ASYMPTOTIC ASPECTS OF RIDGE REGRESSION

by

Myung-Hoe Huh and Ingram Olkin

TECHNICAL REPORT NO. 196
APRIL 1984

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
MCS 81-04262

Ingram Olkin, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ASYMPTOTIC ASPECTS OF RIDGE REGRESSION

by

Myung-Hoe Huh and Ingram Olkin

TECHNICAL REPORT NO. 196
APRIL 1984

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
MCS 81-04262

Ingram Olkin, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ASYMPTOTIC ASPECTS OF RIDGE REGRESSION

Myung-Hoe Huh and Ingram Olkin
Stanford University

ABSTRACT

Ridge regression is considered as an alternative to least squares regression for the analysis of multicollinear data. When the ridge factor is estimated from the data, we obtain the asymptotic distribution of several ridge factors, and propose a new ridge factor based on asymptotic considerations. The mean squared errors of operational ridge regression estimators are evaluated asymptotically, and we propose several statistics useful for the selection of ridge factors.

Key words and phrases: Ridge regression, ridge factor, multicollinearity, asymptotic distribution of ridge factor.
1. **Introduction**

Regression analysis often contains multicollinear data. When this occurs, the least squares estimator (LSE) may not be satisfactory because of its high variability and its computational instability. Hoerl and Kennard (1970) introduced ridge regression in the statistical literature in order to obtain an estimator that dominates the LSE in the presence of multicollinearity.

Consider the standard multiple linear regression model

\[(1.1) \quad y = e\mu + X\beta + \varepsilon ,\]

where \(y\) is an \(n\)-dimensional column vector of responses, \(e = (1, \ldots, 1)'\) is \(n \times 1\), \(\mu\) is an unknown parameter, \(X\) is an \(n \times p\) matrix of known constants such that the columns of \(X\) are orthogonal to the vector \(e\), \(\beta\) is a \(p\)-dimensional column vector of regression coefficients, and \(\varepsilon\) follows an \(n\)-variate normal distribution with mean vector \(0\) and variance-covariance matrix \(\sigma^2 I\). Since the expression \(X'X\) occurs often, we write \(A = X'X\) and \(S = X'X/n\). In this study, \(\sigma^2\) is assumed to be known unless otherwise specified. In practice, \(\sigma^2\) is estimated by

\[(1.2) \quad \hat{\sigma}^2 = (y-e\hat{\mu}-X\hat{\beta})'(y-e\hat{\mu}-X\hat{\beta})/(n-p-1) ,\]

where \(\hat{\mu} = \bar{y}\), and

\[(1.3) \quad \hat{\beta} = (X'X)^{-1}X'y = A^{-1}X'y\]
is the ordinary least squares estimator. When the matrix $X$ is nearly multicollinear, the matrix $A$ will be close to singular.

The ridge estimator of $\beta$ is defined by perturbing the matrix $A$ away from singularity:

\[ \hat{\beta}(k) = (X'X + kI)^{-1}X'y = (A + kI)^{-1}X'y, \]

where $k$ is usually a stochastic quantity, say $\hat{k}$. In the study of ridge regression, it is usual to assume that $X$ is centered and scaled so that the diagonal elements of $A$ are all equal to 1. Instead, we assume that $X$ is centered and scaled so that the diagonal elements of $S$ are all equal to 1. This serves to help establish asymptotic results about operational ridge estimators.

It is sometimes convenient to represent (1.1) in another form. Write $S = PD\lambda P'$, where $D\lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ is the diagonal matrix with ordered eigenvalues, $\lambda_1 \geq \ldots \geq \lambda_p$, of $S = X'X/n$, and $P$ is the $p \times p$ orthogonal matrix of eigenvectors. Let $Z = XP$ and $\alpha = P'\beta$. Thus the columns of $Z$ consist of principal components of $X$. (1.1) can be written as

\[ y = e\mu + Z\alpha + \varepsilon. \]

The variance-covariance matrix of the least squares estimator of $\alpha$ is diagonal and equal to $\sigma^2 D\lambda^{-1}/n$.

The least squares estimator $\hat{\beta}$ of (1.3) is unbiased and has the variance-covariance matrix $A^{-1}\sigma^2$. Define the MSE to be the trace of the matrix of mean squared errors:
\text{(1.6)} \quad \text{MSE}(\hat{\beta}) \equiv E(\hat{\beta} - \beta)'(\hat{\beta} - \beta) = \sigma^2 \text{tr} \ A^{-1} = (\sigma^2 / n) \sum_{i=1}^{p} 1/l_i .

When \( \lambda_p \) is close to zero, \( \text{MSE}(\hat{\beta}) \) becomes very large. On the other hand, the ridge estimator \( \hat{\beta}(k) \) of (1.4), though biased, has smaller variance than the LSE. More specifically, when \( k \) is a fixed constant,

\text{(1.7)} \quad \text{Bias} \ \hat{\beta}(k) \equiv E(\hat{\beta}(k) - \beta) = (A + kI)^{-1}A\beta - \beta = -k(A + kI)^{-1}\beta ,

\text{(1.8)} \quad \text{Var} \ \hat{\beta}(k) = E(\hat{\beta}(k) - E\hat{\beta}(k))(\hat{\beta}(k) - E\hat{\beta}(k))' \\
= \sigma^2 (A + kI)^{-1}A(A + kI)^{-1} = \sigma^2 (A^{1/2} + kA^{-1/2})^2 .

Hence we obtain from (1.7) and (1.8) that

\text{(1.9)} \quad \text{MSE}(\hat{\beta}(k)) = \sigma^2 \text{tr} (A + kI)^{-1}A(A + kI)^{-1} + k^2 \beta'(A + kI)^{-2}\beta .

Hoerl and Kennard (1970) prove the important "existence theorem" to the effect that there exists a value of \( k \) for which ridge regression has a smaller MSE than the LSE. More precisely, there exists a non-negative \( k_0 \) such that

\text{(1.10)} \quad E(\hat{\beta}(k_0) - \beta)'(\hat{\beta}(k_0) - \beta) \leq E(\hat{\beta} - \beta)'(\hat{\beta} - \beta) .

Because the ridge factor is not a constant and has to be estimated from the data, Coniffe and Stone (1973) and Thisted (1976) criticize this "existence theorem" as not being of practical importance. Furthermore, their claims are supported by Monte-Carlo simulation studies. In the
present study, their claims are also verified from the evaluation of mean squared errors asymptotically.

The choice of a stochastic ridge factor $\hat{k}$ is an important component in the study of ridge regression. In Hoerl and Kennard's original paper, they propose using the ridge trace, which is a graph plotting the values of each $\hat{\beta}_i(k)$ against $k$, for $i = 1, \ldots, p$. Although ridge traces are often used in practice (see e.g., Marquardt and Snee, 1975, and Chatterjee and Price, 1977), the choice of a ridge factor based on the ridge trace is essentially subjective.

Again Hoerl, Kennard, and Baldwin (1975) proposed using the ridge factor

\begin{equation}
\hat{k}_{HKB} = \frac{p \sigma^2}{\hat{\beta}' \hat{\beta}},
\end{equation}

where, for convenience, the subscript is named after the authors. Thisted (1976) proposed $\hat{k}_{HKBM}$, a modified version of $\hat{k}_{HKB}$:

\begin{equation}
\hat{k}_{HKBM} = (p-2)\sigma^2/\hat{\beta}' \hat{\beta},
\end{equation}

where the subscript is named after "Hoerl-Kennard-Baldwin-modified". Of course, for large $p$, (1.11) and (1.12) are not too different.

As a consequence of studying some asymptotic properties of the ridge regression as $n$ approaches infinity, we propose another ridge factor, $\hat{k}_A$, where the subscript "A" is an abbreviation of "asymptotic". This ridge factor is derived from the following idea. First, find the value of $k$ that minimizes the asymptotic MSE of $\hat{\beta}(k)$. The minimizer, $k_A$, is a function
of unknown parameters. The proposed ridge factor $\hat{k}_A$ is obtained by replacing the unknown parameters by their estimates. This ridge factor should have good performance asymptotically, and the question is whether it will also perform well for small samples. The specific expression for $\hat{k}_A$ is

$$\hat{k}_A = [\sigma^2 \text{tr} (X'X)^{-2}] / [\hat{\beta}'(X'X)^{-2}\hat{\beta}] = \sigma^2 (\text{tr} \ A^{-2}) / (\hat{\beta}'A^{-2}\hat{\beta})$$

(1.13)

The asymptotic ridge factor, $\hat{k}_A$, is discussed further in Section 2, and in Section 3 asymptotic distributions of different ridge factors are derived. In Section 4, asymptotic MSE's are evaluated for some operational ridge rules, and several statistics are applied (Section 5) to the data of Longley (1967), which consists of six collinear independent variables and one dependent variable. This example shows that $\hat{\beta}(\hat{k}_A)$ is a reasonable estimator in this case.

2. Asymptotic Ridge Factor

Assume that, as $n \to \infty$, $S = X'X/n$ converges to a positive definite matrix $Q$ that has the same eigenvalues $\lambda_1, \ldots, \lambda_p$ as $S$. As a consequence, the LSE $\hat{\beta}$ converges to $\beta$ in probability. From (1.9) we have the asymptotic expression for the MSE $\hat{\beta}(k)$ for a fixed ridge factor $k$.

**Theorem 2.1.** For a fixed ridge factor $k$,

$$\text{MSE}(\hat{\beta}(k)) = \frac{\sigma^2}{n} (\text{tr} \ S^{-1}) + \frac{1}{n^2} [-2\sigma^2 k (\text{tr} \ S^{-2}) + k^2 \hat{\beta}'S^{-2}\hat{\beta}] + o\left(\frac{1}{n^3}\right).$$

(2.1)
Moreover, if

\[ k_A = \frac{\sigma^2 \text{tr} (X'X)^{-2}}{[\beta'(X'X)^{-2} \beta]} \]

then

\[ \text{MSE}(\hat{\beta}(k_A)) = \frac{\sigma^2}{n} \left( \text{tr} \ S^{-1} \right) - \frac{\sigma^4}{n^2} \left( \text{tr} \ S^{-2} \right)^2 / (\beta'S^{-2}\beta) + O\left(\frac{1}{n^3}\right) \]

**Proof.** Recall that $\alpha = P'\beta$; from (1.9),

\[ \text{MSE}(\hat{\beta}(k)) = \frac{\sigma^2}{n} \text{tr}(D + \frac{kl}{n})^{-1}D + \frac{kl}{n})^{-1} + \frac{1}{n} k^2 \alpha'(D + \frac{kl}{n})^{-2} \alpha \]

\[ = (1/n)\sigma^2 \sum_{i=1}^{P} (\ell_i + k/n)^{-2} \ell_i + (1/n)^2 k^2 \sum_{i=1}^{P} \frac{\alpha_i}{(\ell_i + k/n)^2} \]

\[ = (1/n)\sigma^2 \sum_{i=1}^{P} \ell_i^{-1} + (1/n)^2 \sum_{i=1}^{P} \frac{-2\sigma^2 k/\ell_i^2 + \alpha_i^2 k^2/\ell_i^2}{\ell_i^2} + O(n^{-3}) \]

Hence MSE $(\hat{\beta}(k))$ achieves its minimum at $k = k_A$ defined by (2.2), and (2.3) is obtained by evaluating (2.4) at $k = k_A$. \|}

Note that, if $k_A$ is replaced by its estimate $\hat{k}_A$ in Theorem 2.1, then the dominance of the corresponding ridge estimator over the least squares estimator is no longer true, as shown in Theorem 4.1.

**Remark.** From (2.4) the MSE of $\hat{\beta}(k)$ is minimized by choosing $k$ to satisfy
\[(2.5) \quad \sum_{i=1}^{P} \frac{\lambda_i (\alpha_i^2 k - \sigma^2)}{[(\lambda_i + k/n)^3]} = 0.\]

The solution \( k \) of (2.5) is called "SRIDG" factor by Dempster, Schatzoff, and Wermuth (1977). As \( n \) approaches infinity, (2.5) is equivalent to

\[(2.6) \quad \sum_{i=1}^{P} \frac{\alpha_i^2 k - \sigma^2}{\lambda_i} = 0.\]

which can be solved for \( k \) to yield

\[(2.7) \quad k_A = \frac{\sigma^2 \sum_{i=1}^{P} (1/\lambda_i)}{\sum_{i=1}^{P} (\alpha_i^2/\lambda_i^2)} = \frac{[\sigma^2 \text{ tr } S^{-2}]/[\beta ^t S^{-2} \beta]}{[\sigma^2 \text{ tr } S^{-2}]/[\beta ^t S^{-2} \beta]} .\]

Therefore \( k_A \) can be considered to be the "SRIDG" factor in the asymptotic case.

When the eigenvalues of \( S \) are all equal, then \( \hat{k}_{HKB} \) and \( \hat{k}_A \) coincide. As pointed out by Hoerl, Kennard, and Baldwin (1975), \( \hat{k}_{HKB} \) is the harmonic mean of \( \sigma^2/\hat{\alpha}_1^2, \ldots, \sigma^2/\hat{\alpha}_p^2 \). On the other hand, as can be seen in (2.7), \( \hat{k}_A \) is the weighted harmonic mean of \( \sigma^2/\hat{\alpha}_1^2, \ldots, \sigma^2/\hat{\alpha}_p^2 \) with weights proportional to \( \hat{\lambda}_1^{-2}, \ldots, \hat{\lambda}_p^{-2} \).

3. Asymptotic Distribution of Ridge Factors

First we derive the asymptotic distribution of three ridge factors, \( \hat{k}_{HKB}, \hat{k}_{HKB'} \), and \( \hat{k}_A \), defined by (1.11), (1.12), and (1.13), respectively. To avoid the trivial case, assume that \( \beta ^t \beta (= \alpha ^t \alpha) \neq 0 \). Let \( v_1, \ldots, v_p \) be independently and identically distributed standard normal random variates. Then, for \( i = 1, \ldots, p \), we can write
(3.1) \[ \hat{\alpha}_i = \alpha_i + \sigma(n\lambda_i^{-1})^{-1/2} v_i. \]

Hence

(3.2) \[ \hat{k}_A = \left[ \sigma^2 \sum_{i=1}^{p} \lambda_i^{-2} \right] / \left[ \sum_{i=1}^{p} (\hat{\alpha}_i^2 / \lambda_i^2) \right] \]

\[ = \left[ \sigma^2 \sum_{i=1}^{p} \lambda_i^{-2} \right] / \left[ \sum_{i=1}^{p} \alpha_i^2 / \lambda_i^2 + 2n^{-1/2} \sigma \sum_{i=1}^{p} \lambda_i^{-5/2} \alpha_i v_i + n^{-1} \sigma \sum_{i=1}^{p} \lambda_i^{-3} v_i^2 \right] \]

\[ = k_A \left[ 1 - \left\{ 2n^{-1/2} \sigma \sum_{i=1}^{p} \lambda_i^{-5/2} \alpha_i v_i \right\} / \left[ \sum_{i=1}^{p} \alpha_i^2 / \lambda_i^2 \right] \right] + O_p(n^{-1}), \]

where \( k_A \) is defined in (2.2).

Consequently, \( \hat{k}_A \) converges to \( k_A \) in probability, and

(3.3) \[ n^{1/2}(\hat{k}_A - k_A) \overset{d}{\to} N(0, \nu_A^\infty), \]

where

(3.4) \[ \nu_A^\infty = 4\sigma^2 k_A^2 \left[ \left( \sum_{i=1}^{p} \alpha_i^2 / \lambda_i^5 \right) / \left( \sum_{i=1}^{p} \alpha_i^2 / \lambda_i^2 \right)^2 \right]. \]

Because the standard error of \( \hat{k}_A \) is proportional to \( k_A \), it is advantageous to use \( \log \hat{k}_A \) instead of \( \hat{k}_A \) directly. Then \( \log \hat{k}_A \) converges to \( \log k_A \) in probability, and

(3.5) \[ n^{1/2}(\log \hat{k}_A - \log k_A) \overset{d}{\to} N(0, \omega_A^\infty), \]
where

\begin{equation}
\hat{w}^\infty_A = 4\sigma^2 \left[ \frac{\sum_{i=1}^{P} \alpha_i^2 / \lambda_i^5}{\sum_{i=1}^{P} \alpha_i^2 / \lambda_i^2} \right].
\end{equation}

To obtain an approximate confidence interval for \( k_A \), define

\begin{equation}
\hat{w}^\infty_A = 4\sigma^2 \left[ \frac{\sum_{i=1}^{P} \hat{\alpha}_i^2 / \hat{\lambda}_i^5}{\sum_{i=1}^{P} \hat{\alpha}_i^2 / \hat{\lambda}_i^2} \right].
\end{equation}

Since \( \text{plim} \hat{w}^\infty_A = w^\infty_A \),

\begin{equation}
n^{1/2}(\log \hat{k}_A - \log k_A) / (\hat{w}^\infty_A)^{1/2} \overset{d}{\to} N(0,1) .
\end{equation}

Consequently, an approximate confidence interval \((k_L, k_U)\) for \( k_A \) with confidence coefficient \( 1 - \alpha \) is given by

\begin{equation}
k_L = \exp[\log \hat{k}_A - z_{\alpha/2}(\hat{w}^\infty_A)^{1/2}], \quad k_U = \exp[\log \hat{k}_A + z_{\alpha/2}(\hat{w}^\infty_A)^{1/2}] ,
\end{equation}

where \( z_{\alpha/2} \) denotes the upper \( 100(\alpha/2) \) percentile of standard normal distribution.

Similarly, for the Hoerl-Kennard-Baldwin (HKB) ridge factor,

\begin{equation}
\hat{k}_{\text{HKB}} = \frac{p\sigma^2}{\sum_{i=1}^{P} \hat{\alpha}_i^2}
= \frac{p\sigma^2}{\sum_{i=1}^{P} \alpha_i^2 + 2n^{-1/2} \sigma \sum_{i=1}^{P} \frac{1}{\lambda_i} \alpha_i v_i + n^{-1} \sigma^2 \sum_{i=1}^{P} \frac{1}{\lambda_i} v_i^2}
= k_{\text{HKB}} [1 - (2n^{-1/2} \sigma \sum_{i=1}^{P} \frac{1}{\lambda_i} \alpha_i v_i) / (\sum_{i=1}^{P} \alpha_i^2)] + O_p(n^{-1}) .
\end{equation}
Thus \( \lim_{n \to \infty} \log \hat{k}_{\text{HKB}} = \log k_{\text{HKB}} \), and

\[
(3.11) \quad n^{1/2}(\log \hat{k}_{\text{HKB}} - \log k_{\text{HKB}}) \xrightarrow{d} N(0, \omega_{\text{HKB}}^\infty),
\]

where

\[
(3.12) \quad \omega_{\text{HKB}}^\infty = 4\sigma^2 \left[ \left( \frac{\sum_{i=1}^{p} \alpha_i^2 \beta_i}{\sum_{i=1}^{p} \alpha_i^2} \right) - \frac{\sum_{i=1}^{p} \alpha_i^2}{\sum_{i=1}^{p} \alpha_i^2} \right].
\]

For the "Hoerl-Kennard-Baldwin-modified" (HKBM) ridge factor,

\[
(3.13) \quad n^{1/2}(\log \hat{k}_{\text{HKBM}} - \log k_{\text{HKBM}}) \xrightarrow{d} N(0, \omega_{\text{HKBM}}^\infty),
\]

where \( \omega_{\text{HKBM}}^\infty = \omega_{\text{HKB}}^\infty \).

4. **An Asymptotic Evaluation of Several Ridge Rules**

We now obtain the MSE of the ridge estimator using \( \hat{k}_A \).

**Theorem 4.1.**

\[
(4.1) \quad \text{MSE}(\hat{k}_A) = \left( \sigma^2/n \right) \text{tr } S^{-1} - \left( 1/n^2 \right) a + O(n^{-5/2}),
\]

where

\[
(4.2) \quad a = \sigma^2 (\text{tr } S^{-2})^2/\beta' S^{-2} \beta - 4\sigma^4 (\beta' S^{-4} \beta) (\text{tr } S^{-2})/\beta' S^{-2} \beta)^2.
\]
Proof. Note that

\[
(4.3) \quad \hat{\beta}(\hat{k}) = (A + \hat{k}I)^{-1}X'y = \sum_{j=0}^{\infty} (-\hat{k})^j A^{-j} \hat{\beta} = \hat{\beta} - \hat{k}A^{-1} \hat{\beta} + O_p(n^{-2})
\]

and that

\[
(4.4) \quad \hat{\beta}(\hat{k}_A) - \hat{\beta}(k_A) = -(\hat{k}_A - k_A)A^{-1} \hat{\beta} + (\hat{k}_A - k_A)^{-2} \hat{\beta} + \ldots
\]

\[
= O_p(n^{-3/2}) + O_p(n^{-5/2})
\]

Consequently,

\[
(4.5) \quad \text{MSE } \hat{\beta}(\hat{k}_A) = E[(\hat{\beta}(\hat{k}_A) - \hat{\beta}(k_A) - \hat{\beta}(\hat{k}_A))' \{\hat{\beta}(\hat{k}_A) - \hat{\beta}(k_A) - \hat{\beta}(\hat{k}_A)\}]
\]

\[
= \text{MSE } \hat{\beta}(k_A) + 2E[(\hat{\beta}(k_A) - \beta)'(\hat{\beta}(k_A) - \hat{\beta}(k_A))] + O(n^{-3})
\]

Let \( v_1, \ldots, v_p \) be independently and identically distributed standard normal random variates. Then, by (3.1), (3.2), (4.3), and (4.4),

\[
(4.6) \quad E[(\hat{\beta}(k_A) - \beta)'(\hat{\beta}(k_A) - \hat{\beta}(k_A))]
\]

\[
= E[((\hat{k}_A - k_A)^{-1} \hat{\beta} + O_p(n^{-2}))'((k_A - \hat{k}_A)A^{-1} \hat{\beta} + O_p(n^{-5/2}))]
\]

\[
= E[(k_A - \hat{k}_A)(\hat{\beta} - \beta)' A^{-1} \hat{\beta}] + O(n^{-5/2})
\]

\[
= 2E[n^{-1/2} \xi_{\hat{A}}(\sum_{i=1}^{p} \xi_i^{-5/2} \alpha_i \xi_i')/(\sum_{i=1}^{p} \alpha_i^2 / \xi_i^2)]
\]

\[
\cdot \left[ \sum_{i=1}^{p} (n^{-3/2} \xi_i^{-3/2} \sigma \xi_i')(\alpha_i + n^{-1/2} \xi_i^{-1/2} \sigma \xi_i')] + O(n^{-5/2})
\]
\[
= 2k_A \sigma^2 n^{-2} \left( \sum_{i=1}^{P} \frac{\alpha_i^2}{\nu_i} \right) / \left( \sum_{i=1}^{P} \frac{\alpha_i^2}{\nu_i^2} \right) + o(n^{-5/2}) \\
= 2n^{-2} \sigma^2 (\beta' S^{-2} \beta) (\text{tr } S^{-2}) / (\beta' S^{-2} \beta)^2 + o(n^{-5/2}) .
\]

By (2.3), (4.5) and (4.6), the proof of (4.1) is completed.

**Corollary 4.1.** If

\[(4.7) \quad (\text{tr } S^{-2}) (\beta' S^{-2} \beta) > 4 \beta' S^{-2} \beta ,\]

Then \( \hat{\beta}(\hat{k}_A) \) has a smaller asymptotic MSE than that of the LSE. If the inequality in (4.7) is reversed, then \( \hat{\beta}(\hat{k}_A) \) has a larger asymptotic MSE than that of the LSE.

Next we evaluate the MSE of the ridge estimator using \( \hat{k}_{HKB} \). The proof of the following theorem is parallel to that of Theorem 4.1, and is omitted.

**Theorem 4.2.**

\[(4.8) \quad \text{MSE}(\hat{\beta}(\hat{k}_{HKB})) = (\sigma^2 / n) \text{tr } S^{-1} - (1/n^2) p \sigma^4 b / (\beta' \beta)^2 + o(n^{-5/2}) ,\]

where

\[(4.9) \quad b = 2(\beta' \beta) (\text{tr } S^{-2}) - (p+4)(\beta' S^{-2} \beta) .\]

**Corollary 4.2.** If

\[(4.10) \quad 2(\beta' \beta) (\text{tr } S^{-2}) > (p+4)(\beta' S^{-2} \beta) ,\]

12
then \( \hat{\beta}(\hat{k}_{\text{HKBM}}) \) has a smaller asymptotic MSE than that of the LSE. If the inequality in (4.10) is reversed, then \( \hat{\beta}(\hat{k}_{\text{HKBM}}) \) has a larger asymptotic MSE than that of the LSE.

**Theorem 4.3.**

\[
\text{MSE}(\hat{\beta}(\hat{k}_{\text{HKBM}})) = \frac{\sigma^2}{n} \text{tr} \ S^{-1} - \frac{1}{n^2} (p-2)\sigma^4 c/(\beta'\beta)^2 + O(n^{-5/2}) ,
\]

where

\[
(4.12) \quad c = 2(\beta'\beta)(\text{tr} \ S^{-2}) - (p+2)(\beta'S^{-2}\beta) .
\]

**Corollary 4.3.** If

\[
(4.13) \quad 2(\beta'\beta)(\text{tr} \ S^{-2}) > (p+2)(\beta'S^{-2}\beta) ,
\]

then \( \hat{\beta}(\hat{k}_{\text{HKBM}}) \) has a smaller asymptotic MSE than that of the LSE. If the inequality in (4.13) is reversed, then \( \hat{\beta}(\hat{k}_{\text{HKBM}}) \) has a larger asymptotic MSE than that of the LSE.

5. **Choice of Ridge Factors**

Using Corollaries 4.1, 4.2, and 4.3, we can set up guidelines for the pairwise choice between each ridge factor and the LSE.

The asymptotic MSE of \( \hat{\beta}(\hat{k}) \) using \( \hat{k}_A, \hat{k}_{\text{HKBM}}, \) and \( \hat{k}_{\text{HKBM}} \) are smaller than that of the LSE, if and only if,
(5.1) \[ Q_A(\beta) \equiv (\text{tr } S^{-2})(\beta'S^{-2}\beta) - 4\beta'S^{-4}\beta > 0 , \]

(5.2) \[ Q_{\text{HKB}}(\beta) \equiv 2(\text{tr } S^{-2})(\beta'\beta) - (p+4)(\beta'S^{-2}\beta) > 0 , \]

(5.3) \[ Q_{\text{HKBM}}(\beta) \equiv 2(\text{tr } A^{-2})(\beta'\beta) - (p+2)(\beta'S^{-2}\beta) > 0 , \]

respectively.

Unbiased estimators of the three expressions \( Q_A(\beta) \), \( Q_{\text{HKB}}(\beta) \), and \( Q_{\text{HKBM}}(\beta) \) are

(5.4) \[ t_A \equiv Q_A(\hat{\beta}) - (\sigma^2/n)[(\text{tr } S^{-2})(\text{tr } S^{-3}) - (\text{tr } S^{-5})] , \]

(5.5) \[ t_{\text{HKB}} \equiv Q_{\text{HKB}}(\hat{\beta}) - (\sigma^2/n)[2(\text{tr } S^{-1})(\text{tr } S^{-2}) - (p+4)(\text{tr } S^{-3})] , \]

(5.6) \[ t_{\text{HKBM}} \equiv Q_{\text{HKBM}}(\hat{\beta}) - (\sigma^2/n)[2(\text{tr } S^{-1})(\text{tr } S^{-2}) - (p+2)(\text{tr } S^{-3})] , \]

respectively.

In practice, these three statistics can be used as a guide for the choice of ridge factors. If none of these statistics is positive, we recommend using the least squares regression estimator \( \hat{\beta} \). If only one of these is positive, we recommend using the corresponding ridge regression estimator. If two or more of these are positive, then, because ridge estimators often suffer from over-shrinkage, we recommend using the minimum of the corresponding ridge factors.
6. **An Example--Longley Data**

The Longley (1967) data is popular as a data set for multiple regression analysis of multicollinear data. Longley (1967) and Beaton, Rubin, and Barone (1976) studied the numerical accuracy of the LSE using various computing algorithms and machines.

The data consists of one dependent variable and six independent variables.

\[
y: \quad \text{Total Derived Employment (in thousands)}
\]
\[
x_1: \quad \text{Gross National Product Implicit Price Deflator (in tenth)}
\]
\[
x_2: \quad \text{Gross National Product (in millions)}
\]
\[
x_3: \quad \text{Unemployment (in thousands)}
\]
\[
x_4: \quad \text{Size of Armed Forces (in thousands)}
\]
\[
x_5: \quad \text{Noninstitutional Population, 14 Years of Age or Over (in thousands)}
\]
\[
x_6: \quad \text{Year (1947-1962)}
\]

Define the matrix \( X_c \) to be the centered matrix \( X \). The eigenvalues of \( X'_c X_c \) are \( 1.4894 \times 10^{11}, 2.2439 \times 10^7, 2.7507 \times 10^6, 1.7407 \times 10^6, 1.3389 \times 10, \) and \( 4.4918 \times 10^{-1} \). A condition number, the ratio of the largest eigenvalue over the smallest eigenvalue, is \( 3.3 \times 10^{11} \), indicating severe multicollinearity. However, if \( X \) is centered and scaled, we can avoid non-essential ill-conditioning. Let \( X_{cs} \) be obtained from the matrix \( X_c \) by scaling (dividing each column of \( X_c \) by its sample standard deviation). The eigenvalues of \( X'_{cs} X_{cs} \) are \( 69.0505, 17.6301, 3.0515, 0.2240, 0.0383, \) and \( 0.0057 \), and the condition number is reduced to \( 1.2 \times 10^4 \).
Fitting the data to the multiple linear regression model

\[(6.1) \quad y = e \mu + x_{cs} \beta + \varepsilon .\]

The three ridge factors are computed to be

\[(6.2) \quad \hat{k}_{HKB} = 0.0060, \quad \hat{k}_{HKBM} = 0.0040, \quad \hat{k}_A = 0.0020 .\]

The corresponding estimates of regression coefficients are summarized in Table 1.

<table>
<thead>
<tr>
<th>(k)</th>
<th>LSE</th>
<th>A</th>
<th>HKBM</th>
<th>HKB</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{\beta}_1)</td>
<td>164.63</td>
<td>19.71</td>
<td>-33.25</td>
<td>-45.24</td>
</tr>
<tr>
<td>(\hat{\beta}_2)</td>
<td>-3578.68</td>
<td>-2133.63</td>
<td>-1301.28</td>
<td>-764.80</td>
</tr>
<tr>
<td>(\hat{\beta}_3)</td>
<td>-1890.40</td>
<td>-1690.90</td>
<td>-1572.70</td>
<td>-1494.08</td>
</tr>
<tr>
<td>(\hat{\beta}_4)</td>
<td>-719.25</td>
<td>-678.28</td>
<td>-651.53</td>
<td>-631.95</td>
</tr>
<tr>
<td>(\hat{\beta}_5)</td>
<td>-343.28</td>
<td>-743.17</td>
<td>-923.23</td>
<td>-1002.57</td>
</tr>
<tr>
<td>(\hat{\beta}_6)</td>
<td>8714.43</td>
<td>7665.89</td>
<td>6977.94</td>
<td>6473.32</td>
</tr>
</tbody>
</table>

It is interesting to note that the sign of \(\hat{\beta}_1(\hat{k})\) changes from positive to negative and that the absolute value of \(\hat{\beta}_5(\hat{k})\) becomes larger as \(\hat{k}\) increases from 0.0 to 0.006.

For the choice of estimators of \(\beta\), \(t_A\) of (5.4), \(t_{HKB}\) of (5.5), and \(t_{HKBM}\) of (5.6), are computed:
(6.3) \( t_A = 1.4 \times 10^{20}, \quad t_{\text{HKB}} = -2.1 \times 10^{16}, \quad t_{\text{HKBM}} = -8.0 \times 10^{14} \).

According to the guidelines for the choice of ridge factors in Section 5, \( \hat{\beta}(k_A) \) is the most preferred choice.
REFERENCES


