.27694 & 0.99948 & -0.26696 & 0.93056 \\ & 1 & -0.29992 & 0.99966 & 0.09303 \\ & & 1 & -0.28992 & 0.92144 \\ \text{(sym.)} & & & 1 & 0.10230 \\ & & & & 1 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 2.245\times10^{-4} & 0 & 0 & 0 & 0 \\ & 1.94507 & 0 & 0 & 0 \\ & & 3.0538 & 0 & 0 \\ \text{(sym.)} & & & 8.6415\times10^{-50} & \\ & & & & 8.4143\times10^{-4} \end{pmatrix}$$

The variables are standardized by the scheme described in (2). so that the second order polynomial model is given by

Table 1. Example data

Run	w(ppm)	v_1(wt%)	v_2(ph)
1	0.62	37	5.3
2	0.69	37	5.3
3	0.63	37	5.5
4	0.61	37	5.7
5	0.28	29	5.6
6	0.33	29	5.7
7	0.31	29	5.7
8	0.37	29	5.9
9	0.66	29	6.0
10	0.70	29	6.3
11	0.74	29	6.0
12	0.63	29	6.0
13	0.52	27	5.5
14	0.47	27	5.6
15	0.45	27	5.6
16	0.42	27	5.7
17	0.41	27	5.5
18	0.42	27	5.5
19	0.42	27	5.5
20	0.41	27	5.5

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_{11}(x_1^*)^2+\beta_{22}(x_2^*)^2+\beta_{12}x_1^*x_2^*+\epsilon$$

where, for $i=1, 2$,

$$x_i=\frac{v_i-\overline{v_i}}{(S_i)^{1/2}}$$

$$x^{i*}x_j^*=\frac{v_iv_j-\overline{v_iv_j}}{(S_{ij})^{1/2}}$$

$$=\frac{(x_iS_i^{1/2}+\overline{v_i}(x_jS_j^{1/2}+\overline{v_j})-\overline{v_iv_j}}{(S_{ij})^{1/2}}$$

$$S_i=\sum_k(v_{ik}-\overline{v_i})^2$$

$$S_{ij}=\sum_k(v_{ik}v_{jk}-\overline{v_iv_j})^2.$$

We assume that ϵ is identically and independently distributed with mean zero and variance σ^2. For these data, we obtain the following results.

$$P=\begin{pmatrix} 0.75225 & 0.13681 & 0.56168 & -0.19060 & -0.25217 \\ -0.01267 & 0.63840 & -0.26048 & -0.68191 & 0.24380 \\ -0.64949 & 0.11989 & 0.56411 & -0.24515 & -0.43067 \\ 0.03590 & 0.64165 & -0.25525 & 0.51725 & -0.50429 \\ -0.10409 & 0.38424 & 0.48301 & 0.41356 & 0.66121 \end{pmatrix}$$

$$\hat{\underset{\sim}{\alpha}}=\begin{pmatrix} -18.6940 \\ 0.3863 \\ 0.2382 \\ -2.4236 \\ -14\ 7281 \end{pmatrix}$$

$\hat{\sigma}^2=0.018339.$

If we choose the weight lunction, $W(\underset{\sim}{x})$, to be uniform over the region of interest, $R'=\{(x_1,x_2):-1<x_i<1 \text{ for } i=1,2\}$, then the following region moments can be obtained, since

$$\underset{\sim}{x}'=(x_1,x_2(x_1^*)^2,\ (x_2^*)^2,\ x_1^*x_2^*),$$

$$M(\underset{\sim}{x})=\int_{-1}^{1}\int_{-1}^{1}\underset{\sim}{x}\underset{\sim}{x}'dx_1dx_2$$

$$=\begin{pmatrix} 0.33333 & 0 & 0.30846 & 0 & 0.35698 \\ & 0.33333 & 0 & 0.27673 & 0.12604 \\ & & 0.29667 & 0.0019976 & 0.33056 \\ \text{(sym.)} & & & 0.32369 & 0.12398 \\ & & & & 0.43490 \end{pmatrix}$$

$$M(\underset{\sim}{z})=P'M(\underset{\sim}{x})P$$

$$=\begin{pmatrix} 0.005345 & -0.001076 & 0.01897 & 0.001293 & 0.004090 \\ & 0.8137 & 0.1990 & -0.001641 & 0.01121 \\ & & 0.8947 & 0.01542 & 0.02955 \\ \text{(sym.)} & & & 0.001738 & 0.002360 \\ & & & & 0.006433 \end{pmatrix}$$

Using the Newton Raphson method, we find that the $\hat{J}(k)$ in (34) is minimized at $k=0.71965\times10^{-4}$. To see the difference of the choice of k for each proposed rule, these same data are applied to each rule; Hoerl and Kennard(1970), Hoerl, Kennard and Baldwin(1975) and McDonald and Galarueau(1975). The results are summarized in Table 2. The first five columns are the coefficient estimates of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$. and tqe cqoice of k is listed inthe next column. For the selected value of k, the $\hat{J}(k)$ and the residual sum of squares, $SSE()$) are evaluatedand are shown in the last two columns, For comparion purposes, the least squared estimates are included.

Table 2. Comparison of Ridge Estimators.

Rules	$\binom{\alpha_1}{\beta_1}$	$\binom{\alpha_2}{\beta_2}$	$\binom{\alpha_{11}}{\beta_{11}}$	$\binom{\alpha_{22}}{\beta_{22}}$	$\binom{\alpha_{12}}{\beta_{12}}$	k	$\hat{J}(k)$	$SSE(k)$
Least squares	−18.694 (−9.699)	0.3863 (−1.517)	0.2382 (19.26)	−2.424 (5.690)	−14.73 (−8.531)	0	0.9586	0.2751
IMSE	−14.16 (−6.789)	0.3863 (−2.0421)	0.2381 (15.54)	−1.322 (15.837)	−13.57 (−7.781)	0.71965 $\times 10^{-4}$	0.6463	0.2809
Hoerl & Kennard	−15.15 (−7.428)	0.3863 (−1.975)	0.2381 (16.36)	−1.508 (5.845)	−13.86 (−7.950)	0.53475 $\times 10^{-4}$	0.6597	0.2786
Hoerl, Kennard & Baldwin	−10.91 (−4.739)	0.3863 (−2.115)	0.2381 (12.80)	−0.8493 (5.596)	−12.37 (−7.133)	0.16018 $\times 10^{-3}$	0.7854	0.2936
McDonald & Galarneau	−10.41 (−4.428)	0.3863 (−2.107)	0.2381 (12.36)	−0.7897 (5.531)	−12.15 (−7.012)	0.1788 $\times 10^{-3}$	0.8326	0.2963

Table 2 shows that the range of k values is from 0.53475×10^{-4} to 0.16018×10^{-3}, and the IMSE rule gives the k value which is relatively small. The IMSE rule obviously minimizes $\hat{J}(k)$, and it also provides a small value of $SSE(k)$.

6. Conclusion

Ridge regression theory has been mainly concerned with selection of a single value of k such that total mean square error of $\underset{\sim}{B}(k)$ is minimized at k. But, generally for response surface experiments the fitted equation $y(\underset{\sim}{x})$ is intended to be used within-some region of $\underset{\sim}{x}$ space of interest to a researcher. In this paper, we have introduced integrated mean square error(IMSE) of $y(\underset{\sim}{x})$ obtained for polynomial ridge regression models. Thus, using IMSE criterion instead of total mean square error criterion, we have suggested a rule that selects an optimal value of k, and proved that such value of a positive k always exists. To facilitate the use of this rule, a simple example was illustrated.

REFERENCES

* GUILKEY, D.K. and MUPHY, J.L.(1975). Directed ridge regression techniques in cases of multicollinearity. J. Amer. Statist. Assoc., 70, 769-775.

* GUNST, R.F. and MASON, R.L. (1979). Some considerations in the evaluation of alternate prediction equations. Technometrics, 21, 55-64.

* HELMS. R.W. (1974). The average estimated variance criterion for the selection-of-variables problem in general linear models. Technometrics,16, 261-274.

* HOCKING, R.R., SPEED, F.M. and LYNN, M.J. (1976). A class of biased estimator in linear regression. Technometrics, 18, 425-438.

* HOERL, A.E. and KENNARD,R.W. (1970). Ridge regression:2space biased estimator for nonorthogonal problems. Technometrics, 12, 55-67.

* HOERL, A.E., KENNARD, R.W. and BALDWIN, K.F. (1975). Ridge regression: some simulations. Commun. Statist. Theor. Meth., 4, 105-123.

* MARQUARDT, D.W. (1970). Generalized inverses, ridge regression, biased linear estimation and nonorthogonal estimation. Technometrics, 12, 591-612.

* MARQUARDT, D.W. and SNEE, R.D. (1975). Ridge regression in practice. Amer. Statist., 29, 3-19.

* MCDONALD, G.C. and GALARNEAU, D.I.(1975). A Monte Carlo evaluation of some ridge type estimators. J. Amer. Statist. Assoc, 70, 407-416.

* Myers, R.H. (1971). Response Surface Methodology. Boston: Allyn and Bacon.

* PARK, S.H. (1977). Selection of polynomial terms for response surface experiments. Biometrics, 33, 225-229.

* PARK, S.H. (1978). Selecting contrasts among parameter in Scheffé's

mixture models: screening components and model redнction. Technometrics, 20, 273～279.

* STEIN. C.M. (1960) Multiple regression. Contributions to Probability and Statistics. Essays in honor of Horold Hotelling, ed. I. Olkin, Standford University Press, 424～443.

* WICHERN, D.W. and CHURCHILL, G.A. (1978). A comparison of ridge estimators. Technometrics, 20, 301～311.