응용통계연구 제 6권 2호
1993년 9월 pp. 371-381.

Robust Selection Rules of k in Ridge Regression[1)]

Yong B. Lim[2)]

Abstract

When the multicollinearity presents in the standard linear regression model, ridge regression might be used to mitigate the effects of collinearity. As the prediction-oriented criterion, the integrated mean square error criterion $J_w(k)$ was introduced by Lim, Choi & Park(1980). By noting the equivalent relationship between the C_k criterion and $J_w(k)$ with a special choice of weight function $w(x)$, we propose a more reasonable selection rule of k w.r.t. the C_k criterion than that given in Myers(1986). Next, to find the $\hat{\beta}(k)$ which behaves reasonably well w.r.t. competing criteria, we adopt the minimax principle in the sense of maximizing the worst relative efficiency of k among competing criteria.

1. Introduction

Consider the standard linear regression model

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

where Y is an $n\times 1$ vector of observations on the response (or dependent) variable, X is an $n\times p$ fixed matrix of observations on the explanatory (or independent) variables, β is a $p\times 1$ unknown vector of egression coefficients, and ε is an $n\times 1$ vector of random errors with means $E(\varepsilon)=0$ and $Var(\varepsilon)=\sigma^2 I_n$. Here the variables are assumed to be standardized so that $X'X$ is in the form of a correlation matrix and $X'Y$ is the vector of correlation coefficients of the response variable with independent variables. The least squares estimate of β is given by $\hat{\beta}=(X'X)^{-1}X'Y$, and commonly used in practice.

The multicollinearity of X occurs when there exists near linear dependency between columns of X, i.e., $\sum_{i=1}^{p} c_i x_i \simeq 0$, for some c_i, $X=[x_1,\cdots,x_p]$. As a result of the multicollinearity, $X'X$ may have one or more small eigenvalues. It is well known that a difficulty with least squares estimator arises as the multicollinearity occurs. In particular the expected squared distance between $\hat{\beta}$ and β,

$$E(\hat{\beta}-\beta)'(\hat{\beta}-\beta) = \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}, \quad \lambda_j : \text{ eigenvalues of } X'X \tag{1.2}$$

will be large. Thus $\hat{\beta}$ can be expected to be further from β. To overcome this difficulty,

1) Research supported by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1991
2) Department of Statistics, Ewha Womans University, Seoul, 120-750, Korea

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

$$\hat{\beta}(k) = (X'X + kI_p)^{-1}X'Y, \ k>0. \tag{1.3}$$

Note that $\hat{\beta}(0)$ is the least squares estimator.

Hoerl and Kennard(1970) demonstrate that the ridge estimator with the right choice of $k(fixed)$ have smaller total mean square error(TMSE), defined by

$$TMSE(k) = E[(\hat{\beta}(k) - \beta)'(\hat{\beta}(k) - \beta)], \tag{1.4}$$

than the least square estimators. They also establishes a type of admissibility condition, namely, there always exists a $k>0$ such that

$$TMSE(k) < TMSE(0).$$

For response surface experiments in general, the fitted equation $\hat{y}_k(x)=x'\hat{\beta}(k)$ is intended to be used within some region of explanatory variables of interest to a researcher. As the prediction-oriented criterion, Lim, Choi & Park(1980) have introduced the integrated mean squre error ($IMSE$) criterion for selecting k of ridge estimation and made parallel statements to TMSE criterion in Hoerl and Kennard(1970) with $J_w(k)$ defined by

$$J_w(k) = IMSE[\hat{y}_k] = \int_R MSE(\hat{y}_k(x))dw(x), \tag{1.5}$$

where $MSE(\hat{y}_k(x))$ is the mean square error of $\hat{y}_k(x)=x'\hat{\beta}(k)$ and $w(x)$ is a weight function on the region of explanatory variables of interest R. The $w(x)$allows for differential importance of $\hat{y}_k(x)$ at different points in R. When the prediction is equally important at each point in R, $w(x)$ could be taken as the uniform distribution on R. When $w(x)$ takes the uniform mass $1/n$ at each design point, $J_w(k)$ is proportional to the sum of MSE of $\hat{y}_k(x_i)$ at each design point. Thus, $J_w(k)$ is equivalent to the C_k criterion introduced by Myers(1986, p249).

Denote the $J_w(k)$ by $J_u(k)$ for the uniform weight function over R. When $w(x)$ takes the uniform mass $1/n$ at each design point, we denote $J_w(k)$ by $J_n(k)$.

Unfortunatly, the optimal choice of k w.r.t. those two criteria cannot be determined with certainty because it depends on the unknown parameters $\beta\beta'$ and σ^2. In practice, k must be selected from the data. The natural alternative is to replace $\beta\beta'$ and σ^2 by their unbiased estimates in those criteria and then, to find the optimal k. Lim, Choi & Park(1980) have suggested to select k which minimizes $\hat{J}_w(k)$, an unbiased estimate of $J_w(k)$. Similarly, we can select optimal k based on $\hat{TMSE}(k)$, an unbiased estimate of $TMSE(k)$.

The main results of the paper are as follows. First, by noting the equivalent relationship between C_k and $J_n(k)$ and using more reasonable estimates of $\beta\beta'$ rather than unbiased estimates, we propose a more reasonable selection rule of k w.r.t. the C_k criterion than that given in Myers(1986). Next, to find $\hat{\beta}(k)$ which behaves reasonably well w.r.t.

competing criteria, we adopt the minimax principle in the statistical decision theory. In this paper, we suggest a selection rule of k which are robust w.r.t. $J_u(k)$, $J_n(k)$ (or C_k) and $TMSE(k)$ in the sense of maximizing the worst relative efficiency of k among 3 criteria. Since the relative efficiency functions of k w.r.t. three objective functions are continuous respectively, there always exists such a robust k^*. An example will be illustrated to show how to select such a robust k^* from a given set of data for the quadratic polynomial regression model.

2. An unbiased estimate of TMSE and IMSE

Consider the linear model (1.1). Let Λ be a diagonal matrix of eigenvalues of $X'X$ and P be an orthogonal matrix of corresponding eigenvectors. Then $P'(X'X)P = \Lambda$. Letting $Z = XP$ and $\alpha = P'\beta$, the linear model (1.1) is reparametrized as follows;

$$\begin{aligned} Y &= XPP'\beta + \varepsilon \\ &= Z\alpha + \varepsilon, \end{aligned}$$

where $Z'Z = \Lambda$.

The least squares estimator of α is

$$\hat{\alpha} = (Z'Z)^{-1}Z'Y = \Lambda^{-1}Z'Y$$

and a ridge estimator of α is

$$\begin{aligned} \hat{\alpha}(k) &= (Z'Z + kI_p)^{-1}Z'Y \\ &= (\Lambda + kI_p)^{-1}\Lambda\hat{\alpha}, \qquad k>0. \end{aligned}$$

Thus, the residual sum of squares of $\hat{\alpha}(k)$ is

$$\begin{aligned} SSE(k) &= (Y - Z\hat{\alpha}(k))'(Y - Z\hat{\alpha}(k)) \\ &= Y'Y - \hat{\alpha}(k)'Z'Y - k\hat{\alpha}(k)'\hat{\alpha}(k) \end{aligned}$$

Since

$$\frac{dSSE(k)}{dk} = 2kY'Z(\Lambda + kI_p)^{-3}Z'Y > 0,$$

$SSE(k)$ is an increasing function of k.

Noting that $E[\hat{\alpha}] = \alpha$ and $Var(\hat{\alpha}) = \sigma^2\Lambda^{-1}$,

$$\begin{aligned} E[\hat{\alpha}(k)] &= (\Lambda + kI_p)^{-1}\Lambda E[\hat{\alpha}] \\ &= (\Lambda + kI_p)^{-1}\Lambda\alpha \end{aligned}$$

and

$$\begin{aligned} Var(\hat{\alpha}(k)) &= (\Lambda + kI_p)^{-1}(Var(\Lambda\hat{\alpha}))(\Lambda + kI_p)^{-1} \\ &= \sigma^2(\Lambda + kI_p)^{-1}\Lambda(\Lambda + kI_p)^{-1} \\ &= \sigma^2\Lambda(\Lambda + kI_p)^{-2}. \end{aligned}$$

Last equality follows from the fact that the diagonal matrices are commutative.

Now, from Hoerl and Kennard(1970),

$$TMSE(k) = TrE[(\hat{\alpha}(k)-\alpha)(\hat{\alpha}(k)-\alpha)'] = Tr(Var(\hat{\alpha}(k))) + Tr(E[\hat{\alpha}(k)]-\alpha)(E[\hat{\alpha}(k)]-\alpha)' \quad (2.1)$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2} + k^2 Tr(\Lambda+kI_p)^{-2}\alpha\alpha' .$$

Also, from Lim, Choi & Park(1980),

$$J_w(k) = \sigma^2 Tr[\Lambda(\Lambda+kI_p)^{-2}M] + k^2 Tr[(\Lambda+kI_p)^{-1}M(\Lambda+kI_p)^{-1}\alpha\alpha'], \quad (2.2)$$

where

$$M = \int_R zz' dw(z), \quad z = P'x.$$

Letting

$$s^2 = \frac{SSE(0)}{n-p-1}, \quad (2.3)$$

s^2 is an unbiased estimate of σ^2. From $Var(\hat{\alpha}) = E[\hat{\alpha}\hat{\alpha}'] - \alpha\alpha'$,

$$\alpha\alpha' = E[\hat{\alpha}\hat{\alpha}'] - \sigma^2\Lambda^{-1} = E[\hat{\alpha}\hat{\alpha}' - s^2\Lambda^{-1}] \quad (2.4)$$

Using unbiased estimates s^2 and $\hat{\alpha}\hat{\alpha}' - s^2\Lambda^{-1}$ for σ^2 and $\alpha\alpha'$ in (2.1) and (2.2), respectively, we get an unbiased estimate of $TMSE(k)$ and $J_w(k)$ as follows;

$$\hat{TMSE}(k) = s^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2} + k^2 Tr(\Lambda+kI_p)^{-2}(\hat{\alpha}\hat{\alpha}' - s^2\Lambda^{-1}), \quad (2.5)$$

$$\hat{J}_w(k) = s^2 Tr[\Lambda(\Lambda+kI_p)^{-2}M] + k^2 Tr[(\Lambda+kI_p)^{-1}M(\Lambda+kI_p)^{-1}(\hat{\alpha}\hat{\alpha}' - s^2\Lambda^{-1})] . \quad (2.6)$$

Note $\hat{TMSE}(k)$ and $\hat{J}_w(k)$ are continuous function of k. Also it can be easily checked that there exists a $k>0$ such that $\hat{TMSE}(k) < \hat{TMSE}(0)$ and $\hat{J}_w(k) < \hat{J}_w(0)$ by parallel arguments to $TMSE(k)$ in Horel & Kennard(1970) and $J_w(k)$ in Lim, Choi & Park(1980), respectively. Thus, for each of the objective functions $\hat{TMSE}(k)$ and $\hat{J}_w(k)$ there exist optimal k_T and k_w such that

$$\hat{TMSE}(k_T) = \min_k \hat{TMSE}(k)$$

and

$$\hat{J}_w(k_w) = \min_k \hat{J}_w(k).$$

When the w(x) takes the uniform mass $1/n$ at each design point, the $IMSE(\hat{y}_k)$ is

$$J_n(k) = \sum_{i=1}^{n} MSE(\hat{y}_k(z_i)) \frac{1}{n}.$$

But the C_k is defined by

$$C_k = \frac{\sum_{i=1}^{n} MSE(\hat{y}_k(z_i))}{\sigma^2} = \frac{n}{\sigma^2} J_n(k).$$

Thus the C_k criterion is equivalent to the *IMSE* criterion with a special case of weight function. Noting that

$$M = \sum_{i=1}^{n} z_i z_i' \frac{1}{n} = \frac{Z'Z}{n} = \frac{\Lambda}{n}$$

and then, substituting this into equation (2.2),

$$\begin{aligned} J_n(k) &= \frac{\sigma^2}{n} Tr[\Lambda(\Lambda+kI_p)^{-2}\Lambda] + \frac{k^2}{n} Tr[(\Lambda+kI_p)^{-1}\Lambda(\Lambda+kI_p)^{-1}\alpha\alpha'] \\ &= \frac{1}{n}\left[\sigma^2 \sum \frac{\lambda_i^2}{(\lambda_i+k)^2} + k^2 Tr[\Lambda(\Lambda+kI_p)^{-2}\alpha\alpha']\right]. \end{aligned} \quad (2.7)$$

By replacing σ^2 and $\sum_{1}^{n} MSE(\hat{y}_k(z_i))$ by their unbiased estimates, Mallows(1973) and Myers(1986) suggest a C_p-*like* statistic,

$$C_k = \frac{SSE(k)}{s^2} - n + 2 + 2Tr[H_k],$$

where $H_k = Z(Z'Z+kI_p)^{-1}Z'$, which is proportional to

$$\hat{J}_n(k) = \frac{s^2}{n} \sum \frac{\lambda_i^2}{(\lambda_i+k)^2} + \frac{k^2}{n} Tr[\Lambda(\Lambda+kI_p)^{-2}(\hat{\alpha}\hat{\alpha}' - s^2\Lambda^{-1})].$$

Recall that

$$E[\hat{\alpha}\hat{\alpha}' - s^2\Lambda^{-1}] = \alpha\alpha'.$$

For a quite small eigen-value λ, it might be possible that an unbiased estimate of α_i^2 is negative, i.e.,

$$\hat{\alpha}_i^2 - \frac{s^2}{\lambda_i} < 0,$$

which is absurd. Thus, it might be more reasonable to use an estimator Q of $\alpha\alpha'$ defined by

$$Q_{ij} = \begin{cases} \hat{\alpha}_i \hat{\alpha}_j & \text{if } i \neq j \\ \max\{0,\ \hat{\alpha}_i^2 - \frac{s^2}{\lambda_i}\} & \text{if } i = j \end{cases} \quad (2.8)$$

and then, to select k at which

$$\hat{J}_n(k) = \frac{s^2}{n} \sum \frac{\lambda_i^2}{(\lambda_i+k)^2} + \frac{k^2}{n} Tr[\Lambda(\Lambda+kI_p)^{-2}Q] \quad (2.9)$$

is minimized.

3. A robust choice of k

To find a $\hat{\alpha}(k^*)$ which behaves reasonably well w.r.t. $\hat{J}_u(k)$, $\hat{J}_n(k)$ (or C_k) and $\hat{TMSE}(k)$, we adopt the minimax principle in the statistical decision theory. First, we define the relative efficiency function of k in each of the objective function as follows;

$$\hat{Eff}_1(k) = \frac{\hat{J}_u(k_u)}{\hat{J}_u(k)}$$

$$\hat{Eff}_2(k) = \frac{\hat{J}_n(k_n)}{\hat{J}_n(k)}$$

$$\hat{Eff}_3(k) = \frac{\hat{TMSE}(k_T)}{\hat{TMSE}(k)}$$

Now we propose a robust selection rule of k in the sense of maximizing the worst relative efficiency function of k among $\hat{Eff}_1(k)$, $\hat{Eff}_2(k)$ and $\hat{Eff}_3(k)$, i.e., find k^* such that

$$\min_{1\le i\le 3} \hat{Eff}_i(k^*) = \max_k \min_{1\le i\le 3} \hat{Eff}_i(k).$$

Since $\hat{Eff}_i(k), 1\le i\le 3$, are continuous functions of k, there always exists such a robust k^*.

4. Example

To illustrate how to select a robust k^* for a quadratic polynomial regression model, we consider an example which was analyzed in Lim, Choi & Park(1980) for the *IMSE* criterion. The response variable w is the amount (parts per million, ppm) of water soluble in the soil and two explananatory variables are the concentration (weight percent, wt %) of clay(v_1) and the soil P.H.(v_2). Twenty observations were taken. The original data were presented and described by Myers(1971).

All the variables are standardized and the quadratic polynomial regression model in the standardized variables is given by

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^{*2} + \beta_{22} x_2^{*2} + \beta_{12} x_1^* x_2^* + \varepsilon,$$

where

$$y = \frac{w - \bar{w}}{S_w^{1/2}}$$

$$x_i = \frac{v_i - \bar{v}_i}{S_i^{1/2}}, \quad i = 1,2$$

$$x_i^* x_j^* = \frac{v_i v_j - \overline{v_i v_j}}{S_{ij}^{1/2}}, \quad i,j = 1,2$$

$$S_i = \sum_k (v_{ik} - \bar{v}_i)^2 \quad \& \quad S_{ij} = \sum_k (v_{ik} v_{jk} - \overline{v_i v_j})^2.$$

The orthogonal matrix P, $\hat{\alpha}$, s^2, the moments matrix

$$M_x = \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} x x' dx_1 dx_2, \quad x' = (x_1, x_2, x_1^{*3}, x_2^{*3}, x_1^{*}x_2^{*})$$

and $M_z = P' M_x P$ are given in Lim, Choi & Park(1980). An unbiased estimate s^2 given in Lim, Choi & Park(1980) is wrong and the correct value is

$$s^2 = 0.019649.$$

To estimate three criteria, $J_u(k)$, $J_n(k)$ and $TMSE(k)$, first we use unbiased estimates $\hat{\alpha}\hat{\alpha}' - s^2\Lambda^{-1}$ and s^2 for $\alpha\alpha'$ and σ^2, respectively. Next, we use a more reasonable estimates Q for $\alpha\alpha'$. For these two cases, the numerical optimization is performed using the *IMSL* subroutine duvmif. The numerical results are summerized in Table 1 and Table 2.

Table 1: $Eff = Min_{1\le i\le 3} Eff_i$ of selection rules using unbiased estimates $\hat{\alpha}\hat{\alpha}' - s^2\Lambda^{-1}$ for $\alpha\alpha'$

Criteria	Eff(%)
$\hat{J}_u(k)$	7.04
$\hat{J}_n(k)$ (or C_k)	16.61
$\hat{TMSE}(k)$	57.68
Robust	67.97

Table 2: $Eff = Min_{1\le i\le 3} Eff_i$ of selection rules using unbiased estimates Q for $\alpha\alpha'$

Criteria	Eff(%)
$\hat{J}_u(k)$	82.72
$\hat{J}_n(k)$ (or C_k)	94.50
$\hat{TMSE}(k)$	82.40
Robust	95.10

Using an unbiased estimates $\hat{\alpha}\hat{\alpha}' - s^2\Lambda^{-1}$ for $\alpha\alpha'$, the efficiency curve of $\hat{TMSE}(k)$ is quite flat in the neighbourhood of 0, which cause the relative efficiency of robust k^* being quite small, as we can see from Figure 1. robust k^* is $.26484\times 10^{-3}$ and the relative worst efficiency is $Eff = 67.97\%$. On the other hand, using Q instead of unbiased estimates for $\alpha\alpha'$, all three efficiency curves are sharply increased near 0 as we can see from Figure 2. Robust k^* is $.11572\times 10^{-3}$ and the relative worst efficiency is $Eff = 95.10\%$. For the C_k criterion, the optimal k_n using Q is $k_n = .11902\times 10^{-3}$ and the optimal k from the C_k (or $\hat{J}_n(k)$ using unbiased estimats $\hat{\alpha}\hat{\alpha}' - s^2\Lambda^{-1}$) is

$k = .15985 \times 10^{-3}$. In this example, the optimal k for the prediction-oriented criterion is more conservative than the total mean square error criterion of $\hat{\beta}(k)$ and the optimal k for $J_n(k)$ (using Q) is closer to robust k^*.

References

[1] Gunst, R. F. and Mason, R. L.(1977), "Biased estimation in ridge regression: Evaluation using mean squared error," *Journal of the American Statistical Associations* **72**: 616-628

[2] Hoerl, A. E. and Kennard, R. W.(1970), "Ridge regression: biased estimator for nonorthogonal problems," *Technometrics* **12**: 55-67

[3] Lim, Y. B., Choi, C. H. and Park, S. H.(1980), "Estimation of ridge regression under the integrated mean square error criterion," *Journal of Korean Statistical Society* **9**: 61-78

[4] Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics* **15**: 661-675.

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

[6] Myers, R. H. (1986), *Classical and Modern Regression with Applications*. Duxbury Press, Boston.

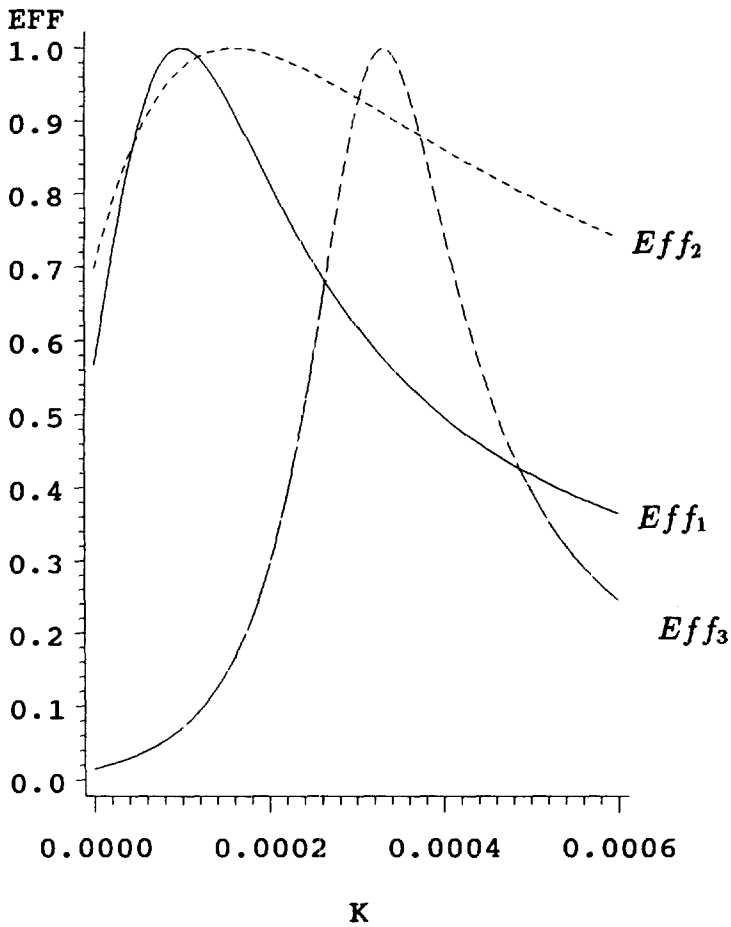

Figure 1. Efficiency Curve of selection rules using unbiased estimates $\hat{\alpha}\hat{\alpha}' - s^2\Lambda^{-1}$ for $\alpha\alpha'$

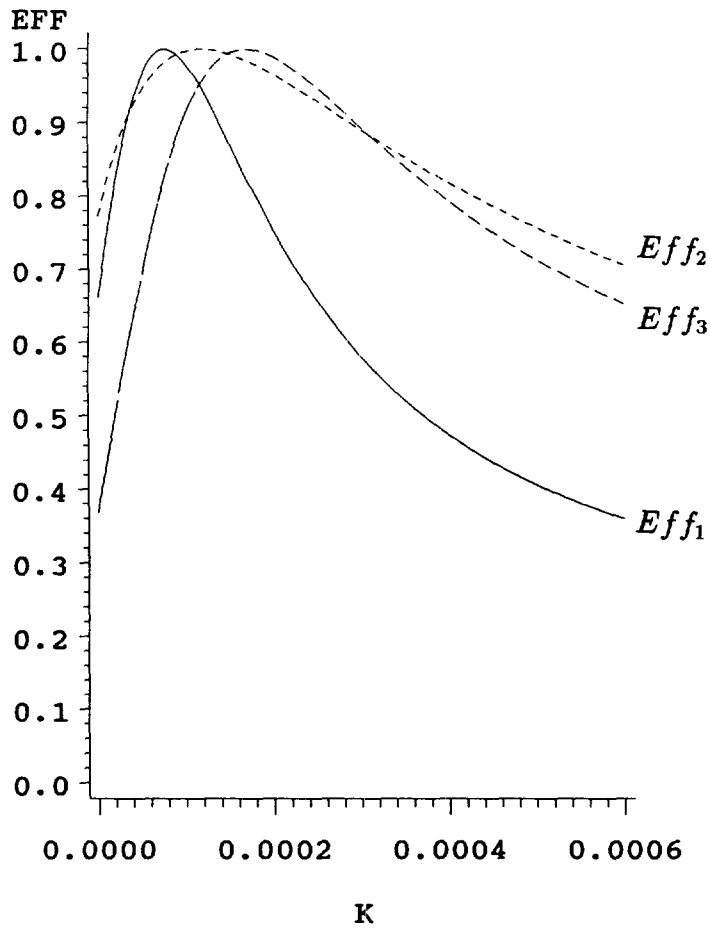

Figure 2. Efficiency Curve of selection rules using Q for $\alpha\alpha'$

능형회귀에서의 로버스트한 k의 선택 방법[1)]

임 용 빈[2)]

요 약

표준화된 중회귀모형에서 다중공선성(multicollinearity)이 존재할 때, 공선성(collinearity)의 영향을 완화하기 위해서 능형회귀가 사용된다. 반응변수의 예측을 위한 기준으로서 반응변수의 예측치의 평균제곱합(MSE)을 설명변수의 관심영역 R에서 적분한(IMSE) $J_w(k)$ 기준이 Lim, Choi & Park(1980)에 의해 소개되었다. C_k기준이 설명변수의 관심영역 R상에서의 가중치 함수인 w(x)가 각각의 자료점에서 등확률 1/n을 갖는 경우의 IMSE 기준인 $J_n(k)$ 기준과 동치라는 관계를 이용함으로 C_k 기준에 대해서 Myers(1986)에 의해 주어진 k의 선택방법 보다 더 합리적이라 기대되는 k의 선택방법이 제시되었다. 다음으로 관심이 있는 모든 기준들에 대해서 상대적으로 효율이 좋은 능형회귀추정량 $\beta(k)$를 선택하기 위해서, 관심이 있는 기준들 간의 가장 나쁜 효율을 최대화한다는 의미에서 MiniMax 원칙을 채택하여 관심이 있는 기준들에 대해서 로버스트한 k^*의 선택방법을 제시하였다.

[1)] 이 논문은 1991년도 교육부 지원 한국학술진흥재단의 자유공모과제 학술연구조성비에 의하여 연구되었음

[2)] (120-750) 서울특별시 서대문구 대현동 이화여자대학교 통계학과