NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

RIDGE ESTIMATORS

Felice du Berger

A Thesis

in

The Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the degree of Master of Science at
Concordia University
Montreal, Quebec, Canada

April 1989

ISBN 0-315-51310-1

Canada

ABSTRACT

Ridge Estimators

Felice du Berger

When the method of least squares is applied to a regression model where the independent variables are nearly collinear, very poor estimates of the regression coefficients result. The variance of the least squares estimates will be considerably inflated and the length of the vector of least squares estimates will be too long on average. This implies that the absolute value of the parameter estimates will be too large and that they will be very unstable. That is, given a different sample, the magnitudes and signs may change considerably. A procedure for obtaining stable and accurate parameter estimates, when the independent variables are nearly collinear, is ridge regression, originally proposed by Hoerl and Kennard (1970 a b). Ridge estimators are biased estimators which achieve a reduction in variance by adding a small amount of bias to the estimation process. This thesis is a review of the various properties of ridge estimators. In addition, several new ridge estimators are proposed which are non-stochastic as they depend only on the eigenvalues of the $X'X$ matrix. A simulation is conducted which compares the performance of the new ridge estimators with other established non-stochastic ridge estimators.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Dr. T. D. Dwivedi, my thesis supervisor, for his guidance and support, during my years at Concordia University, as well as for his comments and suggestions during the preparation of this thesis.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

Consider the following standard linear regression model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots \beta_p x_{i,p} + \varepsilon_i, \text{ for } i = 1, \ldots, n, \tag{1.1}$$

where y_i is the $i-th$ of n observations on the dependent variable, $\beta_0, \ldots, \beta_p$ are unknown regression coefficients, $x_{i,j}$ are observations on the p regressors, and ε_i is an unknown random disturbance term. In matrix notation, we may write model (1.1) including all n observations as

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

where Y is a $(n \times 1)$ vector of observable variables, X is a $(n \times [p+1])$ matrix of known constants the first column of which is the unit vector, β is a $([p+1] \times 1)$ vector of unknown regression coefficients, and ε is an $(n \times 1)$ vector of unknown random errors.

The usual assumptions of the above model are that

1) X is a non-stochastic matrix of regressors,

2) X has full column rank, i.e., Rank(X) $= [p+1]$,

3) ε is unbiased, i.e., $E(\ \varepsilon_i\ ) = 0$ where E denotes the expectation operator,

4) $\varepsilon_i's$ are uncorrelated, i.e., $E(\ \varepsilon_i,\ \varepsilon_j\ ) = 0$ for $i \neq j$,

5) $\varepsilon_i's$ have identical variance, i.e., $E(\varepsilon_i, \varepsilon_i) = \sigma^2$ for $i = 1, \ldots, n$.

Given such a model and the above assumptions, the problem will be to estimate the unknown coefficient vector β . A well known estimator of β , when the distribution of ε is unknown, is the least squares estimator which minimizes the sum of squared errors denoted by Q .

Minimizing

$$\begin{aligned} Q &= (Y - X\beta)'(Y - X\beta) \\ &= Y'Y - 2\beta'X'Y + \beta'\beta, \end{aligned} \tag{1.3}$$

by differentiating Q with respect to β and equating to zero, we have

$$\frac{\partial Q}{\partial \beta} = 2X'X\beta - 2X'Y = 0, \tag{1.4}$$

giving, for $\beta = b$, the least squares normal equations denoted by

$$X'Xb = X'Y. \tag{1.5}$$

Now since X is of rank $[p+1]$, $(X'X)$ is a positive definite invertible matrix. Solving equation (1.5) for b gives the least squares estimator

$$b = (X'X)^{-1}X'Y. \tag{1.6}$$

The estimator b minimizes Q since

$$\frac{\partial^2 Q}{\partial \beta^2} = 2X'X, \tag{1.7}$$

is a positive definite matrix. Using the estimator b , we have now a predictor of the disturbance vector ε , denoted by the residual vector e, where $e = Y - Xb$. The

sum of the squared residuals will provide an unbiased estimator of the disturbance variance σ^2 . To see this consider

$$\begin{aligned} e'e &= (Y - Xb)'(Y - Xb) \\ &= (Y - X(X'X)^{-1}X'Y)'(Y - X(X'X)^{-1}X'Y) \\ &= Y'(I - X(X'X)^{-1}X')'(I - X(X'X)^{-1}X')Y \\ &= Y'(I - X(X'X)^{-1}X')Y \\ &= (X\beta + \varepsilon)'(I - X(X'X)^{-1}X')(X\beta + \varepsilon) \\ &= \varepsilon'(I - X(X'X)^{-1}X')\varepsilon, \end{aligned} \tag{1.8}$$

where I is the identity matrix, and $(I - X(X'X)^{-1}X')$ is an idempotent matrix such that

$$X'(I - X(X'X)^{-1}X') = (I - X(X'X)X')X = 0. \tag{1.9}$$

Now taking expectations of both sides, and using Theorem A3 from Appendix A we have

$$E(e'e) = \text{Tr } \sigma^2(I - X(X'X)^{-1}X') = \sigma^2(n - [p+1]), \tag{1.10}$$

where Tr denotes the trace operator. Thus we have the following unbiased estimator of σ^2 given by

$$s^2 = \frac{e'e}{\nu}, \tag{1.11}$$

where $\nu = n - \text{Rank } X$.

If we were to further assume that the disturbances follow a normal distribution, we could use the method of maximum likelihood to estimate the parameters β and σ^2 . The likelihood function of the disturbances ε_i is given by

$$L(\sigma^2 \mid \varepsilon) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp(-\frac{1}{2\sigma^2}\varepsilon'\varepsilon), \tag{1.12}$$

or equivalently

$$L(\sigma^2, \beta \mid Y) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp(-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)), \qquad (1.13)$$

taking the ln of the likelihood function we have

$$\ln\ L(\sigma^2, \beta \mid Y) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(Y'Y - 2\beta'X'Y + \beta'X'X\beta). \quad (1.14)$$

The maximum likelihood estimates of σ^2 and β are given by the simultaneous solutions of the equations

$$\frac{\partial}{\partial\beta}[\ln\ L(\sigma^2, \beta \mid Y)] = -\frac{1}{2\sigma^2}(-2X'Y + 2X'X\beta) = 0, \qquad (1.15)$$

$$\frac{\partial}{\partial\sigma^2}[\ln\ L(\sigma^2, \beta \mid Y)] = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)'(Y - X\beta) = 0, \qquad (1.16)$$

giving the estimates

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad \text{and} \qquad \hat{\sigma^2} = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n}. \qquad (1.17)$$

That $\hat{\beta}$ and $\hat{\sigma^2}$ determine a maximum is confirmed by the Hessian matrix of the likelihood function evaluated at $\beta = \hat{\beta}$ and $\sigma^2 = \hat{\sigma^2}$ given by

$$H[ln\ L(\hat{\sigma^2}, \hat{\beta} \mid Y)] = \begin{pmatrix} \frac{-X'X}{\hat{\sigma^2}} & 0 \\ 0 & -\frac{n}{2\hat{\sigma^4}} \end{pmatrix}, \qquad (1.18)$$

which is negative definite. Thus when normality of the disturbance vector is assumed, the maximum likelihood estimator of β is identical to the least squares estimator b. The assumption of normality is required for tests of significance on the estimated parameters.

Sampling Properties Of The Least Squares Estimator b .

Since b is a linear function of the random vector Y , it is a random vector with expectation

$$\begin{aligned} E[b] &= E[(X'X)^{-1}X'Y] \\ &= E[(X'X)^{-1}X'(X\beta+\varepsilon)] \\ &= E[\beta+(X'X)^{-1}X'\varepsilon] \\ &= \beta+(X'X)^{-1}X'E[\varepsilon] \\ &= \beta. \end{aligned} \tag{1.19}$$

Thus b is an unbiased estimator of β .

The covariance matrix of b , denoted by $V(b)$, is given by

$$\begin{aligned} V[b] &= E[(b-E[b])(b-E[b])'] \\ &= E[(\ (X'X)^{-1}X'Y-\beta)(\ (X'X)^{-1}X'Y-\beta)'] \\ &= E[(\ \beta+(X'X)^{-1}X'\varepsilon-\beta)(\ \beta+(X'X)^{-1}X'\varepsilon-\beta)'] \\ &= E[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}] \\ &= (X'X)^{-1}X'E[\varepsilon\varepsilon']X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned} \tag{1.20}$$

Recall that an estimator θ, is called the best linear unbiased estimator of β, in short θ is BLUE of β, if θ is a linear unbiased estimator of β , having minimal variance in the class of linear unbiased estimators of β . With respect to the least squares estimator b, we have the following theorem.

Gauss-Markov Theorem: Given model (1.2), the least squares estimator $b=(X'X)^{-1}X'Y$, is the BLUE of β .

To prove the theorem, we have only to consider a linear unbiased estimator of β , say

$$\hat{b} = ((X'X)^{-1}X' + A)Y, \tag{1.21}$$

where A is any $([p+1] \times n)$ non-stochastic matrix. Thus we have

$$\begin{aligned} E[\hat{b}] &= ((X'X)^{-1}X' + A)E[Y] \\ &= ((X'X)^{-1}X' + A)X\beta \\ &= \beta + AX\beta, \end{aligned} \tag{1.22}$$

and since $\hat{b}$ is unbiased, we must have $AX\beta = 0$ for all β, implying $AX = 0$. Hence

$$\begin{aligned} (\hat{b} - \beta) &= ((X'X)^{-1}X' + A)Y - \beta \\ &= ((X'X)^{-1}X' + A)(X\beta + \varepsilon) - \beta \\ &= \beta + (X'X)^{-1}X'\varepsilon + AX\beta + A\varepsilon - \beta \\ &= ((X'X)^{-1}X' + A)\varepsilon. \end{aligned} \tag{1.23}$$

The covariance matrix of $\hat{b}$ is thus given as

$$\begin{aligned} E[(\hat{b} - \beta)(\hat{b} - \beta)'] &= E[((X'X)^{-1}X' + A)\varepsilon\varepsilon'((X'X)^{-1}X' + A)' \\ &= \sigma^2[(X'X)^{-1} + (X'X)^{-1}X'A' + AX(X'X)^{-1} + AA'] \\ &= \sigma^2(X'X)^{-1} + \sigma^2 AA'. \end{aligned} \tag{1.24}$$

But AA' is a positive semi-definite matrix, which shows that the covariance matrix of $\hat{b}$ is the covariance matrix of b plus a positive semi-definite matrix. Thus if $\hat{b}$ has minimum variance, then $A = 0$ and $\hat{b} = b$. Hence b has the minimum variance of all linear unbiased estimators of β .

Coefficient Of Determination.

A useful measure of how well the fitted values $\hat{y}_i$ correspond to the observed values y_i is the sample multiple correlation coefficient R given by

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{[\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2]^{\frac{1}{2}}}. \tag{1.25}$$

A useful theorem which relates the total sum of squares about the mean (SSTO) to the sum of squared error (SSE) and the sum of squares due to the regression (SSR) is

Theorem 1.1 :

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2. \tag{1.26}$$

Proof: The equation is often denoted by

$$SSTO = SSE + SSR. \tag{1.27}$$

Considering

$$\hat{Y} = PY \text{ where } P = X(X'X)^{-1}X', \tag{1.28}$$

we have

$$\hat{Y}'\hat{Y} = Y'P^2Y = Y'PY = Y'\hat{Y}. \tag{1.29}$$

differentiating $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})^2$ with respect to β_0 we have one of the normal equations for b, namely

$$-2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip}) = 0 \tag{1.30}$$

or

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0. \tag{1.31}$$

Thus

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \tag{1.32}$$

since

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i \qquad by\ (1.31)\\ &= 0 \qquad\qquad by\ (1.29).\end{aligned} \tag{1.33}$$

Another useful measure of how well estimated observations fit the actual observations is the coefficient of multiple determination (R^2) which has the following identity

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SSTO}. \tag{1.34}$$

Proof: From (1.31) $\bar{y} = \bar{\hat{y}}$ so that

$$\begin{aligned}&\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\\ &= \sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})\\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})(\hat{y}_i - \bar{y})\\ &= 0 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.\end{aligned} \tag{1.35}$$

Hence

$$\begin{aligned}R &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{[\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2]^{\frac{1}{2}}}\\ &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{[\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2]^{\frac{1}{2}}},\end{aligned} \tag{1.36}$$

and it follows that

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{1.37}$$

Centering and Scaling.

Without loss of generality, we will assume from this point that the X and Y matrices are centered and scaled to unit length so that the matrices $X'X$ and $X'Y$ are in correlation form. The coefficients of the centered and scaled data are often called beta coefficients. Denoting the least squares estimator of the raw data by b^r and the centered and scaled data by b^s , we may always return to our original coefficients by using the identities

$$b_j^r = b_j^s\left(\frac{S_y}{S_{Xj}}\right) \tag{1.38}$$

and

$$b_0^r = \bar{y} - \sum_{j=1}^{p} b_j^r \bar{x}_j, \tag{1.39}$$

where

$$S_y^2 = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}\ ,\ S_{Xj}^2 = \sum_{i=1}^{n} \frac{(X_{i,j} - \bar{X}_j)^2}{n-1}, \tag{1.40}$$

and X_j denotes the $j-th$ column vector of X . From this point on we will drop the superscript s and be using centered and scaled data along with the corresponding beta coefficients denoted by b .

Mean Squared Error.

A measure of how close the estimator b is to β called the mean squared error, is the average squared distance from b to β, defined by

$$\text{MSE } b = E[(b-\beta)'(b-\beta)]. \tag{1.41}$$

An estimator with a low MSE will be close to the true parameter vector β .

The MSE of b is

$$\begin{aligned}
\text{MSE } b &= E[(b-\beta)'(b-\beta)] \\
&= E[(\beta+(X'X)^{-1}X'\varepsilon-\beta)'(\beta+(X'X)^{-1}X'\varepsilon-\beta)] \\
&= E[\varepsilon'X(X'X)^{-1}(X'X)^{-1}X'\varepsilon] \\
&= \sigma^2\text{Tr}[X(X'X)^{-1}(X'X)^{-1}X'] \\
&= \sigma^2\text{Tr}[(X'X)^{-1}],
\end{aligned} \tag{1.42}$$

and since $(X'X)$ is symmetric positive definite, we have $(X'X)^{-1} = G\Lambda^{-1}G'$ where

$$\Lambda = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_p) \tag{1.43}$$

are the eigenvalues of $X'X$ and G is the ($p \times p$) matrix of orthonormal eigenvectors such that $G'G = GG' = I$ and $X'X = G\Lambda G'$. Thus we have

$$\begin{aligned}
\sigma^2\text{Tr}\,(X'X)^{-1} &= \sigma^2\text{Tr}(G\Lambda^{-1}G') \\
&= \sigma^2\sum_{i=1}^{p}\lambda_i^{-1}\,.
\end{aligned} \tag{1.44}$$

And since

$$\begin{aligned}
E[(b-\beta)'(b-\beta)] &= E[b'b-2b'\beta+\beta'\beta] \\
&= E[b'b]-\beta'\beta \\
&= \sigma^2\sum_{i=1}^{p}\lambda_i^{-1},
\end{aligned} \tag{1.45}$$

we also have that

$$E[b'b] = \beta'\beta + \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1}. \tag{1.46}$$

If we assume the disturbances are normaly distributed we also have [Seber 1977 p.16]

$$\begin{aligned} \text{Var}\,[(b-\beta)'(b-\beta)] &= 2\sigma^4 \text{Tr}[X(X'X)^{-2}X']^2 \\ &= 2\sigma^4 \text{Tr}(X'X)^{-2} \\ &= 2\sigma^4 \sum_{i=1}^{p} \lambda_i^{-2}. \end{aligned} \tag{1.47}$$

Problem Of Ill-Conditioned Data.

In the sequel, we will consider the eigenvalues of $X'X$ in the following order

$$\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p = \lambda_{min} > 0 \ .$$

Now if $X'X$, in the form of a correlation matrix, is the identity matrix, the least squares estimator b is a good one in terms of MSE . This is the ideal situation, where the columns of X are perfectly orthogonal, and from (1.46) we have

$$E(b'b) = \beta'\beta + \sigma^2 p \tag{1.48}$$

However problems arise if the $X'X$ matrix is "ill-conditioned", where near multicollinearities between the regressors in the X matrix will cause some of the eigenvalues of Λ to be very close to zero and very small compared to the largest eigenvalue. One measure of ill-conditioning is given by the condition number (ϕ) of the $X'X$ matrix defined by

$$\phi = \frac{\lambda_{max}}{\lambda_{min}}. \tag{1.49}$$

The usual rule of thumb is that condition numbers less than one hundred correspond to weak dependencies and numbers greater than nine hundred are associated with moderate to strong dependencies [Belsley, D.A., Kuh, E., Welsh, R.E. 1980, p. 104].

One consequence of this near-multicollinearity from (1.46), will be that the expected squared distance from b to β will be very large, making the least squares estimates too large in absolute value. Another consequence from (1.44), will be inflated variances of the estimates often accompanied by incorrect signs. To illustrate, consider the two parameter case. The least squares normal equations are

$$X'Xb = X'Y \tag{1.50}$$

$$\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} r_{1y} \\ r_{2y} \end{pmatrix}, \tag{1.51}$$

where

$$r_{12} = \frac{\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{(\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)^2 \sum_{i=1}^{n}(X_{i2} - \bar{X}_2)^2)^{\frac{1}{2}}}. \tag{1.52}$$

Now

$$(X'X)^{-1} = \begin{pmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{pmatrix}. \tag{1.53}$$

Thus as X_1 and X_2 become more collinear $|\, r_{12} \,| \rightarrow 1$, and

$$Var(b_i) = \sigma^2 \frac{1}{1-r_{12}^2} \rightarrow \infty, \text{ and MSE } b = \sigma^2 \left(\frac{1}{1+r_{12}^2} + \frac{1}{1-r_{12}^2}\right) \rightarrow \infty. \tag{1.54}$$

The least squares estimates are

$$\begin{aligned} b_1 &= \frac{r_{1y} - r_{2y}r_{12}}{1-r_{12}^2} \\ b_2 &= \frac{-r_{1y}r_{12} + r_{2y}}{1-r_{12}^2}. \end{aligned} \tag{1.55}$$

Thus the coefficients become very large in absolute value, may even change sign, and have inflated variances as $|r_{12}| \to 1$.

In general for the p variate case, the diagonal elements of the $(X'X)^{-1}$ matrix are the variance inflation factors (VIF) where

$$VIF_j = (X'X)^{-1}_{jj} = \frac{1}{1-r_j^2}, \quad \text{for } j = 1,\ldots,p, \tag{1.56}$$

where r_j^2 is the is the coefficient of multiple determination from regressing X_j on all other remaining regressor variables.

Proof: Let X_j be the j-th column of X and X_o be the X matrix with the j-th column omitted. Then from [Searle 1971 p.27] we have

$$X'X = \begin{pmatrix} X_o'X_o & X_o'X_j \\ X_j'X_o & X_j'X_j \end{pmatrix} = \begin{pmatrix} A & B \\ B' & D \end{pmatrix} \tag{1.57}$$

$$(X'X)^{-1} = \begin{pmatrix} A^{-1} + A^{-1}BWB'A^{-1} & -A^{-1}BW \\ -WB'A^{-1} & W \end{pmatrix} \tag{1.58}$$

where

$$W = (D - B'A^{-1}B)^{-1}. \tag{1.59}$$

Considering only W we have

$$\begin{aligned} W &= (X_j'X_j - X_j'X_o(X_o'X_o)^{-1}X_o'X_j)^{-1} \\ &= (X_j'[I - X_o(X_o'X_o)^{-1}X_o']X_j)^{-1} \\ &= (X_j'[I - X_o(X_o'X_o)^{-1}X_o'][I - X_o(X_o'X_o)^{-1}X_o']X_j)^{-1} \\ &= ([X_j - X_o(X_o'X_o)^{-1}X_o'X_j]'[X_j - X_o(X_o'X_o)^{-1}X_o'X_j])^{-1} \\ &= ([X_j - X_o\hat{b}]'[X_j - X_o\hat{b}])^{-1}, \end{aligned} \tag{1.60}$$

giving the SSE when X_j is regressed on X_o, and $\hat{b}$ is the corresponding vector of parameter estimates for this artificial regression. Assuming X has been centered and scaled and using

$$\begin{aligned} 1 - R_j^2 &= \frac{\sum_{i=1}^{n}(x_j - \hat{x}_j)^2}{\sum_{i=1}^{n}(x_j - \bar{x}_j)^2} \\ &= \frac{SSE_j}{SSTO_j}, \end{aligned} \tag{1.61}$$

from analogy with Theorem 1.1 we have the desired result.

Here r_j^2 is the coefficient of multiple determination from regressing X_j on all other remaining regressor variables. Thus the higher the multiple correlation in this artificial regression, the lower the precision in the estimate of the coefficient b_i. The general rule of thumb is that if any VIF exceeds five [Gunst, Mason, 1980 p.295], one should be concerned that the i-th regressor variable has a strong linear association with the remaining regressor variables.

Thus the least squares estimator, when the X matrix has strong collinearities, will give unstable parameter estimates. This is a serious problem when the objective of the regression analysis is parameter estimation, model specification, or prediction beyond the range of the data base [Gunst, Mason, 1980 p. 311]. In all the above cases, the true theoretical model must be correctly specified and correct parameter estimates are required. In these cases, the option to drop the variables causing the collinearity is not available.

Ridge Regression.

To control the problem of inflated variances and unstable parameter estimates

associated with least squares estimates applied to ill-conditioned data, Hoerl and Kennard (1970 a,b) suggested the ridge estimator, which is a biased estimator, given by

$$b(k) = (X'X + kI)^{-1}X'Y, \quad k \geq 0 . \tag{1.62}$$

Their idea was to trade a small bias for a large reduction in variance to get estimators that would be more stable, have correct signs, and reduced variances in spite of ill-conditioned data. The ridge estimator had the additional advantage that it could portray the sensitivity of the estimates to the ill-conditioning of the data with the ridge trace. The ridge trace is a plot of the ridge coefficients with respect to increases in the biasing parameter k . A typical ridge trace will show the least squares estimates at $k = 0$, and for positive values of k around zero, the ridge coefficients change rapidly, reflecting the coefficient instability due to multicollinearity. However as k increases, variances reduce, and the coefficients become stable. Hoerl and Kennard suggested selecting the ridge parameter k , corresponding to the point where the coefficients begin to stabilize in the ridge trace.

Example Of Ordinary Ridge Regression.

As an example of Ridge regression applied to ill-conditioned data we use a subset of data from Theil (1971) p.456. The raw data is given in following Table 1. The least squares estimates for the raw data are

$$Y_i = 12.4 - 0.17X_{i1} + 1.12X_{i2} \qquad R^2 = .976 \tag{1.63}$$

TABLE 1: Raw Data.

Index	Aggregate Consumption	Aggregate Income From Profits	Aggregate Income From Wages	Year
i	Y_i	X_{i1}	X_{i2}	
1	41.9	12.4	28.2	1921
2	45.0	16.9	32.2	1922
3	59.2	18.4	37.0	1922
4	50.6	19.4	37.0	1923
5	52.6	20.1	38.6	1924
6	55.1	19.6	40.7	1925
7	56.2	19.8	41.5	1926
8	57.3	21.1	42.9	1927
9	57.8	21.7	45.3	1928
Means	51.74	18.82	38.15	
Standard Deviations	5.59	2.79	5.35	

These estimates are unacceptable as they imply the bankruptcy of the wage earner. In particular, if income (b_1) decreases by one dollar, consumption increases by 17 cents. Also, when wage income increases by one dollar, consumption increases by 1.12 dollars.

TABLE 2: Standardized data.

Y_i	X_{i1}	X_{i2}
-0.62	-0.82	-0.66
-0.43	-0.24	-0.39
-0.16	-0.05	-0.08
-0.07	-0.07	-0.08
-0.05	0.16	-0.03
0.21	0.09	0.17
0.28	0.12	0.22
0.35	0.29	0.31
0.38	0.37	0.47

The standardized least squares beta coefficients are given as

$$Y_i = -0.09X_{i1} + 1.07X_{i2} \qquad R^2 = .976 \tag{1.64}$$

with

$r_{12} =$ 0.94
SSE= 0.024
SSR= 0.976
$\lambda_1 =$ 1.94
$\lambda_2 =$ 0.06
$\phi =$ 34.11
$VIF_1 =$ 9.03
$VIF_2 =$ 9.03

TABLE 3: Ridge coefficients.

k	b_1	b_2	SSE
.00	-.09	1.07	.024
.01	-.00	0.98	.025
.05	0.17	0.78	.036
.08	0.23	0.71	.042
.12	0.28	0.65	.049
.20	0.32	0.57	.060
.50	0.33	0.45	.099
.70	0.32	0.41	.127
1.0	0.29	0.36	.171

Although the condition number $\phi = 34.11$ does not indicate any strong dependencies in the $X'X$ matrix, we have from $r_{12} = .94$ a high correlation between X_1 and X_2 making our VIF's exceed the recommended guidelines.

The biasing parameter $k = 0.08$ was chosen from table 3 corresponding to the point where the coefficients begin to stabilize without severely increasing the sum of squared error SSE. The standardized ridge regression estimates are given as

$$Y_i = .23X_{i1} + .71X_{i2} \qquad R^2 = .958 \tag{1.65}$$

and using the identities (1.38) and (1.39), we have the ridge estimates at $k =$ .08 , for the raw data as

$$Y_i = 14.85 + .46X_{i1} + .74X_{i2} \tag{1.66}$$

These estimates are now acceptable since for an increase in income, consumption increases by .46 cents, and for an increase in wages by one dollar, consumption increases by .74 cents.

Justification Of Ridge Regression.

The main theoretical justification for ridge regression is the Hoerl and Kennard (1970a) Existence Theorem which states that there exists a $k > 0$ such that

$$\text{MSE } b(k) < \text{MSE } b. \tag{1.67}$$

In general, the main objective in ridge regression is to obtain point estimates of parameters that have a smaller mean square error than the least square estimates, have reduced variances, and have correct signs and magnitudes. A constrained least squares interpretation of the ridge estimator was given by Hoerl and Kennard (1970a). That is, $b(k)$ minimizes the residual sum of squares subject to a constraint on the length of the estimated coefficient vector. A Baysian interpretation of ridge regression was given by Lindley and Smith (1972). If

$$(Y \mid \beta) \sim N(X\beta, \sigma^2 I) \text{ and } \beta \sim N(0, \sigma_\beta^2 I)\,, \tag{1.68}$$

then $b(k)$ is the bayes estimator where $k = \sigma^2/\sigma_\beta^2$.

These interpretations however do not explicitly define an appropriate k value to use in a specific application. The acceptable range of k values, where the ridge

estimator dominates the least square estimator in mean square error depends on the unknowns β and σ^2. The Bayes interpretation of ridge regression yields a k value that is the ratio of two unknown variances. The constained least squares approach does not define a specific k value in practice, since an explicit constraint on the length of the coefficient vector is unknown in most applications. As a result, several algorithms have been proposed, using the data to select the biasing parameter k. Since the improvement of $b(k)$ over b is dependent on unknown parameters, Monte Carlo experiments are conducted to compare the ridge and least square coefficients in terms of mean square error. Many independent simulations have compared least squares and ridge type estimators. In particular Hoerl,Kennard and Baldwin (1975), Dempster,Schatzoff, and Wermuth (1977), Lawless and Wang (1976),Gibbons (1981) and Gunst and Mason (1978) concluded that ridge type estimators are superior to least squares in the face of ill-conditioned data. However, there is wide spread disagreement about the optimum ridge estimator. The difficulty here is that no one rule is superior under all conditions and the scope of the results is limited to the experimental designs and parameter values considered in the simulation. In spite of these difficulties, ridge regression has become a popular technique for problems where the data is ill-conditioned and accurate parameter estimates are required. In particular, some recent applications of ridge regression have been in criminology (Liu and Bee 1984), economics (Gapinski 1984), mortality estimates (Lawrence, Marsh 1984), robust regression (Askin, Montgomery 1980), subset selection techniques (Hoerl,Schunemeyer, Hoerl 1986), and principle component regression (Baye, Parker 1984).

Chapter 2

ORDINARY RIDGE ESTIMATOR

Ordinary ridge regression, as introduced by Hoerl and Kennard (1970 a), amounts to adding a biasing constant k $(0 \leq k < \infty)$ to the diagonal of the $X'X$ matrix before inverting it for least squares estimation. The ordinary ridge estimator is simple and designed to handle the problem of near multicollinearity between the regressors in the X matrix. The ordinary ridge estimator, denoted by $b(k)$, is defined by the following equation

$$b(k) = (X'X + kI)^{-1}X'Y, \tag{2.1}$$

for $k \geq 0$ and has the following properties.

1) $b(k)$ is a linear transform of the ordinary least squares estimator $b = (X'X)^{-1}X'Y$ where $b(0) = b$ and $\lim_{k\to\infty} b(k) = 0$.

2) $b(k)$ is a biased estimator depending on the unknown parameter vector β. Explicitly, the bias of $b(k)$ is given as

$$\text{BIAS } b(k) = [Eb(k) - \beta] = -k(X'X + kI)^{-1}\beta. \tag{2.2}$$

3) The mean squared error of $b(k)$ a measure of how close $b(k)$ is to the true parameter vector β, denoted by $\text{MSE } b(k)$ is given by

$$\text{MSE } b(k) = E(b(k)-\beta)'(b(k)-\beta) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2}. \tag{2.3}$$

The first term is the sum of the variances of $b(k)$ and is a monotonically decreasing function of k. The second term is the total squared bias of $b(k)$ and is a monotonically increasing function of k.

4) MSE $b(0)$ = MSE b, and $\lim_{k\to\infty}$ MSE $b(k) = \beta'\beta$.

5) There exists a $k > 0$ such that MSE $b(k) <$ MSE b.

6) For $k > 0, b(k)'b(k) < b'b$ making the squared length of $b(k)$ less than b for $k > 0$.

7) The sum of the squared residuals, given by

$$\phi(k) = (Y - Xb(k))'(Y - Xb(k)) \tag{2.4}$$

is an increasing function of k.

8) $b(k)$ minimizes the sum of squared residuals on the sphere centered at the origin whose squared radius is $b(k)'b(k)$.

Proofs:

Property 1:

$$\begin{aligned} b(k) &= (X'X + kI)^{-1}X'Y \\ &= (X'X + kI)^{-1}X'X(X'X)^{-1}X'Y \\ &= (X'X + kI)^{-1}X'Xb \\ &= ((X'X)^{-1}(X'X + kI))^{-1}b \\ &= (I + k(X'X)^{-1})^{-1}b. \end{aligned} \tag{2.5}$$

Thus

$$b(0) = (I + 0I)^{-1}b = b. \tag{2.6}$$

To evaluate $\lim_{k\to\infty} b(k)$, we use the following identity

$$(I + k(X'X)^{-1})^{-1} = I - k(X'X + kI)^{-1}, \tag{2.7}$$

which can be verified by premultiplying both sides by $(I + k(X'X)^{-1})$. Thus,

$$\begin{aligned} b(k) &= (I + k(X'X)^{-1})^{-1}b \\ &= [I - k(X'X + kI)^{-1}]b \\ &= b - (k^{-1}X'X + I)^{-1}b, \end{aligned} \tag{2.8}$$

and taking the limit we have

$$\begin{aligned} \lim_{k\to\infty} b(k) &= \lim_{k\to\infty} [b - (k^{-1}X'X + I)^{-1})b] \\ &= b - (0X'X + I)^{-1}b \\ &= 0. \end{aligned} \tag{2.9}$$

Hence, the ordinary ridge estimator $b(k)$ shrinks the least squares estimator b to the null vector as $k \to \infty$.

Property 2:

$$\begin{aligned} \text{BIAS } b(k) &= E(b(k)) - \beta \\ &= E(I + k(X'X)^{-1})^{-1}b - \beta \\ &= E(I - k(X'X + kI)^{-1}b - \beta \\ &= (I - k(X'X + kI)^{-1})E(b) - \beta \\ &= (I - k(X'X + kI)^{-1})\beta - \beta \\ &= -k(X'X + kI)^{-1}\beta. \end{aligned} \tag{2.10}$$

Property 3:

$$\begin{aligned}\text{MSE } b(k) &= E[(b(k)-\beta)'(b(k)-\beta)] \\ &= E[(b(k)-E(b(k))+E(b(k))-\beta)'(b(k)-E(b(k))+E(b(k))-\beta)] \\ &= E[\,(b(k)-E(b(k))\,)'(b(k)-E(b(k))\,)] \\ &\quad +(E(b(k))-\beta)'(E(b(k))-\beta),\end{aligned} \tag{2.11}$$

where the cross product term is zero since

$$E(b(k)-E(b(k))\,)=0. \tag{2.12}$$

Now the first term is the total variance of $b(k)$ where

$$\begin{aligned}[b(k)-E(b(k))] &= (I+k(X'X)^{-1})^{-1}b-(I+k(X'X)^{-1})^{-1}\beta \\ &= (I+k(X'X)^{-1})^{-1}(b-\beta) \\ &= (I+k(X'X)^{-1})^{-1}((X'X)^{-1}X'Y-\beta) \\ &= (I+k(X'X)^{-1})^{-1}((X'X)^{-1}X'(X\beta+\varepsilon)-\beta) \\ &= (I+k(X'X)^{-1})^{-1}(\beta+(X'X)^{-1}X'\varepsilon-\beta) \\ &= (I+k(X'X)^{-1})^{-1}(X'X)^{-1}X'\varepsilon \\ &= (X'X(I+k(X'X)^{-1}))^{-1}X'\varepsilon \\ &= (X'X+kI)^{-1}X'\varepsilon.\end{aligned} \tag{2.13}$$

So for the first term we have

$$\begin{aligned}&E[(b(k)-E(b(k))\,)'(b(k)-E(b(k))\,)] \\ =&E(\varepsilon'X(X'X+kI)^{-1}(X'X+kI)^{-1}X'\varepsilon),\end{aligned} \tag{2.14}$$

the expected value of a quadratic form in ε. Using Theorem A3 from Appendix A

we have

$$\begin{aligned}
&E(\epsilon' X(X'X+kI)^{-1}(X'X+kI)^{-1}X'\epsilon)\\
&=\mathrm{Tr}(X(X'X+kI)^{-1}(X'X+kI)^{-1}X'\ \mathrm{Var}(\epsilon))\\
&=\sigma^2\ \mathrm{Tr}[(X'X+kI)^{-1}(X'X+kI)^{-1}X'X]\\
&=\sigma^2\ \mathrm{Tr}[G(\Lambda+kI)^{-1}G'G(\Lambda+kI)^{-1}G'G\Lambda G']\\
&=\sigma^2\ \mathrm{Tr}[(\Lambda+kI)^{-1}(\Lambda+kI)^{-1}\Lambda]\\
&=\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^2}.
\end{aligned} \tag{2.15}$$

Here G is the orthonormal matrix of eigenvectors which diagonalizes $X'X$, such that $G'G=I$ and $G\Lambda G'=X'X$ where $\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_p)$ is the diagonal matrix of eigenvalues of $X'X$. For the squared bias term, letting $\alpha=G'\beta$, we have

$$\begin{aligned}
&(E(b(k))-\beta)'(E(b(k))-\beta)\\
&=(-k(X'X+kI)^{-1}\beta)'(-k(X'X+kI)^{-1}\beta)\\
&=k^2\beta'(X'X+kI)^{-1}(X'X+kI)^{-1}\beta\\
&=k^2\beta'G(\Lambda+kI)^{-1}G'G(\Lambda+kI)^{-1}G'\beta\\
&=k^2\alpha'(\Lambda+kI)^{-2}\alpha\\
&=k^2\sum_{i=1}^{p}\frac{\alpha_i^2}{(\lambda_i+k)^2}.
\end{aligned} \tag{2.16}$$

Thus combining our variance and bias terms we have

$$\begin{aligned}
\mathrm{MSE}\ b(k)&=E(b(k)-\beta)'(b(k)-\beta)\\
&=\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^2}+k^2\sum_{i=1}^{p}\frac{\alpha_i^2}{(\lambda_i+k)^2}.
\end{aligned} \tag{2.17}$$

Considering the total variance term we have

$$\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^2},\quad \text{for } k\geq 0. \tag{2.18}$$

Since $\lambda_i > 0$ for all i, each element $(\lambda_i + k)$ is positive and there are no singularities in the sum, and for $k = 0$ we have $\sigma^2 \sum_{i=1}^{p} \lambda_i^{-1}$. Thus the first term is continuous. Considering

$$\frac{d}{dk} \frac{\sigma^2 \sum_{i=1}^{p} \lambda_i}{(\lambda_i + k)^2} = -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^3} < 0 \tag{2.19}$$

we conclude that the total variance is monotone decreasing.

Now the squared bias term

$$k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2} \tag{2.20}$$

is also a sum of rational functions, none of which having any singularity points, and at $k = 0$ the squared bias is zero. Hence the second term is also continuous. For $k > 0$ we have

$$k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2} = \sum_{i=1}^{p} \frac{\alpha_i^2}{(\frac{\lambda_i}{k} + 1)^2}, \tag{2.21}$$

making each term monotonically increasing. Thus the squared bias, being a sum of monotonically increasing functions, is monotonically increasing.

Property 4:

$$\text{MSE } b(k) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2}. \tag{2.22}$$

At $k = 0$ we have

$$\text{MSE } b(0) = \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} = \text{MSE } (b). \tag{2.23}$$

Now as $k \to \infty$ we have

$$\begin{aligned}\lim_{k\to\infty} \text{MSE } b(k) &= \lim_{k\to\infty} \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + \lim_{k\to\infty} \sum_{i=1}^{p} \frac{\alpha_i^2}{(\frac{\lambda_i}{k} + 1)^2} \\ &= 0 + \sum_{i=1}^{p} \alpha_i^2 \\ &= \alpha'\alpha \\ &= \alpha' G G' \alpha \\ &= \beta'\beta. \end{aligned} \tag{2.24}$$

Property 5: Considering the MSE $b(k) = \text{MSE } b$, at $k = 0$, and for $k > 0$, the two terms in MSE $b(k)$ are continuous decreasing and increasing functions respectively. Consequently, if we can show that their always exists a $k > 0$ such that the $\frac{d}{dk}\text{MSE } b(k) < 0$ then the theorem is proved. We require a $k > 0$ such that

$$\frac{d}{dk}\text{MSE } b(k) = 2 \sum_{i=1}^{p} \frac{\lambda_i (k\alpha_i^2 - \sigma^2)}{(\lambda_i + k)^3} < 0 \tag{2.25}$$

Now if $(k\alpha_i^2 - \sigma^2) < 0$ for all i then $k < \frac{\sigma^2}{\alpha_i^2}$ for all i. choosing

$$0 < k < \frac{\sigma^2}{\alpha_{\text{max}}^2} \leq \frac{\sigma^2}{\alpha_i^2} \leq \frac{\sigma^2}{\alpha_{\text{min}}^2} \tag{2.26}$$

we have a positive $k < \frac{\sigma^2}{\alpha_{\text{max}}^2}$ such that

$$\text{MSE } b(k) < \text{MSE } b. \tag{2.27}$$

Property 6: For $k > 0$ we have

$$\begin{aligned} b(k)'b(k) &= b'(I + k(X'X)^{-1})^{-1}(I + k(X'X)^{-1})^{-1}b \\ &= b'G(I + k\Lambda^{-1})^{-1}G'G(I + k\Lambda^{-1})^{-1}G'b \\ &= \alpha'[(\Lambda + kI)\Lambda^{-1}]^{-1}[(\Lambda + kI)\Lambda^{-1}]^{-1}\alpha \\ &= \sum_{i=1}^{p} \alpha_i^2 \frac{\lambda_i^2}{(\lambda_i + k)^2} \\ &< \sum_{i=1}^{p} \alpha_i^2 = \alpha'\alpha = b'GG'b = b'b. \end{aligned} \tag{2.28}$$

Property 7:

$$\begin{aligned} \phi(k) &= (Y - Xb(k))'(Y - Xb(k)) \\ &= (Y - Xb + Xb - Xb(k))'(Y - Xb + Xb - Xb(k)) \\ &= (Y - Xb)'(Y - Xb) + 2(Y - Xb)'X(b - b(k)) + (b - b(k))'X'X(b - b(k)) \\ &= (Y - Xb)'(Y - Xb) + 2Y'(I - X(X'X)^{-1}X')X(b - b(k)) \\ &\quad + (b - b(k))'X'X(b - b(k)) \\ &= (Y - Xb)'(Y - Xb) + 2Y'(X - X)(b - b(k)) + (b - b(k))'X'X(b - b(k)) \\ &= (Y - Xb)'(Y - Xb) + (b - b(k))'X'X(b - b(k)) \\ &= \phi_{(0)} + k^2 b'(X'X + kI)^{-1}X'X(X'X + kI)^{-1}b \\ &= \phi_{(0)} + k^2 b'G(\Lambda + kI)^{-1}G'G\Lambda G'G(\Lambda + kI)^{-1}G'b \\ &= \phi_{(0)} + k^2 \alpha'(\Lambda + kI)^{-2}\lambda\alpha \\ &= \phi_{(0)} + k^2 \sum_{i=1}^{p} \frac{\alpha_i^2 \lambda_i}{(\lambda_i + k)^2} \end{aligned} \tag{2.29}$$

where the second term is a monotonically increasing function of k.

Property 8: Let B be any estimator of β. We minimize

$$F = (Y - XB)'(Y - XB), \quad \text{subject to } B'B = R^2. \tag{2.30}$$

As a Lagrangian problem this is to

$$\text{Minimize } F = (Y - XB)'(Y - XB) + k(B'B - R^2) \tag{2.31}$$

where k is the multiplier, giving

$$\begin{aligned} \frac{\partial F}{\partial B} &= \frac{\partial}{\partial B}[Y'Y - 2B'X'Y + B'X'XB + kB'B - kR^2] \\ &= -2X'Y + 2X'XB + 2kB = 0 \end{aligned} \tag{2.32}$$

when

$$(X'X + kI)B = X'Y \tag{2.33}$$

with solution

$$B = b(k) = (X'X + kI)^{-1}X'Y. \tag{2.34}$$

Since

$$\frac{\partial^2 F}{\partial B^2} = 2(X'X + kI) \text{ is p.d. for } k \geq 0 \tag{2.35}$$

we have the minimum at

$$b(k) = (X'X + kI)X'Y. \tag{2.36}$$

Property 5 is the Existence Theorem and uses the MSE criterion. In general, given two competing estimators of a parameter vector ω say W_1, W_2, if the

$$\text{MSE } W_1 < \text{MSE } W_2 \tag{2.37}$$

then W_1 is the preferred estimator. A stronger criterion is based on the matrix mean squared error, denoted by MTxMSE, which is defined as

$$\text{MTxMSE } b = E(b - \beta)(b - \beta)' \tag{2.38}$$

Say b_1 and b_2 are two competing estimators of the parameter vector β. If

$$\text{MTxMSE } b_2 - \text{MTxMSE } b_1 = \Pi \text{ a p.d. matrix,} \tag{2.39}$$

then b_1 is the preferred estimator. Letting Tr denote the trace operator, we note that $\text{Tr}(\text{MTxMSE } b_1) = \text{MSE } b_1$. Using the stronger criterion of MTxMSE, Theobald(1974) gave the following theorem

$$\text{MTxMSE } b - \text{MTxMSE } b(k) = \Pi \text{ a p.d. matrix} \tag{2.40}$$

if

$$0 < k \leq \frac{2\sigma^2}{\beta'\beta} \tag{2.41}$$

Proof: Consider

$$\begin{aligned}
&\text{WMSE } (b) - \text{WMSE } (b(k)) \\
&= E(b-\beta)'W(b-\beta) - E(b(k)-\beta)'W(b(k)-\beta) \\
&= \text{Tr}[E(b-\beta)'W(b-\beta) - E(b(k)-\beta)'W(b(k)-\beta)] \\
&= \text{Tr}W[E(b-\beta)(b-\beta)' - E(b(k)-\beta)(b(k)-\beta)'] \\
&= \text{Tr}W[\text{MTxMSE } (b) - \text{MTxMSE } (b(k))] \\
&= \text{Tr}W\Pi,
\end{aligned} \tag{2.42}$$

where by Theorem A2 in Appendix A,

$$\text{Tr}W\Delta \geq 0 \text{ for a p.s.d. matrix } W \text{ iff } \Delta \text{ is a p.s.d. matrix.} \tag{2.43}$$

Using

$$\text{MTxMSE } (b(k)) = E[(b(k)-\beta)(b(k)-\beta)'], \tag{2.44}$$

we have

$$\begin{aligned}\text{MTxMSE } &= E[(b(k) - E(b(k))\ )(b(k) - E(b(k))\ )'] \\ &\quad + (E(b(k)) - \beta)(E(b(k)) - \beta)' \\ &= \sigma^2(X'X + kI)^{-1}X'X(X'X + kI)^{-1} \\ &\quad + k^2(X'X + kI)^{-1}\beta\beta'(X'X + kI)^{-1},\end{aligned} \tag{2.45}$$

and

$$\text{MTxMSE } (b) = \sigma^2(X'X)^{-1}, \tag{2.46}$$

giving

$$\begin{aligned}\Pi &= \text{MTxMSE } (b) - \text{MTxMSE } (b(k)) \\ &= \sigma^2(X'X)^{-1} - [\sigma^2(X'X + kI)^{-1}X'X(X'X + kI)^{-1} \\ &\quad + k^2(X'X + kI)^{-1}\beta\beta'(X'X + kI)^{-1}] \\ &= (X'X + kI)^{-1}[\sigma^2(X'X + kI)(X'X)^{-1}(X'X + kI) \\ &\quad - \sigma^2X'X - k^2\beta\beta'](X'X + kI)^{-1} \\ &= (X'X + kI)^{-1}[Q](X'X + kI)^{-1},\end{aligned} \tag{2.47}$$

where

$$Q = [\sigma^2(X'X + kI)(X'X)^{-1}(X'X + kI) - \sigma^2X'X - k^2\beta\beta']. \tag{2.48}$$

Expanding the first term in Q we have

$$\begin{aligned}&\sigma^2(X'X + kI)(X'X)^{-1}(X'X + kI) \\ &= \sigma^2(I + k(X'X)^{-1})(X'X + kI) \\ &= \sigma^2(X'X + 2kI + k^2(X'X)^{-1})\end{aligned} \tag{2.49}$$

Now since $(X'X + kI)^{-1}$ is positive definite, an equivalent condition for Π to be positive definite is that Q be positive definite, where Q can now be written as

$$\begin{aligned}Q &= \sigma^2X'X + \sigma^2 2kI + \sigma^2k^2(X'X)^{-1} - \sigma^2X'X - k^2\beta\beta' \\ &= [\sigma^2k^2(X'X)^{-1} + \sigma^2 2k(I - \frac{k}{2\sigma^2}\beta\beta')\ ],\end{aligned} \tag{2.50}$$

and since $\sigma^2 k^2 (X'X)^{-1}$ is positive definite a sufficient condition is that

$$I - \frac{k}{2\sigma^2}\beta\beta' \text{ a p.s.d. matrix} \tag{2.51}$$

Using Theorem A1 from Appendix A, this is equivalent to

$$k \leq \frac{2\sigma^2}{\beta'\beta} \tag{2.52}$$

Theobald's result, for $\beta'\beta = \sum_{i=1}^{p} \alpha_{max}^2$ is $\frac{2}{p}$ times shorter than the Hoerl and Kennard acceptable $k < \frac{\sigma^2}{\alpha_{max}^2}$. This discrepancy is removed if we consider that the MTxMSE criterion is stronger than the MSE criterion. Further, the MTxMSE criterion is more general since it is also equivalent to the weighted mean square error criterion, denoted by WMSE , where

$$\text{WMSE } b = E(b-\beta)'W(b-\beta), \text{ for } W \text{ a p.s.d. matrix} \tag{2.53}$$

This equivalence is stated explicitly in the following theorem given by Theobald (1974),

Equivalence Theorem. Let two estimators b_1 and b_2 be given. The following two statements are then equivalent:

$$\text{MTxMSE } b_1 - \text{MTxMSE } b_2 = \Delta \quad \text{a p.s.d. matrix} \tag{2.54}$$

$$\text{WMSE } b_1 - \text{WMSE } b_2 \geq 0, \text{ for } W \text{ a p.s.d. matrix} \tag{2.55}$$

The condition that $0 < k \leq \frac{2\sigma^2}{\beta'\beta}$ is sufficient for superiority of $b(k)$ but not necessary. In practice, this condition may be too conservative. Swindel and

Chapman (1973) gave the following sufficient and necessary condition for the superiority of the ordinary ridge estimator $b(k)$ over least squares using the MTxMSE criterion.

Theorem 2.1:

$$\text{MTxMSE } b - \text{MTxMSE } b(k) = \Delta \text{ a p.d. matrix}, \tag{2.56}$$

iff

$$0 < k < \frac{2}{|\min(0, \xi)|}, \tag{2.57}$$

where ξ denotes the minimum eigenvalue of

$$(X'X)^{-1} - \frac{\beta\beta'}{\sigma^2}. \tag{2.58}$$

Considering that the minimum eigenvalue of $(X'X)^{-1}$ is given by λ_1^{-1} and the single non-zero eigenvalue of $\beta\beta'$ is given by $\beta'\beta$, we have

$$\xi = \lambda_1^{-1} - \frac{\beta'\beta}{\sigma^2}. \tag{2.59}$$

The ratio $\frac{\beta'\beta}{\sigma^2}$ is the signal to noise ratio (SNR). Thus if the SNR is sufficiently small to make ξ positive, any k in the interval $(0, +\infty)$ will make $b(k)$ superior to b under the MTxMSE criterion. Conversely, if the SNR is sufficiently large to make ξ negative, then a small biasing parameter k $(0 < k < \frac{2}{|\xi|})$ will be appropriate to make the ridge estimates $b(k)$ superior to b.

Chapter 3

OPERATIONAL ORDINARY RIDGE ESTIMATORS

Up to this point we have not considered how the biasing parameter k should be selected. Much of the controversy surrounding ridge regression revolves around this question. Part of the problem stems from the fact that the acceptable intervals for k, depend on unknown parameter values. We define an acceptable interval for k to be the interval where $b(k)$ dominates b under the MSE or MTxMSE criterion. Many researchers believe that it makes intuitive sense to use the least squares estimates of σ^2 and β to estimate the maximum acceptable k. However, this introduces the problem that if stochastic values are used to estimate k, the MSE or MTxMSE gains ,which assume a fixed k, are no longer guaranteed. In particular, any ridge estimator which depends on the random vector Y, will be a function of the sample data and thus be stochastic. It will not have the same properties as a ridge estimator based on fixed k. Nevertheless, many independent simulations have demonstrated the overall good performance of many different stochastic ridge estimators under the MSE criterion. With these criticisms in mind, we will review various proposals for selecting the biasing parameter k for use in the ordinary ridge estimator $b(k)$.

The Ridge Trace.

Introduced by Hoerl and Kennard (1970b), the ridge trace is one of the simplest and most widely used methods to select the k parameter for ordinary ridge re-

gression. The ridge trace is a two dimensional plot of the ridge coefficient estimates $b(k)_i$ on the vertical axis with k on the horizontal axis. One curve or trace is made for each coefficient. The plot may include the residual sum of squares $\phi(k)$ for corresponding k values. Hoerl and Kennard suggested the following procedure to select k from the ridge trace. Choose the smallest k where the coefficient magnitudes have reached their correct signs, are collectively stable, and the $\phi(k)$ has not increased significantly. Brown and Beattie (1975,p.27) give the following guideline to use the ridge trace. Select k where the last ridge coefficient attains its maximum value after having obtained its correct sign. Here correct sign will be the sign at $k = 0.9$ or the largest appropriate k for the problem in question. One of the criticisms of the ridge trace method is that k is selected from a visual inspection of the random beta coefficients, making k stochastic(Coniffe, Stone, 1973). Another problem is that the ridge trace will have a stability region even for perfectly orthogonal data. Finally, there is no guarantee that the k value selected will be in the acceptable interval. In answer to the first two of the above mentioned criticisms Vinod(1976a) suggested a m scale trace to be used in conjunction with an index of stability of relative magnitudes (ISRM).

The M Scale Trace and ISRM

Vinod (1976a) suggested a rescaling of the horizontal axis to overcome the problem of stability in the ridge trace even for perfectly orthogonal data, where $X'X = I$. The m scale trace is identical to the ridge trace except that the horizontal k axis is compressed to the interval $[0,p]$ where p is the number of

coefficients to be estimated. The rescaled horizontal axis is now the m axis, where

$$m = p - \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)} = p - \sum_{i=1}^{p} \delta_i \tag{3.1}$$

where $\delta_i = \frac{\lambda_i}{(\lambda_i + k)}$, is called the i-th shrinkage factor.

For the case where the $X'X$ matrix is perfectly orthogonal we have $X'X = I$ and $\lambda_i = 1$, for $i = 1, \ldots, p$. In this situation, there is no need for ridge regression as the least squares estimates b will be stable, have correct signs, and the problem of inflated variances will not exist. However, the ridge trace will be unreliable as it will still have a region of stability. To see this consider

$$b(k) = G\Delta G'b = G\text{diag}\left(\frac{\lambda_i}{\lambda_i + k}\right) G'b \tag{3.2}$$

taking the derivative with respect to k we have

$$\begin{aligned} \frac{db(k)}{dk} &= -G\text{diag}\left(\frac{\lambda_i}{(\lambda_i + k)^2}\right) G'b \\ &= -\frac{b}{\sigma^2}\text{Var}(b(k)) \end{aligned} \tag{3.3}$$

and in the orthogonal case we have

$$\begin{aligned} \frac{db(k)}{dk} &= -G\text{diag}\left(\frac{1}{(1 + k)^2}\right) G'b \\ &= -\frac{1}{(1 + k)^2} GIG'b \\ &= -\frac{b}{(1 + k)^2} \end{aligned} \tag{3.4}$$

making the absolute value in the change of $b(k)$ a decreasing function of k . On the other hand, the m scale trace does not have this property since

$$\frac{dm}{dk} = \frac{d}{dk}\left(p - \sum_{i=1}^{p} \delta_i\right) = \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} \tag{3.5}$$

and considering that

$$\frac{db(k)}{dm} = \frac{db(k)}{dk}\frac{dk}{dm} = -b\sigma^{-2}\mathrm{Var}(b(k))\,(\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^2})^{-1} \tag{3.6}$$

when $\lambda_i = 1,\ i = 1,\ldots,p$ we have

$$\frac{db(k)}{dm} = \frac{-b}{(1+k)^2}\frac{(1+k)^2}{p} = \frac{-b}{p} \tag{3.7}$$

showing that the rate of change of $b(k)$ is not a function of m . Thus in the case of orthogonal data the m scale trace will have p straight lines converging to 0 at $m = p$. Each line or trace will have intercept at the least squares estimates $b(0)_i = b_i$, and slope $-b_i p^{-1}$. In the usual case of ill-conditioned data, the m scale trace will show rapidly changing ridge coefficients around $m = 0$. However as m increases the lines will straighten out. The stable region of the m scale trace begins at the point when all the coefficients begin to imitate the straight line behavior of the perfectly orthogonal case. Thus we could select the smallest m or k at the point where the coefficients become straight lines converging to zero. In practice we would calculate the p shrinkage factors δ_i for a range of k values and thus calculate m . Hoerl and Kennard recommended that the ridge trace be truncated at $k = 1$. In the case of severly ill-conditioned data any activity outside this range could be missed. The finite range of the m scale trace would not have this problem. Another advantage is that it could be used to trace the generalized ridge estimates once a rule is given to determine the different k_i . Finally the m scale, where m stands for multicollinearity allowance, has the following interpretation. Suppose the eigenvalues of a two parameter model were $\lambda_1 = 10, \lambda_2 = .001$. Geometrically the data could be viewed as a very flat ellipse with most of the spread in the major axis and the least part in the minor axis.

From the relative smallness in λ_2 we could view the data as being roughly one-dimensional, allowing for the loss of $m = 1$ dimensions due to multicollinearity. Thus the multicollinearity allowance m gives the rank deficiency in the $X'X$ matrix.

Vinod's (1976a) index of stability of relative magnitudes (ISRM), is a non-stochastic measure of stability from which k or m can be selected without the use of any plot. Essentially it quantifies the concept of relative stability of coefficients. The index is non-stochastic as it depends only on the eigenvalues of the $X'X$ matrix. The ISRM is derived by considering the sum of squared differences between

$$\frac{-b}{p} \text{ and } \frac{-b\text{Var}(b(k))}{\sigma^2 \overline{s}}, \text{ from } \frac{db(k)}{dm}, \tag{3.8}$$

for the orthogonal and non-orthogonal cases respectively.

We define

$$\overline{s} = \sum_{i=1}^{p} \delta_i^2 \lambda_i^{-1} = \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2}. \tag{3.9}$$

The sum of squared differences is given as

$$\begin{aligned}
&\left(\frac{-b\text{Var}(b(k))}{\sigma^2\overline{s}} - \frac{-b}{p}\right)' \left(\frac{-b\text{Var}(b(k))}{\sigma^2\overline{s}} - \frac{-b}{p}\right) \\
&= b' \left(\frac{\text{Var}(b(k))}{\sigma^2\overline{s}} - \frac{I}{p}\right)' \left(\frac{\text{Var}(b(k))}{\sigma^2\overline{s}} - \frac{I}{p}\right) b \\
&= b' \left(\frac{G\text{diag}(\delta_i\lambda_i^{-1})G'}{\overline{s}} - \frac{GG'}{p}\right)' \left(\frac{G\text{diag}(\delta_i\lambda_i^{-1})G'}{\overline{s}} - \frac{GG'}{p}\right) b \\
&= b'G \left(\frac{\text{diag}(\delta_i\lambda_i^{-1})}{\overline{s}} - \frac{I}{p}\right)' \left(\frac{\text{diag}(\delta_i\lambda_i^{-1})}{\overline{s}} - \frac{I}{p}\right) G'b \\
&= a'\text{diag}\left(\frac{\delta_i\lambda_i^{-1})}{\overline{s}} - \frac{1}{p}\right)^2 a \\
&= \sum_{i=1}^{p} \frac{a_i^2}{p^2} \left(\frac{p\delta_i^2\lambda_i^{-1}}{\overline{s}} - 1\right)^2,
\end{aligned} \tag{3.10}$$

as a simplification we cancel $\alpha_i^2 p^{-2}$ giving

$$\text{ISRM} = \sum_{i=1}^{p} \left(\frac{p\delta_i^2 \lambda_i^{-1}}{\sum_{i-1}^{p} \delta_i^2 \lambda_i^{-1}} - 1 \right)^2 . \tag{3.11}$$

In practice one evaluates ISRM for a range of k values and selects k at the first local minimum or at a prespecified reduction, say 50% of the ISRM at $k = 0$. Vinod recommends this selection rule to keep the bias small since ISRM's global minimum will emphasize stability without regard to bias. Winston and Churchill (1978) conducted a simulation and found that the ISRM gave k values which tended to be too large. Vinod (1979) rejected their results because of a deficiency in their simulation. Thus the m scale trace used in conjunction with the ISRM will overcome the problems of stochastic k and unreliable stable region of the ridge trace.

The Hoerl-Kennard-Baldwin Estimator.

Hoerl, Kennard , and Baldwin (1975) suggested the following mechanical rule to determine the biasing parameter k given by

$$k_{HKB} = \frac{ps^2}{\sum_{i=1}^{p} b_i^2} \tag{3.12}$$

where

$$s^2 = (Y - Xb^r)'(Y - Xb^r)/v \tag{3.13}$$

is the usual least squares estimate of σ^2 , b are the least squares beta estimates, and $v = n - p$ for the no intercept model or $v = n - p - 1$ otherwise . Many independent simulations have shown the overall good performance of this estimator.

Considering the first order condition for a minimum of the $i-th$ component of the MSE $b_i(k)$ we have

$$\begin{aligned}\frac{d}{dk}\text{MSE } b_i(k) &= \frac{d}{dk}\left(\frac{\lambda_i\sigma^2 + k^2\alpha_i^2}{(\lambda_i+k)^2}\right) \\ &= \frac{2\lambda_i(-\sigma^2 + k\alpha_i^2)}{(\lambda_i+k)^3} = 0\end{aligned} \tag{3.14}$$

at

$$k = \frac{\sigma^2}{\alpha_i^2} \tag{3.15}$$

Consider p different biasing parameters, each given by equation (3.15) . Now the harmonic mean of these p optimal k_i will give

$$\begin{aligned}k_h &= \frac{p}{\sum_{i=1}^{p}\frac{1}{k_i}} \\ &= \frac{p}{\sum_{i=1}^{p}\frac{\alpha_i^2}{\sigma^2}} \\ &= \frac{p\sigma^2}{\alpha'\alpha} \\ &= \frac{p\sigma^2}{\alpha'G'G\alpha} \\ &= \frac{p\sigma^2}{\beta'\beta}.\end{aligned} \tag{3.16}$$

Using the ordinary least squares estimates s^2, b for σ^2, β respectively we have

$$k_{HKB} = \frac{ps^2}{b'b}. \tag{3.17}$$

In a later paper, Hoerl and Kennard (1976) suggested an iterative version of their biasing parameter since $b'b$ tends to overestimate $\beta'\beta$. Specifically they

suggested the following sequence of estimates of β and k.

$$
\begin{aligned}
k^0 &= \frac{ps^2}{b'b} \\
k^1 &= \frac{ps^2}{b'(k^0)b(k^0)} \\
k^2 &= \frac{ps^2}{b'(k^1)b(k^1)} \\
&\cdot \\
&\cdot \\
&\cdot \\
k^t &= \frac{ps^2}{b'(k^{t-1})b(k^{t-1})}.
\end{aligned}
\tag{3.18}
$$

For the above sequence of estimates Hoerl and Kennard recommended the following stopping rule.

$$
\text{If } \frac{k^{j+1} - k^j}{k^j} \leq 20T^{-1.3}, \text{ where } T = \mathrm{Tr}(X'X)^{-1}/p \tag{3.19}
$$

then the iteration should stop at $b(k^j)$.

They cite simulation studies where this termination rule performed well. Another stopping rule used by Gibbons (1981) is to apply a 10^{-4} convergence criterion to the successive k values and default to least squares if convergence is not obtained by 30 iterations. We note that the unstandardized ridge coefficients $b_i(k)$ can always be derived from the standardized ridge coefficients , here denoted by $b_i^s(k)$ using the following identities

$$
b_i(k) = \frac{b_i^s(k)S_y}{S_i}, \text{ for } i = 1, 2, \ldots, p, \tag{3.20}
$$

and

$$
b_0(k) = \overline{y} - b_1(k)\overline{x}_1 - b_2(k)\overline{x}_2 - \cdots - b_p(k)\overline{x}_p, \tag{3.21}
$$

where

$$S_j^2 = \sum_{j=1}^{n} \frac{(x_{ij} - \overline{x}_j)^2}{(n-1)} \tag{3.22}$$

and x_j denotes the $j-th$ column of the unstandardized X matrix.

The Lawless and Wang Estimator.

Lawless and Wang (1976), using a Bayesian approach, suggested the following mechanical rule to select the biasing parameter k.

$$k_{LW} = \frac{ps^2}{\sum_{i=1}^{p} \lambda_i a_i^2} \tag{3.23}$$

where $a = G'b$ are the standardized uncorrelated least squares estimates of $\alpha = G'\beta$, and s^2 is the least squares estimate of σ^2.

The McDonald-Galarneau Estimator.

McDonald and Galarneau(1975) considered the expected squared length of the least squares estimates b to derive their estimator. Since

$$E(b'b) = \beta'\beta + \sigma^2 \mathrm{Tr}(X'X)^{-1} \tag{3.24}$$

they considered the following unbiased estimate of the squared length of the unknown coefficient vector $\beta'\beta$, given as

$$Q = b'b - s^2 \sum_{i=1}^{p} \lambda_i^{-1} \tag{3.25}$$

Taking expectations of both sides we see that

$$E(Q) = \beta'\beta + \sigma^2 \mathrm{Tr}(X'X)^{-1} - \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} = \beta'\beta \tag{3.26}$$

Their idea is to select k at such a value that $b'(k)b(k) = Q$. To use their estimator, we evaluate

$$\mid b'(k)b(k) - Q \mid \tag{3.27}$$

for a range of k values and select k corresponding to the minimum absolute difference. In the case that $b'b - s^2 \sum_{i=1}^{p} \lambda_i^{-1} < 0$, select $k = 0$.

The SRIDG Estimator.

Dempster, Schatzoff, and Wermuth (1977) considered the MSE $b(k)$ to derive their estimator SRIDG. Considering that the MSE $b(k)$ is minimized when

$$\frac{d}{dk}\text{MSE } b(k) = \sum_{i=1}^{p} 2\lambda_i \frac{(k\alpha_i^2 - \sigma^2)}{(\lambda_i + k)^3} = 0, \tag{3.28}$$

they suggested evaluating

$$\mid \sum_{i=1}^{p} 2\lambda_i \frac{(ka_i^2(k) - s^2)}{(\lambda_i + k)^3} \mid, \tag{3.29}$$

for a range of k values and selecting that value of k associated with the observed minimum.

The RIDGM Estimator.

Dempster, Schatzoff, and Wermuth (1977) taking a Bayesian approach, suggested another mechanical rule to select the biasing parameter. For this estimator, we select k so that

$$\sum_{i=1}^{p} \frac{a_i^2}{(\sigma_\beta^2 + s^2\lambda_i^{-1})} = p, \text{ where } \sigma_\beta^2 = \frac{s^2}{k}. \tag{3.30}$$

The Lindley and Smith Estimator .

Taking a Bayesian approach, Lindley and Smith (1972) suggest the following procedure to estimate $k = \frac{\sigma^2}{\sigma_\beta^2}$. Suppose $\beta \sim N(0, \sigma_\beta^2 I)$, and σ^2 and σ_β^2 are independent each having an inverse chi-squared distribution; that is, $\nu\eta/\sigma^2 \sim \chi_\nu^2$ and $\nu_\beta\eta_\beta/\sigma_\beta^2 \sim \chi_{\nu\beta}^2$. From the mode of the posterior distribution , the following equations are used to select k .

$$b_{LS} = (X'X + (s^2/s_\beta^2)I)^{-1}X'Y,$$

$$s^2 = [\nu\eta + (Y - Xb_{LS})'(Y - Xb_{LS})]/(n + \nu + 2), \tag{3.31}$$

$$s_\beta^2 = (\nu_\beta\eta_\beta + b'_{LS}b_{LS})/(p + \nu_\beta + 2).$$

An initial estimate of b_{LS} is b. From this estimate, we find s^2 and s_β^2. The proceedure is repeated until convergence is obtained. Gibbons (1981) uses a 10^{-4} convergence criterion for $k = s^2/s_\beta^2$, which defaults to least squares if convergence is not obtained. If the estimates are not sensitive to small positive values of ν and η_β they can both be set to zero in which case b_{LS} is similar to $b(k)_{HKB}$ iterated estimator.

The above rules and their properties have been studied primarily though the use of Monte Carlo experiments. Other independent simulations by Miller and Tracy (1984), Gibbons(1981), Winston and Churchill(1978), and McDonald and Galarneau (1975) have been conducted and have shown the overall superior performance of various estimators. The problem with such comparisons across simulations is that no one rule can be shown superior to least squares under all conditions. Furthermore, any results from a particular simulation holds only for that experimental design and the parameter values considered.

Chapter 4

DOMINANCE FOR STOCHASTIC k

The MSE or MTxMSE conditions for the dominance of the ordinary ridge estimator over the least squares estimator considered so far all imply a fixed k. However most of the operational ridge estimators use the stochastic estimators b and s^2. We will consider here two such estimators and give conditions for there dominance over least squares.

Consider the following operational ridge estimators. The Hoerl,Kennard, and Baldwin (1975) estimator, where the biasing parameter k is given by

$$k_{HKB} = \frac{ps^2}{(b'b)} = \frac{ps^2}{a'a}, \tag{4.1}$$

and the Lawless and Wang (1976) estimator, where k is given by

$$k_{LW} = \frac{ps^2}{\sum_{i=1}^{p} \lambda_i a_i^2} = \frac{ps^2}{a'\Lambda a}. \tag{4.2}$$

Both these choises for k are stochastic as they depend on the random least squares estimates s and a. However, we can also consider them to be members of the double h class family of k given by

$$k = k_{W,\, h_1,\, h_2} = \frac{h_1 e'e}{a'Wa + h_2 e'e} \tag{4.3}$$

where $Y - Xb = e$, $W = \mathrm{diag}(w_1, w_2, \ldots, w_p)$ is a given diagonal matrix, h_1, h_2 are arbitrary scalars, and $s^2 = e'e/\ \nu$ where

$$\nu = \begin{cases} n-p, & \text{for no intercept model;} \\ n-p-1, & \text{otherwise.} \end{cases} \tag{4.4}$$

Now for $W = I,\ h_1 = p/\nu,\ h_2 = 0$ we have

$$k_{(\ I,\ p/\nu\ ,\ 0)} = \frac{\frac{p}{\nu}e'e}{a'Ia} = \frac{ps^2}{a'a}, \tag{4.5}$$

giving the Hoerl Kennard and Baldwin estimator.

For $W = \Lambda,\quad h_1 = p/\nu,\ h_2 = 0$ we have

$$k_{(\ \Lambda,\ p/\nu\ ,\ 0)} = \frac{\frac{p}{\nu}e'e}{a'\Lambda a} = \frac{ps^2}{\sum_{i=1}^{p}\lambda_i a_i^2}, \tag{4.6}$$

the Lawless and Wang estimator.

In terms of shrinkage factors $\delta_i/(\lambda_i + k)$, we have

$$\begin{aligned}
\delta_i &= \frac{\lambda_i}{(\lambda_i + k)} \\
&= \frac{\lambda_i}{\lambda_i + \frac{h_1 e'e}{a'Wa + h_2 e'e}} \\
&= \frac{a'Wa + h_2 e'e + \frac{h_1}{\lambda_i}e'e - \frac{h_1}{\lambda_i}e'e}{a'Wa + h_2 e'e + \frac{h_1}{\lambda_i}e'e} \\
&= \left[1 - \frac{\frac{h_1}{\lambda_i}e'e}{a'Wa + (\frac{h_1}{\lambda_i} + h_2)e'e}\right] \\
&= \left[1 - \frac{h_{1i}e'e}{a'Wa + h_{2i}e'e}\right];
\end{aligned} \tag{4.7}$$

where $h_{1i} = h_1\lambda_i^{-1},\ h_{2i} = h_1\lambda_i^{-1} + h_2$.

Considering our ridge estimators, we have

$$\begin{aligned}
b(k) &= (X'X + kI)^{-1}X'Y \\
&= (X'X + kI)^{-1}X'X(X'X)^{-1}X'Y \\
&= G(\Lambda + kI)^{-1}\Lambda G'b \\
&= G\Delta G'b
\end{aligned} \tag{4.8}$$

and since $G'b(k) = a(k) = \Delta G'b = \Delta a$, we may write our operational ridge estimators as members of the double h class family of estimators denoted by

$$a_{DHC} = a(W, h_1, h_2) = \hat{\Delta} a \tag{4.9}$$

where

$$\hat{\Delta} = \text{diag}(\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_p) \tag{4.10}$$

and

$$\hat{\delta}_i = \left[1 - \frac{h_{1i} e'e}{a'Wa + h_{2i} e'e}\right]. \tag{4.11}$$

Hence the $i-th$ element of the $a(DHC)$ estimator is given by

$$a(DHC)_i = \hat{\delta}_i a_i = \left[1 - \frac{h_{1i} e'e}{a'Wa + h_{2i} e'e}\right] a_i. \tag{4.12}$$

Vinod and Ullah (1981) give the following theorems for the BIAS , and MSE of the double h class estimators. Their derivation, based on Kadane's (1971) small sigma expansion, is given in Appendix B.

Theorem 4.1: The asymptotic expansion of the bias of the double h class ridge estimator, up to order σ^2 , is given by

$$E(a(DHC) - \alpha)_i = -\frac{h_1}{2\theta} \nu \lambda_i^{-1} \alpha_i \tag{4.13}$$

when $h_1 > 0$ and $\theta = \frac{\alpha' W \alpha}{2\sigma^2}$ is a weighted non-centrality parameter.

Theorem 4.2: The asymptotic expansion of the MSE of the double h class ridge estimator, up to order σ^4 is given by

$$E(a(DHC) - \alpha)_i^2 = \frac{\sigma^2}{\lambda_i} + \frac{\nu h_{1i}}{4\theta^2} \left[\alpha_i^2 \left(4\frac{w_i}{\lambda_i} + h_{1i}(\nu + 2)\right) - 2\frac{\alpha' W \alpha}{\lambda_i}\right], \tag{4.14}$$

where $h_{1i} = h_1\lambda_i^{-1}$, and w_i is the i-th diagonal element of W.

Using theorem 2 we can write the MSE of the DHC estimators as

$$\begin{aligned}\text{MSE } (a(DHC)) &= \sigma^2 \text{Tr}\Lambda^{-1} + \frac{\nu h_1}{4\theta^2}[4\alpha' W\Lambda^{-2}\alpha + h_1(\nu+2)\alpha'\Lambda^{-2}\alpha \\ &\quad - 2\alpha' W\alpha \text{Tr}\Lambda^{-2}] \\ &= \sigma^2 \text{Tr}\Lambda^{-1} + \frac{\nu h_1}{4\theta^2}\alpha'\Lambda^{-2}\alpha \\ &\quad [h_1(\nu+2) - 2\frac{\alpha'\Lambda^{-2}W\alpha}{\alpha\Lambda^{-2}\alpha}(\frac{\alpha' W\alpha}{\alpha'\Lambda^{-2}W\alpha}\text{Tr}\Lambda^{-2} - 2)].\end{aligned} \tag{4.15}$$

Considering the MSE criterion for dominance over least esquares, we have

$$\text{MSE } (a) - \text{MSE } (a(DHC)) > 0 \tag{4.16}$$

$$\begin{aligned}&\text{when } \frac{\nu h_1}{4\theta^2}\alpha'\Lambda^{-2}\alpha \\ &\quad [h_1(\nu+2) - 2\frac{\alpha'\Lambda^{-2}W\alpha}{\alpha\Lambda^{-2}\alpha}(\frac{\alpha' W\alpha}{\alpha'\Lambda^{-2}W\alpha}\text{Tr}\Lambda^{-2} - 2)] < 0.\end{aligned} \tag{4.17}$$

Using the relation $min\frac{x'Ax}{x'Bx}$ is the minimum value of λ in $det(A - \lambda B)$ Rao p. 74, this is equivalent to

$$0 < h_1 < \frac{2w_{\min}}{(\nu+2)}[\min(\lambda_i^2)\text{Tr}(\Lambda^{-2}) - 2], \tag{4.18}$$

$$\text{for } \min(\lambda_i^2)\text{Tr}\Lambda^{-2} > 2. \tag{4.19}$$

Thus we have the following acceptable range for the k_{HKB} estimator

$$0 < h_1 < \frac{2}{(\nu+2)}[\lambda_p^2 \sum_{i=1}^{p}(\lambda_i^{-2}) - 2], \tag{4.20}$$

and for k_{LW} we have

$$0 < h_1 < \frac{2\lambda_p}{(\nu+2)}[\lambda_p^2 \sum_{i=1}^{p}(\lambda_i^{-2}) - 2]. \tag{4.21}$$

The condition

$$\text{for } \min(\lambda_i^2)\text{Tr}\Lambda^{-2} > 2, \tag{4.22}$$

implies $p > 2$. Since h_2 is not included in condition (4.18), the results hold for any $h_2 \geq 0$. That $h_2 \geq 0$ will ensure that the moments of the i-th component of a_{DHC} exist, since the denominator $a'Wa + h_{2i}e'e$ will always be defined.

Chapter 5

RECENT DEVELOPMENTS IN SELECTING k

In this section we will review some new proposals to select k. In particular we will consider a bootstrap method to select k, some non-stochastic or deterministic biasing parameters , and an algorithm to select the optimal biasing parameter.

A Bootstrap Method to Select k.

Delaney and Chatterjee (1986) proposed a new ridge estimator where the biasing parameter k is selected using a combination of cross-validation and bootstrap replications. The procedure is computationally intensive and implemented as follows. Assume a population of n observations on p regressors and one predictor variable. A sample of n observations *with replacement* is taken giving a single bootstrap sample. Using this sample, the ridge regression estimates are computed for a selected range of k values, say k from 0 to 0.1 in steps of 0.002. The observations which were excluded from this sample are then predicted from the estimates obtained. The procedure is then repeated for a large number I of bootstrap samples. Denoting the prediction vector of the missing observations as

$$\hat{y}_{ni}(k_g), \tag{5.1}$$

where ni denotes the number of missing observations in the $i-th$ bootstrap sample, and g denotes the $g-th$ value of the k parameter. The corresponding vector of actual observations for the missing observations in the $i-th$ bootstrap

sample is denoted as

$$y_{ni}. \tag{5.2}$$

The mean square error of prediction MSEP for the $i-th$ sample and the $g-th$ k parameter is defined as

$$\text{MSEP}_i(k_g) = \frac{(\hat{y}_{ni}(k_g) - y_{ni})'(\hat{y}_{ni}(k_g) - y_{ni})}{ni} \tag{5.3}$$

$$\text{for } g = 1, 2, \ldots, G. \tag{5.4}$$

Assuming we have i bootstrap samples, $i = 1, 2, \ldots, I$, a final average MSEP for each k_g is given by

$$\frac{\sum_{i=1}^{I}(\text{MSEP}_i(k_g))ni}{\sum_{i=1}^{I} ni}, \tag{5.5}$$

$$\text{for } g = 1, 2, \ldots, G. \tag{5.6}$$

The bootstrap choice of the ridge parametar k will be k_B where

$$k_B = \min(\text{MSEP}(k_g)). \tag{5.7}$$

The authors carried out a simulation and also illustrated their technique on two sets of previously published data. For their simulation they used the following values for the signal to noise ratio SNR $= \beta'\beta/\sigma^2$, and condition number ϕ of the $X'X$ matrix .

$$\phi = 4\ 25\ 100\ 2,500\ 10,000$$

$$\text{SNR} = 1\ 4\ 9\ 25\ 400$$

$$\beta'\beta = 4\ 16\ 100\ 900$$

From their simulations they gave the following results.

1) The bootstrap MSEP outperformed k_{HKB} estimator in 91% of the models and outperformed the least squares estimates b in 70% of the models under the MSE criterion.

2) Using the criterion of minimum MSEP , the estimators in order of preference were the Bootstrap, HKB, and LS.

3) In cases where the data is collinear and the ridge method does not give a better predictive model the bootstrap estimator will have its minimum MESP at $k = 0$ indicating the least squares estimates b will give the best performance under the MSEP criterion.

4) For each choice of k a measure of uncertainity is provided by the standard error of prediction given by

$$\sqrt{\sum_{i=1}^{I} \frac{(\text{MSEP}_i(k_g) - \overline{\text{MSEP}(k_g)})^2}{I-1}}. \tag{5.8}$$

A Deterministic Ridge estimator.

Lee (1986) conducted a simulation to compare some non-stochastic rules to select the biasing parameter k for the ordinary ridge estimator $b(k)$. From his results he concluded that the biasing parameter $k = \lambda_p$ was the best performer. In particular, if the condition number of the $X'X$ matrix, defined as $\phi = \lambda_1/\lambda_p$, where $\max(\lambda_i) = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p = \min(\lambda_i) > 0$ are the eigen values of the $X'X$ matrix, was at least 1000 then this estimator dominated the least

squares estimates b for the entire range of parameter values considered. The four deterministic ridge rules are given as follows,

$$R1 : k = \sqrt{\lambda_p}$$

$$R2 : k = \lambda_p$$

$$R3 : k = \sqrt{\lambda_p \lambda_{p-1}}$$

$$R4 : k = \sqrt[4]{\lambda_p^3 \lambda_{p-1}}$$

Lee considered the $i-th$ component of the MSE$b(k)$ given as

$$\text{MSE}(b_i(k)) = \frac{\sigma^2 \lambda_i + k^2 \alpha_i^2}{(\lambda_i + k)^2} \tag{5.9}$$

Defining the efficiency of the $i-th$ ridge estimator relative to the $i-th$ least squares estimator as

$$\eta_i = \frac{\text{MSE } b_i}{\text{MSE } b_i(k)} = \frac{E(b_i - \beta_i)^2}{E(b_i(k) - \beta_i)^2} \tag{5.10}$$

Lee gives the following theorem which provides a sufficient condition that the ridge estimator be more efficient than least squares .

Theorem 5.1: If k is chosen deterministically and

$$\frac{\lambda_i \alpha_i^2}{\sigma^2} < 1 + 2\lambda_i k^{-1}, \text{ for } i = 1, 2, \ldots, p, \tag{5.11}$$

$$\text{then } \eta_i > 1, \text{ for all i .} \tag{5.12}$$

Proof: If

$$\eta_i = \frac{\text{MSE } b_i}{\text{MSE } b_i(k)} = \frac{\sigma^2}{\lambda_i} \frac{(\lambda_i + k)^2}{\lambda_i \sigma^2 + \alpha_i^2 k^2} > 1 \tag{5.13}$$

$$\text{then} \quad 2k\lambda_i\sigma^2 + k^2\sigma^2 > k^2\lambda_i\alpha_i \tag{5.14}$$

$$\text{or} \quad \frac{2\lambda_i}{k} + 1 > \frac{\lambda_i\alpha_i^2}{\sigma^2}. \tag{5.15}$$

Considering this criterion for dominance, $R2$ gives the widest possible range for the condition (5.11) to be satisfied. Thus we would expect $R2$ to have the best performance in simulation. Lee considered three models with condition numbers given by $\phi = 54,\ 924,\ 1397$. The SNR was varied from 1 to $10,000$. The various estimators were compared with the least squares estimates under the MSE and PMSE criterion, where

$$\text{PMSE } b = E(b-\beta)'X'X(b-\beta). \tag{5.16}$$

Of all the estimators, $R2$ dominated the least squares estimates under the MSE criterion for all values of ϕ and SNR considered with the sole exception of one case where $\phi = 54$ and SNR $= 10,000$. Under the PMSE criterion, $R2$ performed the best as long as SNR ≤ 400, and $\phi > 54$. In light of these observations, Lee recommended the use of

$$k = \lambda_p, \tag{5.17}$$

over all the other estimators, especially when $\phi \geq 1000$.

An Algorithm for Optimum Ridge parameter Selection.

Lee (1987) suggested a computational procedure to find the optimal k using the least squares estimates in the minimization of the MSE or PMSE . The

procedure uses the Newton-Raphson method to minimize the functions. Considering

$$\begin{aligned} \text{MSE } b(k) &= \sum_{i=1}^{p} \frac{\sigma^2\lambda_i + k^2\alpha_i^2}{(\lambda_i + k)^2}, \\ \text{PMSE } b(k) &= \sum_{i=1}^{p} \lambda_i \frac{(\sigma^2\lambda_i + k^2\alpha_i^2)}{(\lambda_i + k)^2} \end{aligned} \tag{5.18}$$

if we replace α_i^2 and σ^2 by their least squares estimates a^2 and s^2 we have the following estimates to minimize

$$\begin{aligned} \hat{f}_1(k) &= \sum_{i=1}^{p} \frac{s^2\lambda_i + k^2a_i^2}{(\lambda_i + k)^2} \\ \hat{f}_2(k) &= \sum_{i=1}^{p} \lambda_i \frac{(s^2\lambda_i + k^2a_i^2)}{(\lambda_i + k)^2} \end{aligned} \tag{5.19}$$

Lee gave the following algorithm to minimize $\hat{f}_j(k)$

Step 1. Set $k^0 = 0$ and $i = 0$.

Step 2. Compute k^{i+1} from

$$k^{i+1} = k^i - \frac{\hat{f}_j'(k^i)}{\hat{f}_j''(k^i)} \tag{5.20}$$

Where $\hat{f}_j'(k)$, $\hat{f}_j''(k)$ for $j = 1, 2.$ are the first and second derivatives of $\hat{f}_j$ with respect to k.

Step 3 . If

$$| k^{i+1} - k^i | < \delta \tag{5.21}$$

for a given $\delta > 0$ stop. Otherwise set $i = i + 1$ and go to step 2.

Although no proof of the convergence of (5.20) is available, Lee found the algorithm converged in all his simulations, with $\delta_i = 10^{-6}$, under 15 iterations.

His algorithm gave a ridge estimator which was dominant over least squares under the MSE and PMSE criterions for condition numbers ranging from 50 to 1400. His paper includes a FORTRAN program to compute the optimal k.

Chapter 6

GENERALIZED RIDGE ESTIMATORS

The singular value decomposition (SVD) of the matrix X is often used to simplify analysis of the generalized ridge estimators. The SVD of the matrix X is given by $X = H\Lambda^{\frac{1}{2}}G'$ where H is the $(n \times p)$ matrix of eigenvectors corresponding to the non-zero eigenvalues of the XX' matrix, standardized so that $H'H = I$, $\Lambda^{\frac{1}{2}}$ is the square root matrix of the eigenvalues of the $X'X$ matrix, and G is the $(p \times p)$ matrix of eigenvectors corresponding to the eigenvalues of $X'X$ such that $G\Lambda G' = X'X$, and $G'G = I$.

Using the SVD we can write $X'X = G\Lambda^{\frac{1}{2}}H'H\Lambda^{\frac{1}{2}}G' = G\Lambda G'$,

and

$$\begin{aligned} Y &= X\beta + \varepsilon \\ &= H\Lambda^{\frac{1}{2}}G'\beta + \varepsilon \\ &= H\Lambda^{\frac{1}{2}}\alpha + \varepsilon. \end{aligned} \tag{6.1}$$

Hence the least square estimates of $\alpha = G'\beta$, denoted by a, are given by

$$\begin{aligned} a &= (\Lambda^{\frac{1}{2}}H'H\Lambda^{\frac{1}{2}})^{-1}\Lambda^{\frac{1}{2}}H'Y \\ &= \Lambda^{-1}\Lambda^{\frac{1}{2}}H'Y \\ &= \Lambda^{-\frac{1}{2}}H'Y, \end{aligned} \tag{6.2}$$

and since $Ga = b$ we have the least squares estimator as $b = G\Lambda^{-\frac{1}{2}}H'Y$.

Using the SVD for the ordinary ridge estimator, we have

$$\begin{aligned} b(k) &= (X'X + kI)^{-1}X'Y \\ &= G(\Lambda + kI)^{-1}G'G\Lambda^{\frac{1}{2}}H'Y \\ &= G(\Lambda + kI)^{-1}\Lambda G'G\Lambda^{-1}\Lambda^{\frac{1}{2}}H'Y \\ &= G(\Lambda + kI)^{-1}\Lambda G'Ga \\ &= G\Delta G'b. \end{aligned} \tag{6.3}$$

Here $\Delta = \text{diag}(\delta_1, \delta_2, \ldots, \delta_p)$ and $\delta_i = \lambda_i/(\lambda_i + k)$ are the shrinkage factors. We note that for any fixed positive k and declining eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$, we have

$$1 > \frac{\lambda_1}{\lambda_1 + k} \geq \frac{\lambda_2}{\lambda_2 + k} \geq \cdots \geq \frac{\lambda_p}{\lambda_p + k} > 0, \tag{6.4}$$

with the smallest shrinkage for the minor axis λ_p .

For the variance of the least square estimator b we have

$$\begin{aligned} \text{Var}(b) &= \sigma^2(X'X)^{-1} \\ &= \sigma^2 G\Lambda^{-1}G' \\ &= G\ \text{Var}(a)\ G' \\ &= \text{Var}(Ga). \end{aligned} \tag{6.5}$$

Noting that

$$\text{Var}(a) = \sigma^2 \text{diag}(\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_p^{-1}), \tag{6.6}$$

we see that covariance matrix of a is diagonal making the components of a_i uncorrelated. Also, since $0 < \lambda_1^{-1} \leq \lambda_2^{-1} \leq \cdots \leq \lambda_p^{-1}$, we have that $0 < \text{Var } a_1 \leq \text{Var } a_2 \leq \cdots \leq \text{Var } a_p$. Hence $\text{Var}(a_p) = \sigma^2\lambda_p^{-1}$ is the uncorrelated component with the largest variance. Using the property of declining shrinkage factors δ_i

(6.4) , we also see that δ_p is the smallest shrinkage factor corresponding to the component with the largest variance.

The generalized ridge estimator, introduced by Hoerl and Kennard (1970), is similar to the ordinary ridge estimator where each diagonal element of Λ is augmented by a positive k_i . Thus the diagonal matrix Λ is augmented by the diagonal matrix $K = \text{diag}(k_1, k_2, \ldots, k_p)$ before inverting for least squares. The generalized ridge estimator of β for the model

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim (0, \sigma^2 I) \tag{6.7}$$

is given by

$$b(K) = (X'X + GKG')^{-1}X'Y, \tag{6.8}$$

$$\text{where } K = \text{diag}(k_1, \ldots, k_p). \tag{6.9}$$

To derive the generalized ridge estimator of β we use the SVD of the matrix X in the model (1.2) giving

$$\begin{aligned} Y &= H\Lambda^{\frac{1}{2}}G'\beta + \varepsilon \\ &= H\Lambda^{\frac{1}{2}}\alpha + \varepsilon, \end{aligned} \tag{6.10}$$

where the least square estimates of α , given by

$$\begin{aligned} a &= (\Lambda^{\frac{1}{2}}H'H\Lambda^{\frac{1}{2}})^{-1}\Lambda^{\frac{1}{2}}H'Y \\ &= (\Lambda)^{-1}\Lambda^{\frac{1}{2}}H'Y, \end{aligned} \tag{6.11}$$

is augmented by the matrix K before inversion giving the generalized ridge estimator

$$a(K) = (\Lambda + K)^{-1}\Lambda^{\frac{1}{2}}H'Y, \tag{6.12}$$

of the uncorrelated components α .

With respect to $b(K)$ we have

$$\begin{aligned} b(K) &= Ga(K) \\ &= G(\Lambda + K)^{-1}\Lambda^{\frac{1}{2}}H'Y \\ &= G(\Lambda + K)^{-1}G'G\Lambda^{\frac{1}{2}}H'Y \\ &= (X'X + GKG')^{-1}X'Y. \end{aligned} \tag{6.13}$$

We may also write

$$\begin{aligned} b(K) &= G(\Lambda + K)^{-1}\Lambda^{\frac{1}{2}}H'Y \\ &= G(\Lambda + K)^{-1}\Lambda G'G\Lambda^{-1}\Lambda^{\frac{1}{2}}H'Y \\ &= G(\Lambda + K)^{-1}\Lambda G'G\Lambda^{-\frac{1}{2}}H'Y \\ &= G(\Lambda + K)^{-1}\Lambda G'b \\ &= G\Delta G'b, \end{aligned} \tag{6.14}$$

where

$$\Delta = \mathrm{diag}(\delta_1, \delta_2, \ldots, \delta_p), \tag{6.15}$$

and

$$\delta_i = \frac{\lambda_i}{(\lambda_i + k_i)}, i = 1, 2, \ldots, p. \tag{6.16}$$

Comparing with (6.3) we can see that the ordinary ridge estimator is a special case of the generalized ridge estimator where all $k_i = k$.

The MSE Of The Generalized Ridge Estimator.

To derive the MSE of the estimator $b(K)$ we consider

$$\begin{aligned} b(K) &= (X'X + GKG')^{-1}X'Y \\ &= G\Delta G'b, \end{aligned} \tag{6.17}$$

with

$$E(b(K)) = G\Delta G'\beta, \tag{6.18}$$

$$\begin{aligned} \text{BIAS}\,(b(K)) &= E(b(K)) - \beta \\ &= G(\Delta - I)G'\beta, \end{aligned} \tag{6.19}$$

and

$$\begin{aligned} [b(K) - E(b(K))] &= G\Delta G'b - G\Delta G'\beta \\ &= G\Delta G'(b - \beta). \end{aligned} \tag{6.20}$$

Using

$$\text{MSE}\; b(K) = E[b(K) - E(b(K))]'[b(K) - E(b(K))] + [E(b(K)) - \beta]'[E(b(K)) - \beta], \tag{6.21}$$

we have for the first term

$$\begin{aligned} & E[b(K) - E(b(K))]'[b(K) - E(b(K))] \\ &= E[(b - \beta)'G\Delta G'G\Delta G'(b - \beta)] \\ &= \text{Tr}[G\Delta G'G\Delta G'\text{Var}(b)] \\ &\quad + E(b - \beta)'G\Delta G'G\Delta G'E(b - \beta) \\ &= \sigma^2\text{Tr}[(X'X)^{-1}G\Delta^2 G'] \\ &= \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1}\delta_i^2 \\ &= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k_i)^2}. \end{aligned} \tag{6.22}$$

For the second term in the MSE we have

$$\begin{aligned} (\ \text{BIAS}\,(b(K))\ )^2 &= \beta'G(\Delta - I)G'G(\Delta - I)G'\beta \\ &= \alpha'(\Delta - I)^2\alpha \\ &= \sum_{i=1}^{p} \alpha_i^2(\delta_i - 1)^2 \\ &= \sum_{i=1}^{p} \frac{\alpha_i^2 k_i^2}{(\lambda_i + k_i)^2}\ . \end{aligned} \tag{6.23}$$

Combining (6.22) and (6.23) we have

$$\text{MSE } b(K) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k_i)^2} + \sum_{i=1}^{p} \frac{\alpha_i^2 k_i^2}{(\lambda_i + k_i)^2}. \tag{6.24}$$

Showing that the MSE of $b(K)$ is similar to MSE of $b(k)$ except for the differing k_i values.

The MTxMSE of the Generalized Ridge Estimator.

Using the results of the equations (6.19) (6.20) and (2.38) we have

$$\begin{aligned}
\text{MTxMSE } b(K) &= E[G\Delta G'(b-\beta)(b-\beta)'G\Delta G'] \\
&\quad + G(\Delta - I)G'\beta\beta'G(\Delta - I)G' \\
&= G\Delta G'\text{Var}(b)G\Delta G' + G(\Delta - I)\alpha\alpha'(\Delta - I)G' \\
&= \sigma^2 G\Delta G'G\Lambda^{-1}G'G\Delta G' + G(\Delta - I)\alpha\alpha'(\Delta - I)G' \\
&= \sigma^2 G[\Delta\Lambda^{-1}\Delta + \sigma^{-2}(\Delta - I)\alpha\alpha'(\Delta - I)]G'.
\end{aligned} \tag{6.25}$$

Considering the conditions for dominance of $b(K)$ over b under the MTxMSE criterion, and using MTxMSE $b = \sigma^2 G\Lambda^{-1}G'$ we have

$$\begin{aligned}
\Pi &= \text{MTxMSE } b - \text{MTxMSE } b(K) \\
&= \sigma^2 G[\Lambda^{-1} - \Delta\Lambda^{-1}\Delta - \sigma^{-2}(\Delta - I)\alpha\alpha'(\Delta - I)]G' \\
&= \sigma^2 G[\Lambda^{-1} - (\Lambda + K)^{-1}\Lambda(\Lambda + K)^{-1} - \sigma^{-2}(\Lambda + K)^{-1}K\alpha\alpha'K(\Lambda + K)^{-1}]G' \\
&= \sigma^2 G(\Lambda + K)^{-1}[(\Lambda + K)\Lambda^{-1}(\Lambda + K) - \Lambda - \sigma^{-2}K\alpha\alpha'K](\Lambda + K)^{-1}G' \\
&= \sigma^2 G(\Lambda + K)^{-1}[\Lambda + 2K + K^2\Lambda^{-1} - \Lambda - \sigma^{-2}K\alpha\alpha'K](\Lambda + K)^{-1}G' \\
&= \sigma^2 G(\Lambda + K)^{-1}K[2K^{-1} + \Lambda^{-1} - \sigma^{-2}\alpha\alpha']K(\Lambda + K)^{-1}G' \\
&= G(\Lambda + K)^{-1}K[\ T\ ]K(\Lambda + K)^{-1}G',
\end{aligned} \tag{6.26}$$

where $T = [\sigma^2(2K^{-1} + \Lambda^{-1}) - \alpha\alpha']$. Letting K be a diagonal matrix of positive constants, we have that $G(\Lambda + K)^{-1}K$ is positive definite, and we need only to ensure the T is positive definite for Π to be positive definite. We will consider the following theorems and corollaries of Chawla (1988) such that the generalized ridge estimator would improve on the least squares estimator under the MTxMSE criterion.

Theorem 6.1: A necessary and sufficient condition that the generalized ridge estimator is better than the least squares estimator , under MTxMSE , is given by

$$\alpha'(2K^{-1} + \Lambda^{-1})^{-1}\alpha < \sigma^2. \tag{6.27}$$

Proof: We have

$$\Pi\ , \quad \text{a p.d. matrix} \tag{6.28}$$

$$\text{iff } T = [\sigma^2(2K^{-1} + \Lambda^{-1}) - \alpha\alpha'] \text{ is a p.d. matrix.} \tag{6.29}$$

Now for any square positive definite matrix A we have by definition that $c'Ac > 0$ for all non-zero conformable c vectors. Thus T is positive definite

$$\text{iff } \quad c'[\sigma^2(2K^{-1} + \Lambda^{-1}) - \alpha\alpha']\ c > 0, \tag{6.30}$$

for all non-zero $(p \times 1)$ vectors c . Equivalently we have

$$c'\sigma^2(2K^{-1} + \Lambda^{-1})\ c > c'\alpha\alpha'c, \tag{6.31}$$

or

$$\sigma^2 > \frac{c'\alpha\alpha'c}{c'(2K^{-1} + \Lambda^{-1})c}. \tag{6.32}$$

Now $(2K^{-1}+\Lambda^{-1})$, being positive definite, has an inverse. Dividing both sides of (6.32) by $\alpha'(2K^{-1}+\Lambda^{-1})^{-1}\alpha$ we have the equivalent inequality

$$\frac{\sigma^2}{\alpha'(2K^{-1}+\Lambda^{-1})^{-1}\alpha} > \frac{c'\alpha\alpha'c}{c'(2K^{-1}+\Lambda^{-1})c\alpha'(2K^{-1}+\Lambda^{-1})^{-1}\alpha}. \tag{6.33}$$

Using Theorem A4 (ii) from the appendix, we have

$$\frac{\sigma^2}{\alpha'(2K^{-1}+\Lambda^{-1})^{-1}\alpha} > 1 \geq \frac{c'\alpha\alpha'c}{c'(2K^{-1}+\Lambda^{-1})c\alpha'(2K^{-1}+\Lambda^{-1})^{-1}\alpha}. \tag{6.34}$$

Using the inequality on the left hand side we have that T is a p.d. matrix

$$\text{iff } \sigma^2 > \alpha'(2K^{-1}+\Lambda^{-1})^{-1}\alpha. \tag{6.35}$$

Corollary 6.1: A sufficient condition that the generalized ridge estimator improves on the least squares estimator under the MTxMSE criterion is that

$$\alpha'\Lambda\alpha \leq \sigma^2. \tag{6.36}$$

Proof: Considering that $T = [\sigma^2(2K^{-1}+\Lambda^{-1}) - \alpha\alpha']$ is a p.d. matrix if

$$\sigma^2\Lambda^{-1} - \alpha\alpha' \text{ is a p.s.d. matrix,} \tag{6.37}$$

we have the equivalent condition that

$$c'(\sigma^2\Lambda^{-1} - \alpha\alpha')c \geq 0, \tag{6.38}$$

for all non-zero $(p \times 1)$ vectors c . Thus we have

$$c'\sigma^2\Lambda^{-1}c \geq c'\alpha\alpha'c, \tag{6.39}$$

or

$$\frac{\sigma^2}{\alpha'\Lambda\alpha} \geq 1 \geq \frac{c'\alpha\alpha'c}{c'\Lambda^{-1}c\alpha'\Lambda\alpha}, \tag{6.40}$$

where

$$\frac{c'\alpha\alpha'c}{c'\Lambda^{-1}c\alpha'\Lambda\alpha} = 1 \text{ when } \alpha \propto \Lambda^{-1}c. \tag{6.41}$$

Thus if

$$\frac{\sigma^2}{\alpha'\Lambda\alpha} \geq 1, \tag{6.42}$$

or

$$\sigma^2 \geq \alpha'\Lambda\alpha, \tag{6.43}$$

then we have a sufficient condition that T and so Π , be a p.d. matrix .

Corrollary 6.2: A sufficient condition that the generalized ridge estimator improves on the least squares estimator under the MTxMSE criterion is that

$$\alpha' K\alpha \leq 2\sigma^2. \tag{6.44}$$

Proof: If $[2\sigma^2 K^{-1} - \alpha\alpha']$ is a p.s.d. matrix then

$$T = [\sigma^2(2K^{-1} + \Lambda^{-1}) - \alpha\alpha'] \text{ is a p.d. matrix.} \tag{6.45}$$

Using the same procedure as in the previous corollary, if

$$[2\sigma^2 K^{-1} - \alpha\alpha'] \text{ is a p.s.d. matrix,} \tag{6.46}$$

then by definition

$$c'[2\sigma^2 K^{-1} - \alpha\alpha']c \geq 0, \tag{6.47}$$

or

$$c'2\sigma^2 K^{-1}c \geq c'\alpha\alpha'c. \tag{6.48}$$

This is equivalent to

$$\frac{2\sigma^2}{\alpha' K \alpha} \geq 1 \geq \frac{c'\alpha\alpha' c}{c' K^{-1} c \alpha' K \alpha}, \tag{6.49}$$

where the right hand side attains a maximum at $\alpha \propto K^{-1}c$. Thus if $2\sigma^2 \geq \alpha' K \alpha$ then we have a sufficient condition that Π be a p.d. matrix .

Theorem 6.2: A necessary and sufficient condition for Π to be positive definite is that

$$\theta_i > 1, \quad i = 1, 2, \ldots, p, \tag{6.50}$$

where the $\theta_i' s$ are the roots of the equation

$$\det(\Lambda^{-1} - \alpha\alpha'\sigma^{-2} - \theta 2K^{-1}) = 0 . \tag{6.51}$$

Proof: If

$$(2K^{-1} + \Lambda^{-1} - \alpha\alpha'\sigma^{-2}) \text{ is a p.d. matrix} \tag{6.52}$$

then for a p.d. matrix Q we also have that

$$Q'[2K^{-1} + \Lambda^{-1} - \alpha\alpha'\sigma^{-2}]Q \text{ is a p.d. matrix.} \tag{6.53}$$

Let Q be such that

$$Q' 2K^{-1} Q = I, \text{ and } Q'(\Lambda^{-1} - \alpha\alpha'\sigma^{-2})Q = \Theta \tag{6.54}$$

where

$$\Theta = \text{diag}(\theta_1, \theta_2, \ldots, \theta_p) \tag{6.55}$$

and each θ_i is a root of equation (6.51). Then we have

$$Q'[2K^{-1} + \Lambda^{-1} - \alpha\alpha'\sigma^{-2}]Q = I + \Theta, \text{ is a p.d. matrix} \tag{6.56}$$

iff

$$\theta_i > -1 \text{ for } i = 1, 2, \ldots, p. \tag{6.57}$$

To see that a p.d. matrix Q exists with the required properties, we have

$$\begin{aligned} &\det\,(\Lambda^{-1} - \alpha\alpha'\sigma^{-2} - \theta 2K^{-1}) = 0 \\ \Leftrightarrow\ &\det\,((\frac{1}{2}K)^{\frac{1}{2}}(\Lambda^{-1} - \alpha\alpha'\sigma^{-2})(\frac{1}{2}K)^{\frac{1}{2}} - \theta I) = 0 \\ \Leftrightarrow\ &\det\,(P'(\frac{1}{2}K)^{\frac{1}{2}}(\Lambda^{-1} - \alpha\alpha'\sigma^{-2})(\frac{1}{2}K)^{\frac{1}{2}}P - \theta P'P) \\ &= \det\,(\Theta - \theta I) = 0, \end{aligned} \tag{6.58}$$

where P is a matrix which diagonalizes the symmetric matrix

$$(\frac{1}{2}K)^{\frac{1}{2}}(\Lambda^{-1} - \alpha\alpha'\sigma^{-2})(\frac{1}{2}K)^{\frac{1}{2}} \tag{6.59}$$

normalized so that $P'P = I$. Hence the roots of equation (6.51) are given by the roots of $\det\,(\Theta - \theta I) = 0$ which are $\Theta = \text{diag}(\theta_1, \theta_2, \ldots, \theta_p)$. For the matrix Q, setting $Q = (\frac{1}{2}K)^{\frac{1}{2}}P$ will give a a p.d. matrix with the required properties.

Chapter 7

A BAYESIAN APPROACH TO RIDGE ESTIMATORS

Within the classical approach to the linear regression model, we assume the parameter vector β to be a vector of fixed constants. Conversely, the Bayesian interpretation of the linear model takes the parameter vector β to be a random vector with some known prior distribution function. This distibution expresses the state of knowledge about β before the sample data is to be analyzed.

To derive the Bayesian estimator of β we begin with the prior distribution of β denoted by $f(\beta)$, and the probability model $f(Y \mid \beta) = L(\beta \mid Y)$ where L denotes the likelihood function for β given the data Y. Using Bayes Theorem where

$$f(\beta \mid Y) = \frac{f(Y \mid \beta) f(\beta)}{f(Y)}, \tag{7.1}$$

we may derive the posterior distribution of β given Y denoted by $f(\beta \mid Y)$. In general the marginal distribution of Y, given by

$$f(Y) = \int f(Y \mid \beta) f(\beta) d\beta, \tag{7.2}$$

is treated as the normalizing constant. Thus we have

$$\begin{aligned} f(\beta \mid Y) &= \frac{f(Y \mid \beta) f(\beta)}{f(Y)} \\ &\propto f(Y \mid \beta) f(\beta) \\ &\propto L(\beta \mid Y) f(\beta). \end{aligned} \tag{7.3}$$

From the mean or mode of the posterior distribution we have the bayes estimator of β. In the case that there are other unknown parameters, so-called nuisance

parameters, in the posterior distribution, they are intergrated out and the bayes estimator of β will be the mean or mode of the marginal posterior distribution. Thus

$$\hat{\beta} = E(\beta \mid Y) = \int_{\beta} \beta f(\beta \mid Y) d\beta, \tag{7.4}$$

is a bayes estimator of β .

In particular, let $Y \sim N(X\beta, \sigma^2 I)$ for known σ^2 . The likelihood function of β, σ^2 for a given data set Y is given as

$$\begin{aligned} & L(\beta, \sigma^2 \mid Y) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} exp[-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)] \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} exp[-\frac{1}{2\sigma^2}((Y - Xb)'(Y - Xb) + (\beta - b)'X'X(\beta - b))] \\ &\propto exp[-\frac{1}{2\sigma^2}((\beta - b)'X'X(\beta - b))]. \end{aligned} \tag{7.5}$$

since $\frac{1}{2\sigma^2}(Y - Xb)'(Y - Xb)$ does not involve β , it is treated as a constant of proportionality.

Assuming the random vector β has the following prior and known covariance matrix,

$$\beta \sim N(\beta_0, \sigma^2\Omega), \tag{7.6}$$

we may write the likelihood function as

$$\begin{aligned} L(\beta, \sigma^2\Omega \mid Y) &= \frac{1}{(2\pi)^{\frac{p}{2}}\sigma^p} exp[-\frac{1}{2\sigma^2}(\beta - \beta_0)'\Omega^{-1}(\beta - \beta_0)] \\ &\propto exp[-\frac{1}{2\sigma^2}(\beta - \beta_0)'\Omega^{-1}(\beta - \beta_0)], \end{aligned} \tag{7.7}$$

which is multivariate normal with prior mean vector β_0 . Using Bayes Theorem,

the posterior is proportional to the likelihood times the prior and is given as

$$\begin{aligned} f(\beta \mid Y) &\propto L(\beta \mid Y) f(\beta) \\ &\propto exp[-\frac{1}{2\sigma^2}((\beta - b)'X'X(\beta - b) + (\beta - \beta_0)'\Omega^{-1}(\beta - \beta_0))]. \end{aligned} \tag{7.8}$$

Considering the two quadratic forms in the exponent we have

$$\begin{aligned} &(\beta - b)'X'X(\beta - b) + (\beta - \beta_0)'\Omega^{-1}(\beta - \beta_0) \\ &= \beta'X'X\beta + b'X'Xb - 2\beta'X'Xb + \beta'\Omega^{-1}\beta + \beta_0'\Omega^{-1}\beta_0 - 2\beta'\Omega^{-1}\beta_0 \\ &= \beta'(X'X + \Omega^{-1})\beta - 2\beta'(X'Xb + \Omega^{-1}\beta_0) + \beta_0'\Omega^{-1}\beta_0 + b'X'Xb \\ &= \beta'(X'X + \Omega^{-1})\beta - 2\beta'(X'X + \Omega^{-1})(X'X + \Omega^{-1})^{-1}(X'Xb + \Omega^{-1}\beta_0) \\ &+ (X'Xb + \Omega^{-1}\beta_0) \\ &\quad (X'X + \Omega^{-1})^{-1}(X'X + \Omega^{-1})(X'X + \Omega^{-1})^{-1}(X'Xb + \Omega^{-1}\beta_0) \\ &- (X'Xb + \Omega^{-1}\beta_0) \\ &\quad (X'X + \Omega^{-1})^{-1}(X'X + \Omega^{-1})(X'X + \Omega^{-1})^{-1}(X'Xb + \Omega^{-1}\beta_0) \\ &+ \beta_0'\Omega^{-1}\beta_0 + b'X'Xb \\ &= (\beta - b^*)'\Omega^{-1*}(\beta - b^*) + C^*, \end{aligned} \tag{7.9}$$

where

$$\begin{aligned} b^* &= (X'X + \Omega^{-1})^{-1}(X'Xb + \Omega^{-1}\beta_0) \\ \Omega^* &= (X'X + \Omega^{-1})^{-1}, \end{aligned} \tag{7.10}$$

and since

$$\begin{aligned} C^* &= \beta_0'\Omega^{-1}\beta_0 + b'X'Xb - (X'Xb + \Omega^{-1}\beta_0) \\ &\quad (X'X + \Omega^{-1})^{-1}(X'X + \Omega^{-1})(X'X + \Omega^{-1})^{-1}(X'Xb + \Omega^{-1}\beta_0), \end{aligned} \tag{7.11}$$

has no β term it is treated as a constant of proportionality. Thus the posterior density β can written as

$$f(\beta \mid Y) \propto exp[-\frac{1}{2\sigma^2}(\beta - b^*)'\Omega^{-1*}(\beta - b^*)], \tag{7.12}$$

which has a multivariate normal distribution with mean vector b^* . Thus our bayes estimator of β is given by

$$\hat{\beta} = E(\beta \mid Y) = b^* = (X'X + \Omega^{-1})^{-1}(X'Xb + \Omega^{-1}\beta_0). \tag{7.13}$$

More generaly, if the prior of $\beta \sim N(\beta_0, \sigma_\beta^2\Omega)$, then the bayes estimator of β is given by

$$\begin{aligned} b^* &= (\frac{1}{\sigma^2}X'X + \frac{1}{\sigma_\beta^2}\Omega^{-1})^{-1}(\frac{1}{\sigma^2}X'Xb + \frac{1}{\sigma_\beta^2}\Omega^{-1}\beta_0) \\ &= (\frac{1}{\sigma^2}X'X + \frac{1}{\sigma_\beta^2}\Omega^{-1})^{-1}\sigma^2\sigma^{-2}(\frac{1}{\sigma^2}X'Xb + \frac{1}{\sigma_\beta^2}\Omega^{-1}\beta_0) \\ &= (X'X + \frac{\sigma^2}{\sigma_\beta^2}\Omega^{-1})^{-1}(X'Xb + \frac{\sigma^2}{\sigma_\beta^2}\Omega^{-1}\beta_0). \end{aligned} \tag{7.14}$$

Thus the ordinary ridge estimator has the following Bayesian interpretation. If

$$\beta \sim N(0, \sigma_\beta^2 I), \text{ and } Y \sim N(X\beta, \sigma^2 I) \tag{7.15}$$

then,using equation (7.14), the posterior mean of β is given by

$$\begin{aligned} b^* &= [X'X + \frac{\sigma^2}{\sigma_\beta^2}I]^{-1}(X'Xb) \\ &= [I + \frac{\sigma^2}{\sigma_\beta^2}(X'X)^{-1}]^{-1}b, \end{aligned} \tag{7.16}$$

which is the ordinary ridge estimator with $k = \frac{\sigma^2}{\sigma_\beta^2}$. Thus the Bayesian interpretation of the estimator $b(k)$ assumes a prior mean vector of zero and a prior constant variance of σ_β^2 for each component of β .

For the generalized ridge estimator we would consider the prior

$$\beta \sim N(0, \sigma^2 GK^{-1}G'), \tag{7.17}$$

and

$$Y \sim N(X\beta, \sigma^2 I). \tag{7.18}$$

Then the posterior mean, bayes estimator of β , would be

$$b^* = [X'X + GKG']X'Xb, \tag{7.19}$$

which is the generalized ridge estimator . Unfortunately, it is usually the case that σ^2, and σ_β^2 are usually unknown.

Lawless and Wang (1976) proposed the following operational estimator of

$$k = \frac{\sigma^2}{\sigma_\beta^2}, \tag{7.20}$$

given by

$$k_{LW} = \frac{ps^2}{\sum_{i=1}^{p} \lambda_i a_i^2}. \tag{7.21}$$

Assuming that

$$\alpha \sim N(0, \sigma_\beta^2 I), \tag{7.22}$$

they argued that unconditionally, the expectation of $\lambda_i a_i^2$ is given by

$$\begin{aligned} E(\lambda_i a_i^2) &= E[E(\lambda_i a_i^2 \mid \alpha_i)] \\ &= E[\mathrm{Var}(\lambda_i^{\frac{1}{2}} a_i \mid \alpha_i) + E(\lambda_i^{\frac{1}{2}} a_i \mid \alpha_i)^2] \\ &= \lambda_i \frac{\sigma^2}{\lambda_i} + E(\lambda_i \alpha_i^2) \\ &= \sigma^2 + \mathrm{Var}(\lambda_i^{\frac{1}{2}} \alpha_i) + E(\lambda_i^{\frac{1}{2}} \alpha_i)^2 \\ &= \sigma^2 + \lambda_i \sigma_\beta^2. \end{aligned} \tag{7.23}$$

Thus

$$\begin{aligned} \sum_{i=1}^{p} E\left(\frac{\lambda_i a_i^2}{\sigma^2}\right) &= \frac{1}{\sigma^2} \sum_{i=1}^{p} (\sigma^2 + \lambda_i \sigma_\beta^2) \\ &= p + \sum_{i=1}^{p} \frac{\lambda_i \sigma_\beta^2}{\sigma^2} \\ &= p + p\frac{\sigma_\beta^2}{\sigma^2}, \end{aligned} \tag{7.24}$$

since $X'X$ is in correlation form, $\mathrm{Tr}(X'X) = p$.

Using the relation that

$$\sum_{i=1}^{p} E\left(\frac{\lambda_i a_i^2}{p\sigma^2}\right) - 1 = \frac{\sigma_\beta^2}{\sigma^2}, \tag{7.25}$$

they choose the operational biasing parameter

$$k = \frac{ps^2}{\sum_{i=1}^{p} \lambda_i a_i^2}, \tag{7.26}$$

to be a reasonable estimator of

$$\frac{\sigma^2}{\sigma_\beta^2}. \tag{7.27}$$

In numerous simulations, Lawless and Wang (1976), Wichern and Churchill (1978), Galarneau (1981), the Lawless and Wang estimator has performed well with respect to the least squares estimator .

The Lindley and Smith Estimator .

Taking a Bayesian approach, Lindley and Smith (1972) proposed the following estimator of $k = \frac{\sigma^2}{\sigma_\beta^2}$. Here the prior for β is given as $\beta \sim N(0, \sigma_\beta^2 I)$, and $Y \sim N(X\beta, \sigma^2 I)$ where σ^2 and σ_β^2 are both unknown. Assuming that σ^2 and σ_β^2 are independent each having an inverse chi-squared distribution; that is, $\nu\eta/\sigma^2 \sim \chi_\nu^2$ and $\nu_\beta\eta_\beta/\sigma_\beta^2 \sim \chi_{\nu\beta}^2$, they give the posterior distribution of β, σ^2, and σ_β^2 as proportional to

$$\begin{aligned}(\sigma^2)^{-\frac{1}{2}(n+\nu+2)} exp[-\frac{1}{2\sigma^2}(\nu\eta + (Y - X\beta)'(Y - X\beta)\ )] \\ \times (\sigma_\beta^2)^{-\frac{1}{2}(p+\nu_\beta+2)} exp[-\frac{1}{2\sigma_\beta^2}(\nu_\beta\eta_\beta + \beta'\beta\ )].\end{aligned} \tag{7.28}$$

From the mode of the posterior distribution , the following equations are used to select k .

$$b_{LS} = (X'X + (s^2/s_\beta^2)I)^{-1}X'Y, \tag{7.29}$$

$$s^2 = [\nu\eta + (Y - Xb_{LS})'(Y - Xb_{LS})]/(n + \nu + 2), \tag{7.30}$$

$$s_\beta^2 = (\nu_\beta\eta_\beta + b'_{LS}b_{LS})/(p + \nu_\beta + 2). \tag{7.31}$$

To use this estimator, begin with $k = \frac{s^2}{s_\beta^2} = 0$ in (7.29). Use the initial estimates $b_{LS} = b$ in (7.30) and (7.31) to obtain s^2 and s_β^2 . The procedure is repeated until convergence is obtained. Gibbons (1981) uses a 10^{-4} convergence criterion for $k = s^2/s_\beta^2$, which defaults to least squares if convergence is not obtained. When the solution is insensitive to small positive values of ν and ν_β , they can both be set to zero.

Chapter 8

THEORETICAL AND OPERATIONAL GENERALIZED RIDGE ESTIMATORS.

Estimating The Acceptable Range Of K

We define the acceptable range for k to be the interval that the following inequality holds

$$\text{MSE } b(k) < \text{MSE } b. \tag{8.1}$$

Using this definition, we define the acceptable interval of k to be given as

$$0 < k < k_{\max}. \tag{8.2}$$

For example, using Theobalds (1974) result, for the ordinary ridge estimator, where $K = kI$, we have the following acceptable interval where

$$\text{MSE } b - \text{MSE } b(k) > 0, \tag{8.3}$$

if

$$0 < k < k_{\max}, \tag{8.4}$$

where

$$k_{\max} = 2\sigma^2/\beta'\beta. \tag{8.5}$$

In the case of the generalized ridge estimator however, we have a separate acceptable interval for each k_i given by

$$0 < k_i < k_{i,\max}. \tag{8.6}$$

Correspondingly, the shrinkage factors given by $\delta_i = \lambda_i/(\lambda_i + k_i)$ will also have separate acceptable ranges given by

$$\delta_{i,min} < \delta_i < 1. \tag{8.7}$$

Now the MSE of the least squares estimator is unchanged if we are using the canonical model and uncorrelated components since

$$\begin{aligned} \text{MSE } b &= E[(b-\beta)'(b-\beta)] \\ &= E[(b-\beta)'GG'(b-\beta)] \\ &= E[(G'b-G'\beta)'(G'b-G'\beta)] \\ &= E[(a-\alpha)'(a-\alpha)] \\ &= \text{MSE } a. \end{aligned} \tag{8.8}$$

Considering the difference in MSE between the least squares estimator and the generalized ridge estimator of the uncorrelated components α we have

$$\begin{aligned} &\text{MSE } a - \text{MSE } a(K) \\ &= \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} - \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1}\delta_i^2 - \sum_{i=1}^{p} \alpha_i^2(\delta_i - 1)^2. \end{aligned} \tag{8.9}$$

In terms of the individual uncorrelated components a_i we have

$$\begin{aligned} &\text{MSE } a_i - \text{MSE } a_i(K) \\ &= \sigma^2\lambda_i^{-1} - \sigma^2\lambda_i^{-1}\delta_i^2 - \alpha_i^2(\delta_i - 1)^2 \\ &= \sigma^2\lambda_i^{-1}(1-\delta_i^2) - \alpha_i^2(\delta_i - 1)^2 \\ &= (1-\delta_i)[\sigma^2\lambda_i^{-1}(1+\delta_i) - \alpha_i^2(1-\delta_i)] > 0. \end{aligned} \tag{8.10}$$

Considering the second factor we have

$$\text{MSE } a_i - \text{MSE } a_i(K) > 0, \tag{8.11}$$

if

$$\delta_i[\sigma^2\lambda_i^{-1} + \alpha_i^2] > -\sigma^2\lambda_i^{-1} + \alpha_i^2, \tag{8.12}$$

or

$$\delta_i > \frac{\alpha_i^2 - \sigma^2\lambda_i^{-1} + \sigma^2\lambda_i^{-1} - \sigma^2\lambda_i^{-1}}{\sigma^2\lambda_i^{-1} + \alpha_i^2} = 1 - 2\sigma^2(\sigma^2 + \lambda_i\alpha_i^2). \tag{8.13}$$

Now since $\delta_i = \lambda_i(\lambda_i + k_i)^{-1}$ is non-negative we set

$$\delta_{i,\min} = \begin{cases} 0, & \text{if } \alpha_i^2 \leq \sigma^2\lambda_i^{-1}; \\ 1 - \frac{2\sigma^2}{\sigma^2 + \lambda_i\alpha_i^2}, & \text{otherwise.} \end{cases} \tag{8.14}$$

Thus our minimum shrinkage factors are given by

$$\delta_{i,\min} = \left[\max(0, 1 - \frac{2\sigma^2}{\sigma^2 + \lambda_i\alpha_i^2})\right], \tag{8.15}$$

giving our acceptable k_i from

$$\delta_{i,\min} = \frac{\lambda_i}{(\lambda_i + k_{i,\max})}. \tag{8.16}$$

Thus we have

$$k_{i,\max} = \begin{cases} \infty, & \text{if } \alpha_i^2 \leq \sigma^2\lambda_i^{-1}; \\ 2\sigma^2(\alpha_i^2 - \sigma^2\lambda_i^{-1})^{-1}, & \text{otherwise.} \end{cases} \tag{8.17}$$

This result was given by Vinod (1978).

Optimal Values of K for the Generalized Ridge Estimator.

The optimal values for the biasing parameters K will be those values which minimize the MSE $a(K)$. We will consider the MSE $a(K)$ in terms of the shrinkage factors δ_i where

$$\begin{aligned} \text{MSE } a(K) &= E[(a(K)-\alpha)'(a(K)-\alpha)] \\ &= \sigma^2 \sum_{i=1}^{p} \delta_i^2 \lambda_i^{-1} + \sum_{i=1}^{p} (\delta_i - 1)^2 \alpha_i^2. \end{aligned} \tag{8.18}$$

The first order condition for a minimum of the $i-th$ component of MSE $a(K)$ is that its derivative with respect to δ_i be zero. We note that

$$\begin{aligned} &\frac{d\text{MSE } a_i(K)}{dk} = \frac{d\text{MSE } a_i(K)}{d\delta_i}\frac{d\delta_i}{dk} = 0, \\ &\text{when } \frac{d\text{MSE } a_i(K)}{d\delta_i} = 0, \text{ since} \frac{d\delta_i}{dk} \neq 0. \end{aligned} \tag{8.19}$$

Thus we set

$$\begin{aligned} \frac{d\text{MSE } a_i(K)}{d\delta_i} &= 2\sigma^2 \delta_i \lambda_i^{-1} + 2(\delta_i - 1)\alpha_i^2 \\ &= \delta_i(2\sigma^2 \lambda_i^{-1} + 2\alpha_i^2) - 2\alpha_i^2 = 0 \end{aligned} \tag{8.20}$$

giving the first order condition for a minimum to be

$$\begin{aligned} \delta_i &= \alpha_i^2(\sigma^2 \lambda_i^{-1} + \alpha_i^2)^{-1} \\ &= \alpha_i^2\left(\left[\frac{\lambda_i}{\alpha_i^2}\frac{\alpha_i^2}{\lambda_i}\right](\sigma^2 \lambda_i^{-1} + \alpha_i^2)\right)^{-1} \\ &= \frac{\lambda_i}{\alpha_i^2}\alpha_i^2\left(\frac{\lambda_i}{\alpha_i^2}(\sigma^2 \lambda_i^{-1} + \alpha_i^2)\right)^{-1} \\ &= \lambda_i\left(\frac{\sigma^2}{\alpha_i^2} + \lambda_i\right)^{-1}. \end{aligned} \tag{8.21}$$

The corresponding optimal MSE values for k_i are given by using the relation

$$\delta_i = \frac{\lambda_i}{(\lambda_i + k_i)} = \lambda_i\left(\frac{\sigma^2}{\alpha_i^2} + \lambda_i\right)^{-1}, \tag{8.22}$$

and solving for k_i giving

$$k_i = \frac{\sigma^2}{\alpha_i^2}. \tag{8.23}$$

Methods To Select The k_i Parameter Of The Generalized Ridge Estimator.

Hoerl and Kennards First Iteration .

Using the least squares estimates of σ^2 and β to estimate the optimal $k_i's$, we have

$$k_i = \frac{s^2}{a_i^2}. \tag{8.24}$$

Hoerl and Kennards Iterative Proceedure

To estimates the optimal k_i, Hoerl and Kennard (1970 a) suggested the following procedure to compute k_i . Since the least squares estimator a tends to overestimate α they recommended an iterative procedure. Beginning with the initial estimates

$$k_i^0 = \frac{s^2}{a_i^2} \text{ for } i = 1, 2, \ldots, p, \tag{8.25}$$

the first iteration generalized ridge estimator are computed from

$$a(K^0) = (\Lambda + K^0)^{-1} G' X' Y, \tag{8.26}$$

where

$$K^0 = \text{diag}(k_1^0, k_2^0, \ldots, k_p^0). \tag{8.27}$$

These initial generalized ridge estimator $a(K^0)$ are then used to revise he estimates of k_i where

$$k_i^1 = \frac{s^2}{(a(K^0)_i)^2} \text{ for } i = 1, 2, \ldots, p \tag{8.28}$$

at the $(j-th+1)$ iteration we have

$$a(K^j) = (\Lambda + K^j)^{-1} G' X' Y \tag{8.29}$$

where

$$k_i^j = \frac{s^2}{(a(K^{j-1})_i)^2} \text{ for } i = 1, 2, \ldots, p. \tag{8.30}$$

The iterative process should continue until stable parameter estimates result. The final generalized ridge estimator $b(K)$ are obtained by the relation

$$b(K) = G' a(K). \tag{8.31}$$

Hemmerle's and Brantle's estimator.

Considering that $E(a_i^2 - s^2 \lambda_i^{-1}) = \alpha_i^2$, we have the following rule to select k_i

$$k_i = \frac{s^2}{a_i^2 - s^2 \lambda_i^{-1}}. \tag{8.32}$$

Hemmerle's Fully Iterative procedure for Generalized Ridge estimates.

Hemmerle (1975), showed that Hoerl and Kennards iterative procedure for estimating the biasing parameters k_i has an explicit closed form solution so that in general, iteration is unnecessary. Hocking (1976) showed that Hemmerle's result is to choose a shrinkage coefficient e_i such that

$$e_i = \begin{cases} 0, & \text{if } \tau_i^2 < 4; \\ .5 + [.25 - (1/\tau_i^2)], & \text{otherwise.} \end{cases} \tag{8.33}$$

where

$$a(K)_i = e_i a_i \text{ for } a = G'b, \tag{8.34}$$

and

$$\tau_i^2 = \frac{a_i^2 \lambda_i}{s^2}. \tag{8.35}$$

Here τ_i is the t-statistic associated with the $i-th$ regressor. Thus if the t-statistic is small, the corresponding generalized ridge estimate will be set to zero. However, if the t-statistic is large, the ridge coefficient will be a fraction e_i of the least squares coefficient. In short, this means that insignificant coefficients will be shrunk to zero while the significant coefficients will be reduced less severly. Hemmerle noted that the fully iterated generalized ridge estimator may give a poor fit through the introduction of too much bias. This led him to consider a modification of his shrinkage fractions e_i, taking into account a constraint on the total reduction of R^2. His modified shrinkage fractions $\hat{e}_i$ are given by

$$\hat{e}_i = 1 - \sqrt{m}(1 - e_i) \tag{8.36}$$

where m is the ratio of the allowable loss in R^2 if e_i is used. Hocking, et. all. (1976) objected to the use of the modified shrinkage fractions $\hat{e}_i$ since they would force all $a(K)_i$ to be non-zero and thus retain the strong influence of a small eigenvalue on the variance inflation factors.

Chapter 9

DOMINANCE FOR STOCHASTIC K

In this chapter we will review the conditions for MSE dominance of a few generalized ridge estimators. In particular the results from Vinod, Ullah, and Kadyala (1979) for the double f class generalized ridge estimator, and the results from Dwivedi, Srivastava, and Hall (1980) will be considered.

Double f class Generalized Ridge estimators.

We define an operational estimator, to be any formula for the biasing parameter k_i, which does not depend on unknown parameters. Thus for the generalized ridge estimator, we have the following operational k_i.

Hoerl and Kennard (1970) first iteration, where

$$(K_{HKFI})_i = \frac{s^2}{a_i^2}. \tag{9.1}$$

Vinod (1977), proposed the following operational k_i, based on his upper bound.

$$(K_{UB})_i = \frac{2s^2}{a_i^2 - s^2\lambda_i^{-1}}. \tag{9.2}$$

Hemmerle and Brantle (1978) suggested the following operational k_i , derived

by minimizing an unbiased estimate of MSE given by

$$(K_{HB})_i = \frac{s^2}{a_i^2 - s^2\lambda_i^{-1}}. \tag{9.3}$$

Another operational k_i , given by Vinod (1977), based on Stein's unbiased estimate of MSE is given by

$$(K_V)_i = \frac{2s^2}{a_i^2 - 2s^2\lambda_i^{-1}}. \tag{9.4}$$

The above operational k_i are all stochastic since they are functions of the least square estimates s^2, a_i^2 . However they are also members of the double f class family of k_i given by

$$(K_{DFC})_i = (K_{f_1,f_2})_i = \frac{f_1 s^2}{a_i^2 - f_2 s^2\lambda_i^{-1}}, \tag{9.5}$$

where f_1, f_2 are arbitrary scalars. Thus

$$(K_{1,0})_i = \frac{s^2}{a_i^2}. \tag{9.6}$$

gives Hoerl and Kennards first iteration.

$$(K_{2,1})_i = \frac{2s^2}{a_i^2 - s^2\lambda_i^{-1}}. \tag{9.7}$$

gives Vinod's upper bound.

$$(K_{1,1})_i = \frac{s^2}{a_i^2 - s^2\lambda_i^{-1}}. \tag{9.8}$$

gives Hemmerle and Brantle estimator.

$$(K_{2,2})_i = \frac{2s^2}{a_i^2 - 2s^2\lambda_i^{-1}}. \tag{9.9}$$

gives Vinod's unbaised estimator.

Using $(K_{f_1,f_2})_i$ we may write the shrinkage factors δ_i as

$$\begin{aligned}
\delta_i &= \left(\frac{\lambda_i}{(\lambda_i + k_i)}\right) \\
&= \left(\frac{\lambda_i}{\lambda_i + \frac{f_1 s^2}{a_i^2 - \frac{f_2 s^2}{\lambda_i}}}\right) \\
&= \left(\frac{\lambda_i}{\lambda_i + \frac{\lambda_i f_1 s^2}{\lambda_i a_i^2 - f_2 s^2}}\right) \\
&= \left(\frac{\lambda_i a_i^2 - f_2 s^2 + f_1 s^2 - f_1 s^2}{\lambda_i a_i^2 - f_2 s^2 + f_1 s^2}\right) \\
&= \left(1 - \frac{f_1 s^2}{\lambda_i a_i^2 + (f_1 - f_2)s^2}\right).
\end{aligned} \tag{9.10}$$

Considering our generalized ridge estimator $b(K)$ we have

$$\begin{aligned}
b(K) &= (X'X + GKG')^{-1}X'y \\
&= G(\Lambda + K)^{-1}G'X'y \\
&= G(\Lambda + K)^{-1}G'X'X(X'X)^{-1}X'y \\
&= G(\Lambda + K)^{-1}G'G\Lambda G'b \\
&= G\Delta G'b \\
&= G\Delta a,
\end{aligned} \tag{9.11}$$

$a(K) = G'b(K) = G'Ga = \Delta a$. Thus the family of double f-class estimators is given by

$$a(DFC) = a(K)_{f_1,f_2} = \hat{\Delta} a, \tag{9.12}$$

where

$$\hat{\Delta} = \text{diag}(\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_p), \tag{9.13}$$

and

$$\hat{\delta}_i = \left[1 - \frac{f_1 s^2}{\lambda_i a_i^2 + (f_1 - f_2)s^2}\right], \tag{9.14}$$

for arbitrary scalars f_1, f_2 and operational k_i. Since the MSE $b(K)$ = MSE $a(K)$ the MSE properties of the double f-class estimator will be considered through the $i-th$ uncorrelated component $a(DFC)$ given by

$$a(DFC)_i = \hat{\delta}_i a_i = \left[1 - \frac{f_1 s^2}{\lambda_i a_i^2 + (f_1 - f_2)s^2}\right] a_i. \tag{9.15}$$

Now at $f_1 = 0$, we have

$$a(DFC)_i = a(0, f_2)_i = [1 - \frac{0}{\lambda_i a_i^2 - f_2 s^2}] a_i = a_i \tag{9.16}$$

and the double f-class estimator reduces to the least square estimator a_i at $f_1 = 0$.

At $f_1 = f_2$ we have

$$\begin{aligned} a(DFC)_i = a(f_1, f_1)_i &= \left[1 - \frac{f_1 s^2}{\lambda_i a_i^2 + (f_1 - f_1)s^2}\right] a_i \\ &= \left[a_i - \frac{f_1 s^2}{\lambda_i a_i}\right], \end{aligned} \tag{9.17}$$

and since a_i may be zero, the moments of any order do not exist for the double f-class estimator when $f_1 = f_2$.

Vinod, Ullah, and Kadyala (1979) gave the exact expressions for the bias and MSE of the double f-class generalized ridge estimator using confluent hypergeometric functions. From these exact expressions they derive asymptotic expressions

for the bias and MSE for large non-centrality parameter given by $\Theta_i = \lambda_i \alpha_i^2 / 2\sigma^2$. Their approximations of the bias and MSE expressions are given in the following two theorems.

Theorem 9.1: The asymptotic expansion for the exact bias of a component of $a(DFC)$, up to the order of σ^2, or Θ_i^{-1} is given by

$$E(a(DFC) - \alpha)_i = -\frac{f_1 \alpha_i}{2\Theta_i} = -\frac{f_1 \alpha_i \sigma^2}{\lambda_i \alpha_i^2} = -\frac{f_1 \sigma^2}{\lambda_i \alpha_i} \tag{9.18}$$

from here we can see that the double f-class estimator will be biased in a direction opposite the sign of α_i.

Theorem 9.2: The asymptotic expansion of the exact MSE of a component of $a(DFC)$, up to the order of σ^6, or Θ_i^{-3} is given by

$$E(a(DFC) - \alpha)_i^2 = \frac{\sigma^2}{\lambda_i} + \frac{f_1 \alpha_i^2}{4\Theta_i^2(\nu)} A_1 + \frac{f_1 \alpha_i^2}{8\Theta_i^3(\nu)} [3A_1 - 2(f_1 - f_2)A_2], \tag{9.19}$$

where

$$A_1 = f_1(\nu + 2) \quad , \tag{9.20}$$

$$A_2 = \frac{n+2}{n}[f_1(\nu + 4) + 3\nu], \tag{9.21}$$

and $(f_1 - f_2) > 0$ for the existence of the Bias and MSE. Both theorems assume a large non-centrality parameter $\Theta_i = \frac{\lambda_i \alpha_i^2}{2\sigma^2}$. Substituting for the non-centrality parameter we can write $E(a(DFC) - \alpha)_i^2$ as

$$\begin{aligned} \text{MSE } a(DFC)_i &= \frac{\sigma^2}{\lambda_i} + \frac{f_1 \sigma^4}{\lambda_i^2 \alpha_i^2(\nu)} A_1 + \frac{f_1 \sigma^6}{\alpha_i^4 \lambda_i^3(\nu)} [3A_1 - 2(f_1 - f_2)A_2] \\ &= \frac{\sigma^2}{\lambda_i} + f_1 P_1 A_1 + f_1 P_2 A_1 - f_1(f_1 - f_2) P_3 P_4 A_3, \end{aligned} \tag{9.22}$$

where

$$
\begin{aligned}
P_1 &= \frac{\sigma^4}{(\nu)\lambda_i^2\alpha_i^2}, \\
P_2 &= \frac{3\sigma^6}{(\nu)\lambda_i^3\alpha_i^4}, \\
P_3 &= \frac{6\sigma^6}{(\nu)\lambda_i^3\alpha_i^4}, \\
P_4 &= \frac{(\nu+2)}{(\nu)}, \\
A_3 &= f_1(\nu+4)+3(\nu),
\end{aligned}
\tag{9.23}
$$

and $P_i > 0$ for $i = 1, 2, \ldots, 4$. Now MSE $a_i = \frac{\sigma^2}{\lambda_i}$, so the double f-class estimater will dominate the least squares estimator if

$$\text{MSE } a_i - \text{MSE } a(DHC)_i = -f_1 P_1 A_1 - f_1 P_2 A_1 + f_1(f_1 - f_2)P_3 P_4 A_3 > 0, \tag{9.24}$$

or equivalently if

$$f_1(f_1 - f_2) > f_1 \frac{A_1}{A_3}\frac{1}{P_4}\left(\frac{P_1 + P_2}{P_3}\right), \tag{9.25}$$

$$f_1(f_1 - f_2) > f_1 \frac{A_1}{A_3}\frac{1}{P_4}(\Theta_i + 3/2), \tag{9.26}$$

for $A_3 > 0$. We may further simplify the inequality by assuming large n and fixing f_1 to be positive. This gives

$$\frac{A_1}{A_3 P_4} = \frac{f_1(\nu+2)+2n}{f_1(\nu+4)+3n}\left(\frac{n}{\nu+2}\right) \approx \frac{f_1+2}{f_1+3}, \tag{9.27}$$

so our condition for dominance over least squares, for large n, is given by

$$\left(\frac{f_1+2}{f_1+3}\right)(\Theta_i + 3/2) < (f_1 - f_2). \tag{9.28}$$

Considering the Hoerl and Kennard (1970) first iteration choice $f_1 = 1$, $f_2 = 0$ the condition for superiority over the least squares estimator for large n becomes

$$(\Theta_i + 3/2)\frac{3}{4} < 1, \tag{9.29}$$

which will not hold since $\Theta_i \geq 0$.

Hemmerle and Brantle's choice $f_1 = f_2 = 1$ or Vinod's choice of $f_1 = f_2 = 2$ would give an infinite MSE . In the present context of large n the condition for dominance over least squares estimator becomes

$$(\Theta_i + 3/2)\frac{3}{4} < 0 \text{ for } f_1 = f_2 = 1 , \tag{9.30}$$

$$(\Theta_i + 3/2)\frac{4}{5} < 0 \text{ for } f_1 = f_2 = 2 , \tag{9.31}$$

neither condition being true since $\Theta_i \geq 0$.

For Vinod's upper bound choice $f_1 = 2,\ f_2 = 1$ the condition for dominance becomes

$$(\Theta_i + 3/2)\frac{4}{5} < 1 , \tag{9.32}$$

which again is not possible.

A favorable region of the parameter space for large n is given by

$$\Theta_i < (f_1 - f_2)\frac{(f_1 + 3)}{(f_1 + 2)} - \frac{3}{2}. \tag{9.33}$$

Fixing $f_1 = 1,\ f_2$ must satisfy

$$\Theta_i < -\frac{1}{6} - \frac{4}{3}f_2, \tag{9.34}$$

or

$$2\Theta_i < -\frac{1}{3} - \frac{8}{3}f_2. \tag{9.35}$$

Considering

$$2\Theta_i = \frac{\alpha_i^2 \lambda_i}{\sigma^2}, \tag{9.36}$$

to be the true unknown value of the corresponding F statistic given by

$$F_i = \frac{\lambda_i a_i^2}{s^2} \sim F_{1,\nu}, \tag{9.37}$$

Vinod and Ullah (1981) suggest replacing $2\Theta_i$ with a tabulated $F_{\alpha\ 1,\nu}$ of the F statistic and solving for f_2 giving the condition

$$f_2 \leq -\frac{3(F_{\alpha\ 1,\nu}) + 1}{8}, \tag{9.38}$$

which will provide an operational generalized ridge estimator that will have a lower MSE than the least squares estimator for $(1-\alpha)100\%$ of the region along each dimension of the parameter space for large n. These conditions for dominance have the drawback however that they require the assumptions of large n, and large non-centrality parameter Θ_i .

Finite Sample Properties of a Ridge Estimator.

The results of Dwivedi,Srivastava, and Hall (1980) , apply to the Hoerl and Kennard first iteration, where

$$k_i = \frac{s^2}{a_i^2}. \tag{9.39}$$

In particular, they derived the first and second moments of the estimator using the following assumptions.

$$a \sim N(\alpha, \sigma^2 \Lambda^{-1}), \tag{9.40}$$

$$z_i = \frac{\sqrt{\lambda_i} a_i}{\sigma} \sim N(\frac{\sqrt{\lambda_i}\alpha_i}{\sigma}, 1), \tag{9.41}$$

$$\rho = \frac{\nu s^2}{\sigma^2} \sim \chi^2_\nu, \tag{9.42}$$

where ρ is independent of z_i,

$$\frac{z_i^2}{\rho} \sim F'(1, \nu, \frac{\lambda_i a_i^2}{\sigma^2}), \tag{9.43}$$

where F' is the noncentral F with non-centrality parameter $\frac{\lambda_i a_i^2}{\sigma^2}$. They write the i-th uncorrelated generalized ridge estimator as

$$\begin{aligned}
(a(K)_{HKFI})_i = \hat{\Delta}_i a_i &= \left(\frac{\lambda_i}{\lambda_i + \frac{s^2}{a_i^2}}\right) a_i \\
&= \left(\frac{\lambda_i a_i^3}{\lambda_i a_i^2 + s^2}\right) = \frac{\sigma}{\sqrt{\lambda_i}}\left(\frac{\frac{\sqrt{\lambda_i}}{\sigma}\lambda_i a_i^3}{\lambda_i a_i^2 + s^2}\right) \\
&= \frac{\sigma}{\sqrt{\lambda_i}}\left(\frac{\frac{\sqrt{\lambda_i}}{\sigma^3}\lambda_i a_i^3}{\frac{\lambda_i a_i^2}{\sigma^2} + \frac{s^2}{\sigma^2}}\right) = \frac{\sigma}{\sqrt{\lambda_i}}\left(\frac{z_i^3}{z_i^2 + \frac{\rho}{\nu}}\right) \\
&= \frac{\sigma}{\sqrt{\lambda_i}}\left(\frac{z_i^3}{(z_i^2 + \rho)}\right)\left(\frac{\nu(z_i^2 + \rho)}{(\nu z_i^2 + \rho)}\right) \\
&= \frac{\sigma}{\sqrt{\lambda_i}}\left(\frac{z_i^3}{(z_i^2 + \rho)}\right)\left[\frac{\nu z_i^2 + \nu\rho - \nu\rho + \rho}{\nu(z_i^2 + \rho)}\right]^{-1} \\
&= \frac{\sigma}{\sqrt{\lambda_i}}\left(\frac{z_i^3}{(z_i^2 + \rho)}\right)\left[1 - \left(\frac{\nu - 1}{\nu}\right)\left(\frac{\rho}{z_i^2 + \rho}\right)\right]^{-1},
\end{aligned} \tag{9.44}$$

and since

$$\left(\frac{\nu - 1}{\nu}\right)\left(\frac{\rho}{z_i^2 + \rho}\right) < 1, \tag{9.45}$$

we have

$$\begin{aligned}
(a(K)_{HKFI})_i &= \frac{\sigma}{\sqrt{\lambda_i}}\left(\frac{z_i^3}{(z_i^2 + \rho)}\right)\sum_{j=0}^{\infty}\left(\frac{\nu - 1}{\nu}\right)^j\left(\frac{\rho}{z_i^2 + \rho}\right)^j \\
&= \frac{\sigma}{\sqrt{\lambda_i}}\sum_{j=0}^{\infty}\left(\frac{\nu - 1}{\nu}\right)^j \frac{z_i^3 \rho^j}{(z_i^2 + \rho)^{j+1}},
\end{aligned} \tag{9.46}$$

Similarly for the second moment about the origin we have

$$\begin{aligned}
(a(K)_{HKFI})_i^2 &= \frac{\sigma^2}{\lambda_i}\left(\frac{z_i^6}{(z_i^2 + \rho)^2}\right)\left[1 - \left(\frac{\nu - 1}{\nu}\right]\left(\frac{\rho}{z_i^2 + \rho}\right)\right]^{-2} \\
&= \frac{\sigma^2}{\lambda_i}\left(\frac{z_i^6}{(z_i^2 + \rho)^2}\right)\sum_{j=0}^{\infty}(j+1)\left(\frac{\nu - 1}{\nu}\right)^j\left(\frac{\rho}{(z_i^2 + \rho)}\right)^j \\
&= \frac{\sigma^2}{\lambda_i}\sum_{j=0}^{\infty}(j+1)\left(\frac{\nu - 1}{\nu}\right)^j\left(\frac{z_i^6 \rho^j}{(z_i^2 + \rho)^{j+2}}\right).
\end{aligned} \tag{9.47}$$

Taking the expectations of both moments we have

$$E[(a(K)_{HKFI})_i] = \frac{\sigma}{\sqrt{\lambda_i}} \sum_{j=0}^{\infty} (\frac{\nu-1}{\nu})^j E[\frac{z_i^3 \rho^j}{(z_i^2+\rho)^{j+1}}]. \tag{9.48}$$

$$E[(a(K)_{HKFI})_i^2] = \frac{\sigma^2}{\lambda_i} \sum_{j=0}^{\infty} (j+1)(\frac{\nu-1}{\nu})^j E[\frac{z_i^6 \rho^j}{(z_i^2+\rho)^{j+2}}]. \tag{9.49}$$

Dwivedi, Srivastava, and Hall showed that the expectations, for $z_i \sim N(\Theta,1)$ and independent $\rho \sim \chi^2_\nu$, with integers m,q,r are given by

$$E[\frac{z_i^m \rho^q}{(z_i^2+\rho)^r}] = \frac{2^{q-r+\frac{m}{2}} \Gamma(q+\frac{\nu}{2}) e^{-\frac{\Theta^2}{2}}}{\Gamma(\frac{\nu}{2})} \times \sum_{j=0}^{\infty} \frac{\Gamma(q-r+j+\frac{m+\nu+1}{2})\Gamma(j+\frac{m+1}{2})}{\Gamma(q+j+\frac{m+\nu+1}{2})\Gamma(j+\frac{1}{2})} \frac{(\frac{\Theta^2}{2})^j}{j!} \quad \text{for even } m \tag{9.50}$$

$$\text{or} \quad = \frac{2^{q-r+m-\frac{1}{2}} \Theta \Gamma(q+\frac{\nu}{2}) e^{-\frac{\Theta^2}{2}}}{\Gamma(\frac{\nu}{2})} \times \sum_{j=0}^{\infty} \frac{\Gamma(q-r+j+1+\frac{m+\nu}{2})\Gamma(j+1+\frac{m}{2})}{\Gamma(q+j+1+\frac{m+\nu}{2})\Gamma(j+\frac{3}{2})} \frac{(\frac{\Theta^2}{2})^j}{j!} \quad \text{for odd } m \tag{9.51}$$

Thus they derived the first and second moments of $(a(K)_{HKFI})_i$ from which they evaluated the BIAS and MSE for differing values of $\Theta_i = \frac{\lambda_i \alpha_i^2}{\sigma^2}$ and ν. From their results they gave the following observations.

1) The efficientcy of a relative to $(a(K)_{HKFI})_i$ given by $\frac{\text{MSE}_{(a(K)_{HKFI})_i}}{\text{MSE}_a} \times 100$ is a function of ν and the non-centrality parameter $\Theta_i = \frac{\lambda_i \alpha_i^2}{\sigma^2}$ which is unknown.

2) The $i-th$ component of the generalized ridge estimator $(a(K)_{HKFI})_i$ is biased in the direction opposite to a_i.

3) The absolute value of the relative BIAS, where the relative bias is given as

$$E\left(\frac{(a(K)_{HKFI})_i - \alpha_i}{\alpha_i}\right), \tag{9.58}$$

is a decreasing function of Θ and an increasing function of ν.

4) The $i-th$ component of the generalized ridge estimator $(a(K)_{HKFI})_i$ will dominate the least squares estimator a_i under the MSE criterion when the non-centrality parameter $\Theta_i = \frac{\lambda_i \alpha_i^2}{\sigma^2} \leq 2$.

Chapter 10

DEVELOPMENTS IN GENERALIZED RIDGE ESTIMATORS

A Pre-Test Estimator.

Using the results of Dwivedi, Srivastava, and Hall (1980), Srivastava and Giles (1984), developed a pre-test generalized ridge estimator . In particular they use the result that the Hoerl and Kennard first iteration estimator, given by

$$a(HKFI)_i = \frac{\lambda_i}{(\lambda_i + k_i))} a_i \ , \text{ where } k_i = \frac{s^2}{a_i^2} \ , \tag{10.1}$$

will dominate the least squares estimator a_i if $\Theta_i = \frac{\lambda_i \alpha_i^2}{\sigma^2} \leq 2$.

Under the same assumptions used by Dwivedi et.al., Srivastava and Giles (1984), proposed the statistic

$$W = \frac{z_i^2}{z_i^2 + \rho} = \left(\frac{\frac{a_i^2 \lambda_i}{\sigma^2}}{\frac{a_i^2 \lambda_i}{\sigma^2} + \frac{\nu s^2}{\sigma^2}} \right) = \left(\frac{a_i^2 \lambda_i}{a_i^2 \lambda_i + \nu s^2} \right), \tag{10.2}$$

which has a non-central beta distribution, with non-centrality parameter $\Theta_i = \frac{\lambda_i \alpha_i^2}{\sigma^2}$, given by

$$f_{\Theta_i}(W) = e^{-\Theta_i} \sum_{j=0}^{\infty} \frac{(\Theta_i)^j}{j!} \frac{\Gamma(j + \frac{1}{2} + \frac{\nu}{2})}{\Gamma(j + \frac{1}{2})\Gamma(\frac{\nu}{2})} W^{j - \frac{1}{2}} (1 - W)^{\frac{\nu}{2} - 1}. \tag{10.3}$$

They suggest the following test of the hypothesis that

$$\text{MSE } a(HKFI)_i < \text{MSE } a_i. \tag{10.4}$$

The null and alternate hypothesis are given by

$$H_0 : \Theta_i \leq 2, \qquad H_a : \Theta_i > 2. \tag{10.5}$$

Using the test statistic W , the rejection criterion is to reject H_0 if

$$W \geq W_\alpha, \tag{10.6}$$

where

$$\int_0^{W_\alpha} f_2(W)dW = (1-\alpha), \tag{10.7}$$

for a given significance level α .

Thus they define their pre-test generalized ridge estimator to be

$$a(PTGRE)_i = \begin{cases} a(HKFI)_i, & \text{if } \frac{z_i^2}{(z_i^2+\rho)} < W_\alpha; \\ a_i, & \text{otherwise.} \end{cases} \tag{10.8}$$

For this estimator they derive the BIAS and MSE , showing that they are functions of ν, Θ, α. For different values of Θ, ν, and α, they evaluate the relative BIAS , and relative MSE of the PTGRE. Concerning the relative efficiency, they evaluated the PTGRE with respect to both the least squares estimator and the HKFI generalized ridge estimator for different values of Θ, ν, α. The main results of their study are given as follows.

1) The HKFI generalized ridge estimator dominated the least squares estimator and the PTGRE in terms of MSE when $\Theta_i \leq 2$. However, when Θ_i is relatively large, the PTGRE outperformed the generalized ridge estimator .

2) The loss in efficiency by using PTGRE over generalized ridge estimator is very small for $\Theta_i \leq 2$. For larger values of the non-centrality parameter the efficiency gain can be as large as 30% .

3) The PTGRE is biased is the direction opposite in sign to that of the corresponding parameter α_i .

4) For a significance level of $\alpha = .05$, the PTGRE dominates the least squares estimator over large parts of the parameter space, and its efficiency with respect to the least square estimator is better than or only slightly worse than the efficiency of the HKFI estimator.

5) The main advantage of the PTGRE is that a significance level α can be attached to the hypothesis that the HKFI generalized ridge estimator is in the correct non-centrality interval.

The results of Dwivedi, Srivastava, and Hall (1980), showed that the HKFI generalized ridge estimator would dominate the least squares estimator a_i if $\Theta_i \leq 2$. However the problem was how to test that the HKFI generalized ridge estimator was in the correct non-centrality interval. The PTGRE provides an answer to this problem by using a test of the hypothesis $\Theta_i \leq 2$ which is equivalent to a test that MSE $a(HKFI)_i <$ MSE a_i .

An Almost Unbiased Ridge Estimator.

Singh, Chaubey, and Dwivedi (1986), used a Jack-knife proceedure to develop a generalized ridge estimator which has a reduced bias. In particular, considering the canonical form of the model (1), we have

$$\begin{aligned} Y &= X\beta + \varepsilon \\ &= XGG'\beta + \varepsilon \\ &= Z\alpha + \varepsilon, \end{aligned} \tag{10.9}$$

where $Z'Z = \Lambda$, $Z = XG$ is a $(n \times p)$ matrix of non-stochastic regressors such that $\lim_{n\to\infty} \frac{Z'Z}{n}$ is finite. Now

$$\text{BIAS } a(K) = -K(\Lambda + K)^{-1}\alpha = O(\frac{1}{n}), \tag{10.10}$$

under the assumption that $(Z'Z)^{-1}$ is of order $\frac{1}{n}$.

Let Z_i be the matrix Z with its i-th row, denoted by z_i , deleted, and Y_i be the observation vector with its i-th observation deleted. The ridge estimator, with its i-th observation deleted is given as

$$\begin{aligned} a(K)_{-i} &= (Z_i'Z_i + K)^{-1}Z_i'Y_i \\ &= (Z'Z - z_i'z_i + K)^{-1}Z_i'Y_i \\ &= (A - z_i'z_i)^{-1}(Z'Y - z_iy_i), \end{aligned} \tag{10.11}$$

where $A = \Lambda + K$ and from Theorem A5 in the appendix we have

$$\begin{aligned} a(K)_{-i} &= (A^{-1} + \frac{A^{-1}z_iz_i'A^{-1}}{1-w_i})(Z'Y - z_iy_i), \quad \text{where } w_i = z_i'A^{-1}z_i \\ &= a(K) - \frac{A^{-1}z_iy_i}{1-w_i} + \frac{A^{-1}z_iy_iw_i}{1-w_i} + \frac{A^{-1}z_iz_i'A^{-1}Z'Y}{1-w_i} - \frac{A^{-1}z_iw_iy_i}{1-w_i} \\ &= a(K) + \frac{A^{-1}z_i}{1-w_i}[z_i'a(K) - y_i] \\ &= a(K) - \frac{A^{-1}z_i}{1-w_i}[e_i]. \end{aligned} \tag{10.12}$$

The weighted pseudo-values are then given by

$$Q_i = a(K) + n(1 - w_i)(a(K) - a(K)_{-i}), \tag{10.13}$$

providing the weighted jacknife estimator of α as

$$\begin{aligned}
\tilde{a} &= \sum_{i=1}^{n} \frac{Q_i}{n} \\
&= a(K) + A^{-1}Z'(Y - Za(K)) \\
&= [2I - A^{-1}\Lambda]a(K) \\
&= [2I - [I - A^{-1}K]\,]a(K) \\
&= [I + A^{-1}K]\,]a(K) \\
&= [I + A^{-1}K]\,][A^{-1}\Lambda]a \\
&= [I + A^{-1}K]\,][I - KA^{-1}]a \\
&= [I - (A^{-1}K)^{-2}\,]a.
\end{aligned} \tag{10.14}$$

So that the

$$\text{Bias } \tilde{a} = -(A^{-1}K)^2\alpha = O(\frac{1}{n^2})\,, \tag{10.15}$$

is of a smaller order in magnitude than $a(K)$. Considering the difference in total squared bias, we have

$$\begin{aligned}
&\sum_{i=1}^{p}[\text{ BIAS }^2 a(K) - \text{ BIAS }^2\tilde{a}] \\
&= \alpha' A^{-1}KKA^{-1}\alpha - \alpha'(A^{-1}K)^4\alpha \\
&= \alpha' A^{-1}K[\,I - (A^{-1}K)^2]KA^{-1}\alpha > 0,
\end{aligned} \tag{10.16}$$

showing that the total squared BIAS in $\tilde{a}$ is smaller than $a(K)$.

The jack-knife technique offers a method in providing confidence intervals for $\tilde{b}$ depending on the property that

$$\sqrt{n}(\tilde{b} - \beta) \to N(0, \sigma^2\Sigma^{-1}), \text{where } \Sigma = \lim_{n\to\infty}(\frac{X'X}{n}). \tag{10.17}$$

Using a consistent estimator for the Var $\tilde{b}$ given by

$$V = \frac{1}{n(n-p)} \sum_{i=1}^{n} (Q_i - \tilde{b})(Q_i - \tilde{b})', \tag{10.18}$$

the confidence interval for $\tilde{b}_i$ is given by

$$\tilde{b}_i \pm t(1 - \frac{\alpha}{2}; n-p)\sqrt{v_{ii}}, \tag{10.19}$$

where $t(1 - \frac{\alpha}{2}; n-p)$ is the upper $\frac{\alpha}{2} \times 100\%$ point of the student's t-distibution.

Improved Ridge Estimator.

In a later paper Singh and Chaubey (1987), suggested the following improved ridge estimator $\bar{a}(K)$, given by

$$\bar{a}(K) = La + \bar{L}a(K), \tag{10.20}$$

$$\text{where } L = \text{diag}(l_1, l_2, \ldots, l_p) \text{ is a p.s.d. matrix .} \tag{10.21}$$

The constants l_i are unknown where $0 \leq l_i \leq 1$, $\bar{L} = I - L$ and the $l_i's$ are chosen such that the i-th component of the MSE $\bar{a}(K)$ is minimum.

Considering the BIAS and Var of the estimator $\bar{a}(K)$ we have

$$\begin{aligned} \bar{a}(K) &= La + \bar{L}\Delta a \\ &= [L + I - I + \bar{L}\Delta]a \\ &= [I - \bar{L} + \bar{L}\Delta]a \\ &= [I - \bar{L}(I - \Delta)]a \\ &= [I - \bar{L}\bar{\Delta}]a, \end{aligned} \tag{10.22}$$

giving

$$\text{BIAS } \bar{a} = -\bar{L}\bar{\Delta}\alpha, \quad \text{and} \quad \text{Var } \bar{a} = \sigma^2(I - \bar{L}\bar{\Delta})\Lambda^{-1}(I - \bar{L}\bar{\Delta}). \tag{10.23}$$

Considering the i-th component of the MSE of $\bar{a}$ we have

$$\text{MSE } \bar{a}(K)_i = \frac{\sigma^2}{\lambda_i}[(1 - \bar{l}_i\bar{\delta}_i)^2] + \alpha_i\bar{\delta}_i^2\bar{l}_i^2, \tag{10.24}$$

taking the minimum with respect to $\bar{l}_i$ gives

$$\frac{\partial \text{MSE } \bar{a}(K)_i}{\partial \bar{l}_i} = 2\alpha_i^2\bar{\delta}_i^2\bar{l}_i - \frac{\sigma^2}{\lambda_i}2(1 - \bar{l}_i\bar{\delta}_i)\bar{\delta}_i = 0, \tag{10.25}$$

when

$$\bar{l}_i = \frac{t_i(\lambda_i + k_i)}{k_i(\lambda_i + k_i)}, \tag{10.26}$$

where $t_i = \frac{\sigma^2}{\alpha_i^2}$, and $k_i \geq \frac{t_i}{\lambda_i}$ for $0 \leq \bar{l}_i \leq 1$.

By comparing the MSE of $\bar{a}(K)$ with $a(K)$ and $\tilde{a}$, the authors show that

1) The improved ridge estimator has smaller MSE and Bias than the generalized ridge estimator for $k_i \neq t_i$.

2) The improved ridge estimator has a larger Bias but smaller MSE than the almost unbiased ridge estimator for $\delta_i^2 \neq \frac{t_i}{t_i+\lambda_i}$.

Using operational versions of their estimators a simulation was conducted which demonstated the overall superior performance of the improved ridge estimator.

Chapter 11

DETERMINISTIC RIDGE ESTIMATORS.

Using model (1.2) in canonical form we have

$$Y = Z\alpha + \varepsilon, \tag{11.1}$$

where

$$Z = XG,\ Z'Z = \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p),\ \text{ and }\ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0. \tag{11.2}$$

Consider the class of generalized ridge estimators given by

$$a(K) = (\Lambda + K)^{-1}\Lambda a,\ \text{ for } K = \mathrm{diag}(k_1, k_2, \ldots, k_p), k_i > 0. \tag{11.3}$$

When all the biasing parameters k_i are chosen to be the same we have the ordinary ridge estimator.

Of the many ridge rules that have been proposed to select the biasing parameter k_i , many have the problem that they depend on the unknown values of β and σ^2. Many reserchers opt for the operational versions of their rules using the stochastic estimates b and s^2. However this opens up new problems since the MSE properties of the ridge estimators are only valid for fixed non-random $k_i's$. The purpose of this chapter is to consider some deterministic ridge estimators which depend only on the eigen values of the $X'X$ matrix. Several new estimators are proposed and their performance, under the MSE criterion, is investigated in a simulation.

Six ridge estimators.

The following six ridge rules were considered in this study. Rule R5 was suggested by Conniffe abd Stone (1973) and rule R1 was suggested by Lee (1986). The other four estimators are considered here for the first time.

R1: $k_i = \lambda_p$

R2: $k_i = \frac{\sqrt{\lambda_i}}{\lambda_1}$

R3: $k_i = \frac{1}{p}\lambda_p$

R4: $k_i = \frac{\sqrt{\lambda_i}}{\lambda_1}(\sqrt{\lambda_1} + \sqrt{\lambda_p} - \sqrt{\lambda_i})$

R5: $k_i = \sqrt{\lambda_i}$

R6: $k_i = \lambda_p$ if $\frac{\lambda_1}{\lambda_i} \leq 50$,

else, $k_i = p \times \lambda_p$

Models considered in the simulation.

There are three basic models considered in this study. The first, denoted by model 1, is the four factor model Hald(1952). Model 2 is a ten factor model Gorman and Toman (1966). Model 3 is a fifteen factor model McDonald and Schwing (1973). The parameters of the models are listed below.

MODEL 1

$p =$ 4

$n =$ 12

$\lambda_i =$ 2.2357, 1.5761, .1866, .0016 .

$\phi =$ 1397

MODEL 2

$p =$ 10

$n =$ 36

$\lambda_i =$ 3.6923, 1.5418, 1.2927, 1.0457, .9719, .6587, .3574, .2197, .1513, .0681.

$\phi =$ 54

MODEL 3

$p =$ 15

$n =$ 60

$\lambda_i =$ 4.5272, 2.7547, 2.0545, 1.3487, 1.2277, .9605, .6124, .4729, .3708, .2163, .1665, .1275, .1142, .0460, .0049.

$\phi =$ 924

Using the guidelines of Belsley,Kuh,Welsh,(1980) p. 104 , condition numbers less than 100 are indications of weak dependencies and condition numbers greater than 900 will correspond with moderate to strong dependencies. Thus, model 2 without any other information will be considerably less ill-conditioned than model 1 or 3.

Simulation design.

To obtain a set of coefficients, p random numbers $\{\alpha_i\}_1^P$ are chosen from a uniform distribution on $(0,1)$. The coefficients are subsequently normalized to unit length such that $\sum_{i=1}^{p} \alpha_i^2 = 1$. In all the models, six different values of σ^2 are considered making the signal to noise ratio $(SNR) = \frac{\alpha'\alpha}{\sigma^2}$ range from 1 to 10,000. In particular we have

$\sigma^{-2} =$ 1 25 100 400 2500 10,000

$SNR =$ 1 25 100 2500 10,000

For each of the models and the SNR's considered, 1000 different replicates were generated by first selecting $\alpha_i's$ and then obtaining the estimates via

$$a_i \sim N(\alpha_i, \frac{\sigma^2}{\lambda_i}),\ i = 1, 2, \ldots, p. \tag{11.4}$$

Each of the estimators R1 to R6 was then computed componentwize using

$$a(Ri)_i = (\frac{\lambda_i}{\lambda_i + k_i})a_i \tag{11.5}$$

where k_i is determined by using the rules R1 to R6, respectively. The simulation was programmed with the MAPLE4.2 package and is included in Appendix C.

Evaluation Criterion.

For each estimator, including least squares given by R0, the sum of squared error is computed using

$$L(Ri) = \sum_{i=1}^{p}(a(Ri)_i - \alpha_i)^2, \tag{11.6}$$

where $a(Ri)_i$ is given by equation (11.5). As an estimate of the MSE , we use the average squared error over the 1000 replicates given by

$$M(Ri) = \sum_{j=1}^{1000} \frac{L(Ri)_j}{1000}. \tag{11.7}$$

The average sum of squared error is reported in tables 1 to 3.

A measure of the relative improvement obtained by using a specific ridge estimator is given by

$$RM(Ri) = \frac{M(Ri)}{M(R0)} \times 100, \tag{11.8}$$

where $M(Ri)$ is the average sum of squared error of a specified ridge estimator and $M(R0)$ is the average sum of squared error of the least squares estimator. The relative improvement over least squares is reported in tables 4 to 6.

The true MSE for the least squares estimator a , given by

$$\sigma^2 \sum_{i=1}^{p} \lambda_i^{-1}, \tag{11.9}$$

is used as a check on the adequacy of the simulations.

Results of the Simulations.

1) The potential improvement of the ridge estimators is greatest when the condition number is large and the SNR is small.

2) None of the ridge estimators outperform the least squares estimator in all situations. In particular, all the ridge estimators perform quite poorly when the condition number is low $\phi = 54$, and the signal to noise ratio is high $SNR \geq 2,500$.

3) The ridge estimators R2, R4, R5, and R6, which perform very well in low SNR regions, perform poorly in high SNR regions.

4) The ridge estimators R1, and R3, which are not the best performers in low

SNR regions, outperform least squares even in high SNR regions, as long as the condition number is high, say $\phi \geq 900$.

5) From a conservative point of view, good performance in worst possible situations, the best all round ridge estimators are R1 and R3.

Conclusion.

None of the estimators considered here outperform least squares over all parameter values considered. The estimator R3 outperforms R1 in regions where the SNR is high, say $SNR \geq 2,500$, and in the case where the condition number is less than 900, R3 is the recommended estimator since it performs the least poorly in high SNR regions, eve . though exceeding least squares. However in situations where the condition number of the correlation matrix is high, say $\phi \geq 900$, an indication of a serious multicollinearity problem, the ridge estimators R1 and R3 outperformed least squares in every case. From the results of this simulation, either of these deterministic ridge estimators are recommended for problems where the condition number is high, and R3 is the recommended estimator for problems with low or medium condition numbers.

TABLE 1: Average sum of squared error - 4 factor model.

	SNR					
RULE	1	25	100	400	2,500	10,000
R0	625.33	25.01	6.25	1.56	.25	.06
R1	161.02	6.45	1.70	.53	.19	.05
R2	6.09	.58	.55	.60	.50	.22
R3	402.48	16.06	4.01	1.01	.18	.04
R4	4.08	.50	.55	.62	.50	.19
R5	1.91	.53	.61	.70	.61	.36
R6	31.04	1.43	.65	.47	.35	.08

TABLE 2: Average sum of squared error - 10 factor model.

	SNR					
RULE	1	25	100	400	2,500	10,000
R0	33.51	1.34	.34	.08	.01	0
R1	15.81	.68	.18	.06	.11	.05
R2	9.56	.56	.17	.10	.26	.16
R3	25.95	1.04	.26	.07	.02	.01
R4	8.11	.44	.20	.15	.27	.15
R5	3.59	.43	.34	.33	.48	.33
R6	12.76	.63	.16	.06	.20	.11

TABLE 3: Average sum of squared error - 15 factor model.

	SNR					
RULE	1	25	100	400	2,500	10,000
R0	260.61	10.42	2.61	.65	.10	.03
R1	103.08	4.12	1.05	.27	.05	.03
R2	33.81	1.45	.45	.21	.10	.13
R3	237.98	9.51	2.38	.59	.09	.02
R4	17.55	.88	.34	.25	.16	.21
R5	5.96	.61	.40	.43	.36	.43
R6	43.97	1.84	.55	.20	.08	.15

TABLE 4: Relative improvement over least squares - 4 factor model.

	SNR					
RULE	1	25	100	400	2,500	10,000
R0	100	100	100	100	100	100
R1	26	26	27	34	75	73
R2	1	2	9	38	198	345
R3	64	64	64	65	70	70
R4	1	2	9	40	201	306
R5	0	2	10	45	243	578
R6	5	6	10	30	138	135

TABLE 5: Relative improvement over least squares - 10 factor model.

	SNR					
RULE	1	25	100	400	2,500	10,000
R0	100	100	100	100	100	100
R1	47	50	54	76	791	1558
R2	29	42	51	119	1928	4862
R3	77	78	78	79	172	277
R4	24	33	59	177	2047	4495
R5	11	32	101	398	3544	9907
R6	38	47	48	69	1469	3187

TABLE 6: Relative improvement over least squares - 15 factor model.

	SNR					
RULE	1	25	100	400	2,500	10,000
R0	100	100	100	100	100	100
R1	40	40	40	42	45	98
R2	13	14	17	32	100	502
R3	91	91	91	91	91	91
R4	7	8	13	38	156	823
R5	2	6	15	66	346	1648
R6	17	18	21	31	80	568

APPENDIX A

Theorem A1 : Let I be identity matrix of order $(n \times n)$ and g an $(n \times 1)$ vector. Then

$$I - gg' \text{ is a p.s.d. matrix iff } g'g \leq 1. \tag{1}$$

Proof: Let C be an orthonormal matrix such that $C'C = I$ and

$$[C'g]' = [a, 0, \ldots, 0]. \tag{2}$$

Now $I - gg'$ is a p.s.d. matrix

$$\text{iff} \qquad C'[I - gg']C \text{ , a p.s.d. matrix,} \tag{3}$$

i.e., when

$$I - \begin{pmatrix} \alpha^2 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix} \qquad \text{a p.s.d. matrix,} \tag{4}$$

which is equivalent to

$$\alpha^2 = \mathrm{Tr}[C'gg'C] = g'g \leq 1. \tag{5}$$

Theorem A2 : A symmetric $(p \times p)$ matrix D is a p.s.d. matrix iff

$$\mathrm{Tr}CD \geq 0 \text{ for all p.s.d. } C. \tag{6}$$

Proof: Let

$$D = P\Lambda P' = \sum_{i=1}^{p} \lambda_i P_i P_i', \tag{7}$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ is the diagonal matrix of eigenvalues of D, and $P = (P_1, P_2, \ldots, P_p)$ is the corresponding matrix of orthonormal eigenvectors, where $PP' = I$. Now

$$\begin{aligned} \text{Tr}(CD) &= \text{Tr}(C \sum_{i=1}^{p} \lambda_i P_i P_i') \\ &= \text{Tr}(\sum_{i=1}^{p} \lambda_i C P_i P_i') \\ &= \text{Tr}(C \sum_{i=1}^{p} \lambda_i P_i' C P_i) \\ &= \sum_{i=1}^{p} \lambda_i P_i' C P_i, \end{aligned} \tag{8}$$

and each $P_i' C P_i \geq 0$ for all p.s.d. C. Also, if D is a p.s.d. matrix then all λ_i are non negative. Thus

$$\text{Tr}(CD) \geq 0 \text{ for any } C \text{ a p.s.d. matrix.} \tag{9}$$

Conversely, if $\text{Tr}(CD) \geq 0$ for a p.s.d. matrix C, consider

$$C = P_j P_j' \text{ for } j = 1, 2, \ldots, p. \tag{10}$$

Then we have

$$\begin{aligned} \text{Tr}(CD) &= \text{Tr}[P_j P_j' (\sum_{i=1}^{p} \lambda_i P_i P_i')] \\ &= \sum_{i=1}^{p} \lambda_i P_i P_i' P_j P_j' \\ &= \lambda_j \end{aligned} \tag{11}$$

$$\geq 0, \text{ for } j = 1, 2, \ldots, p,$$

making D a p.s.d. matrix. (12)

Theorem A3 : If A is a symmetric $(n \times n)$ matrix and Y an $(n \times 1)$ random vector with $E(Y) = \mu$ and $\text{Var}(Y) = \Sigma$ then

$$E(Y'AY) = \text{trace}(A\Sigma) + \mu' A \mu. \tag{13}$$

Theorem A4 : For any two column vectors X, and Y of real elements

i)

$$(X'Y)^2 \leq (X'X)(Y'Y), \tag{14}$$

with equality when and only when $aX + bY = 0$ for real scalars a, and b.

ii) for $A = B'B$, then

$$(X'Y)^2 \leq X'AXY'A^{-1}Y \qquad \text{if } A^{-1} \text{ exists} \tag{15}$$

with equality when $X \propto A^{-1}Y$.

Proof:

i) The quadratic form in a, b, given by

$$(aX + bY)'(aX + bY) = a^2X'X + 2abX'Y + b^2Y'Y, \tag{16}$$

is non-negative, so the

$$det\begin{pmatrix} X'X & X'Y \\ Y'X & Y'Y \end{pmatrix} \geq 0 \text{ , or } (X'Y)^2 \leq (X'X)(Y'Y), \tag{17}$$

where the equality is atained when the determinant is zero which implies that $aX + bY = 0$.

ii) Let $U = (B^{-1})'Y$, $V = BX$ and using A4.i on U, V we have

$$(X'Y)^2 \leq Y'A^{-1}YX'AX \tag{18}$$

and when $X \propto A^{-1}Y$ we have

$$(Y'A^{-1}Y)^2 = (Y'A^{-1}Y)(Y'A^{-1}Y). \tag{19}$$

Theorem A5 : Given a square non-singular $(p \times p)$ matrix M , and a p dimensional column vector Z , then

$$(M - ZZ')^{-1} = M^{-1} + \frac{M^{-1}ZZ'M^{-1}}{1 - Z'M^{-1}Z} \ . \tag{20}$$

Proof: Multiply both sides by $(M - ZZ')$.

APPENDIX B

Kadane's Small σ^2 Expansion.

For the model

$$Y = X\beta + \sigma u \ , \ \sigma u \sim N(0, \sigma^2 I), \ u \sim N(0, I) \tag{1}$$

we write $\varepsilon = \sigma u$ to take explicitly into account the magnitude of the variation in ε . When σ is small, the model is good in the sense that the variation in Y_i is small. However when σ is large, the model is not well explained by X as the variation in Y_i is large. Thus the small sigma assumption implies that the model has a good fit.

Kadane's (1971) small sigma expansion for the moments of an estimator involves

1) Assuming σ to approach zero.

2) Expanding the sampling error $(\tilde{\beta} - \beta)$ of the estimator in higher orders of σ .

3) Taking term by term expectations of the sampling error expansion.

As an illustration, consider the sampling error of the i-th component of the double h class estimator given by

$$\begin{aligned} (\alpha_{DHC} - \alpha)_i &= \delta_i a_i - \alpha_i \\ &= (1 - \frac{h_{1i} e'e}{a'Wa + h_{2i} e'e}) a_i - \alpha_i. \end{aligned} \tag{2}$$

To write the sampling error in terms of σ, α we use the model

$$\begin{aligned} Y &= X\beta + \sigma u \\ &= XGG'\beta + \sigma u \\ &= Z\alpha + \sigma u. \end{aligned} \tag{3}$$

Giving

$$\begin{aligned} a &= (Z'Z)^{-1}Z'Y \\ &= (Z'Z)^{-1}Z'[Z\alpha + \sigma u] \\ &= \alpha + \sigma\Lambda^{-1}Z'u, \end{aligned} \tag{4}$$

and

$$a_i = \alpha_i + \sigma\lambda_i^{-1}Z_i'u, \tag{5}$$

where Z_i is the i-th column of Z .

Considering

$$\begin{aligned} e &= Y - Za \\ &= Y - Z(Z'Z)^{-1}Z'Y \\ &= (I - Z(Z'Z)^{-1}Z')Y \\ &= MY, \quad \text{where } M = (I - Z(Z'Z)^{-1}Z'), \end{aligned} \tag{6}$$

we have

$$\begin{aligned} e'e &= Y'MY \\ &= (Z\alpha + \sigma u)'M(Z\alpha + \sigma u) \\ &= \sigma^2 u'Mu, \qquad \text{since } Z'M = 0. \end{aligned} \tag{7}$$

For the denominator in (a) we have

$$\begin{aligned} h_{2i}e'e + a'Wa &= (\alpha + \sigma\Lambda^{-1}Z'u)'W(\alpha + \sigma\Lambda^{-1}Z'u) + h_{2i}\sigma^2 u'Mu \\ &= \alpha'W\alpha + \sigma[2\alpha'W\Lambda^{-1}Z'u] + \sigma^2[u'Z\Lambda^{-1}W\Lambda^{-1}Z'u + h_{2i}u'Mu] \\ &= \alpha'W\alpha + \sigma[A] + \sigma^2[B] = g, \end{aligned} \tag{8}$$

where

$$A = 2\alpha' W \Lambda^{-1} Z' u \quad \text{and } B = u' Z \Lambda^{-1} W \Lambda^{-1} Z' u + h_{2i} u' M u. \tag{9}$$

Thus we may write equation (2) as

$$\begin{aligned}(\alpha_{DHC} - \alpha)_i &= \alpha_i + \sigma \lambda_i^{-1} Z_i' u - \frac{\sigma^2 h_{1i} u' M u (\alpha_i + \sigma \lambda_i^{-1} Z_i' u)}{\alpha' W \alpha + \sigma A + \sigma^2 B} - \alpha_i \\ &= \sigma \lambda_i^{-1} Z_i' u - \sigma^2 h_{1i} u' M u (\alpha_i + \sigma \lambda_i^{-1} Z_i' u) \frac{1}{\alpha' W \alpha} \left[1 + \frac{\sigma A + \sigma^2 B}{\alpha' W \alpha}\right].\end{aligned} \tag{10}$$

Using the following expansions, for small sigma

$$(1+R)^{-1} = 1 - R + R^2 - \cdots, \text{ and } (1+R)^{-2} = 1 - 2R + 3R^2 - \cdots, \tag{11}$$

we have

$$\begin{aligned}\frac{1}{g} &= \frac{1}{\alpha' W \alpha}\left[1 - \frac{\sigma A + \sigma^2 B}{\alpha' W \alpha} + \frac{\sigma^2 A^2 + 2\sigma^3 AB + \sigma^4 B^2}{(\alpha' W \alpha)^2} - \cdots\right] \\ \frac{1}{g^2} &= \frac{1}{\alpha' W \alpha}\left[1 - 2\left(\frac{\sigma A + \sigma^2 B}{\alpha' W \alpha}\right) + 3\left(\frac{\sigma^2 A^2 + 2\sigma^3 AB + \sigma^4 B^2}{(\alpha' W \alpha)^2}\right) - \cdots\right].\end{aligned} \tag{12}$$

Thus we have for equation (10) , collecting terms up to σ^2 and taking expectations

$$\begin{aligned}E(\alpha_{DHC} - \alpha)_i &= E\left[\sigma \lambda_i^{-1} Z_i' u - \frac{\sigma^2 h_{1i} u' M u \alpha_i}{\alpha' W \alpha}\right] \\ &= -\sigma^2 h_{1i}[\nu] \alpha_i \frac{1}{\alpha' W \alpha} \\ &= -\frac{h_1 \lambda_i^{-1} \nu \alpha_i}{2\theta},\end{aligned} \tag{13}$$

where

$$\theta = \frac{\alpha' W \alpha}{2\sigma^2}, \quad \text{and } \nu = \mathrm{Tr}(M) \ . \tag{14}$$

Similarly for the i-th component of the squared sampling error we have

$$
\begin{aligned}
E[(\alpha_{DHC}-\alpha)_i]^2 &= E[\sigma\lambda_i^{-1}Z_i'u - \frac{\sigma^2 h_{1i}u'Mu(\alpha_i+\sigma\lambda_i^{-1}Z_i'u)}{g}]^2 \\
&= E[\sigma^2(\lambda_i^{-1}Z_i'u)^2 - 2\sigma^3\lambda_i^{-1}Z_i'uh_{1i}u'Mu\alpha_i\frac{1}{g} \\
&\quad - 2\sigma^4(\lambda_i^{-1}Z_i'u)^2h_{1i}u'Mu\frac{1}{g} \\
&\quad + h_{1i}^2\sigma^4(u'Mu)^2(\alpha_i^2+2\alpha_i\lambda_i^{-1}Z_i'\sigma u+\sigma^2(\lambda_i^{-1}Z_i'u)^2)\,\frac{1}{g^2}].
\end{aligned}
\tag{15}
$$

Collecting terms up to σ^4 and taking expectations we have

$$
\begin{aligned}
E[(\alpha_{DHC}-\alpha)_i]^2 &= E[\sigma^2(\lambda_i^{-1}Z_i'u)^2 - 2\sigma^3\lambda_i^{-1}Z_i'uh_{1i}u'Mu\alpha_i\frac{1}{\alpha'W\alpha}[1-\frac{\sigma A}{\alpha'W\alpha}] \\
&\quad - 2\sigma^4(\lambda_i^{-1}Z_i'u)^2h_{1i}u'Mu\frac{1}{\alpha'W\alpha} \\
&\quad + h_{1i}^2\sigma^4(u'Mu)^2\alpha_i^2]\frac{1}{\alpha'W\alpha} \\
&= \sigma^2\lambda_i^{-2}E[Z_i'u]^2 \\
&\quad + 4\sigma^4\lambda_i^{-1}h_{1i}\alpha_i\alpha'W\Lambda^{-1}E(Z_i'u,Z'u)E(u'Mu)\frac{1}{(\alpha'W\alpha)^2} \\
&\quad - 2\sigma^4\lambda_i^{-2}h_{1i}\frac{1}{\alpha'W\alpha}E(Z_i'u)^2E(u'Mu) \\
&\quad + h_{1i}^2\sigma^4\alpha_i^2\frac{1}{\alpha'W\alpha}E(u'Mu)^2 \\
&= \sigma^2\lambda_i^{-1} + 4\sigma^4\lambda_i^{-1}h_{1i}\alpha_i^2\frac{W_{ii}\lambda_i^{-1}\lambda_i\nu}{(\alpha'W\alpha)^2} \\
&\quad - 2\sigma^4\lambda_i^{-2}\frac{h_{1i}}{\alpha'W\alpha}\lambda_i\nu \\
&\quad + h_{1i}^2\sigma^4\frac{\alpha_i}{\alpha'W\alpha}\nu(\nu+2) \\
&= \frac{\sigma^2}{\lambda_i} + \frac{\nu h_{1i}}{4\theta^2}[\alpha_i^2(4\frac{W_{ii}}{\lambda_i}+h_{1i}(\nu+2)\,) - 2\frac{\alpha'W\alpha}{\lambda_i}],
\end{aligned}
\tag{16}
$$

where $MZ=0$, making $u'Mu$, and Zu independent [Seber p.33], $E(u'u)=I, E(u'Mu)=\nu$, $E(u'Mu)^2=\nu(\nu+2)$, and $u'Mu\sim\chi_\nu^2$.

APPENDIX C

```
    % program runsimt
% overall job calling sequence
% for a maple4.2 interactive session.
% note ** has replaced the maple hat command
writeto('t1a');
!date;
writeto('terminal');
read sim1;
read sim2;
read sim3ff;
writeto('m110');
read simpri;
writeto('terminal');
writeto('t1b');
!date;
writeto('terminal');
writeto('t2a');
!date;
writeto('terminal');
read sim1;
read sim44;
read sim4;
writeto('m210');
read simpri;
writeto('terminal');
writeto('t2b');
!date;
writeto('terminal');
writeto('t3a');
!date;
writeto('terminal');
read sim1;
read sim55;
read sim5;
writeto('m310');
read simpri;
writeto('terminal');
writeto('t3b');
!date;
writeto('terminal');
quit;
% program gen10, gen100, gen200, gennum=1000
% calls fill of standerd normal variables
% the sample gen10 is given below

aa := array ( 1 .. 10, 1 .. 1, [
[-.1925298938],
```

```
[-.3443982579],
[-.05780795793],
[1.360936501],
[-.1562774456],
[.07838367178],
[1.048381610],
[.3076559117],
[.1210047933],
[-.4796100917]]);
% program sim1 which generates each random number using a random generator
and the central limit theorem.

yin:=proc()
local k,a;
r1:=rand(1001);
a:=0;
for k from 1 to 50 do
a:= a+ (r1()/1000);
od;
a:= ((a/50)- .5)*((12*50)**(1/2));
end;

% program sim2, which generates the regression coefficients
% sim44, and sim55 are identical except the size of the
%parameter vector changes.

with(linalg,mulcol,multiply,transpose);
ns1:=array(1..1,1..6);
b11:=array(1..4,1..6);

iden:=array(1..6,1..6,[[1,0,0,0,0,0], [0,1,0,0,0,0],lb [0,0,1,0,0,0], [0,0,0,1,0,0],
[0,0,0,0,1,0],
[0,0,0,0,0,1] ]);
r2:=rand(1001);
for j from 1 to 6 do
s1:=0;
for i from 1 to 4 do
b11[i,j]:=r2();
s1:=s1 + b11[i,j]**2;
od;
iden[j,j]:= 1 / (evalf(s1**(1/2)));
od;
t:=transpose(b11);
b1:=multiply(iden,t);
b1:=transpose(b1);

% program sim3ff, which gives the mse matrix model 1
% except for parameter differences for the vector dimensions
```

```
% identical to
% program sim4, which gives the mse matrix model 2
% program sim5, which gives the mse matrix model 3
with(linalg,mulcol,multiply,transpose);
readlib(evalm);
fmss1:=array(1..7,1..6);
genmat:=proc()
l1:=2.2357; l2:=1.5761; l3:=.1866; l4:=.0016; ss1:= 1; ss2:= 1/25; ss3:= 1/100;
ss4:= 1/400; ss5:= 1/2500; ss6:= 1/10000;
s1:=array(1..4,1..6,[
[(ss1 /l1)**(1/2),(ss2 /l1)**(1/2),(ss3/l1)**(1/2),(ss4/l1)**(1/2),(ss5/l1)**(1/2),
(ss6/l1)**(1/2)],
[(ss1 /l2)**(1/2),(ss2 /l2)**(1/2),(ss3/l2)**(1/2),(ss4/l2)**(1/2),(ss5/l2)**(1/2),
(ss6/l2)**(1/2)],
[(ss1 /l3)**(1/2),(ss2 /l3)**(1/2),(ss3/l3)**(1/2),(ss4/l3)**(1/2),(ss5/l3)**(1/2),
(ss6/l3)**(1/2)],
[(ss1 /l4)**(1/2),(ss2 /l4)**(1/2),(ss3/l4)**(1/2),(ss4/l4)**(1/2),(ss5/l4)**(1/2),
(ss6/l4)**(1/2)] ]);
s1k0:=array(1..4,1..1,[
[1/l1],
[1/l2 ],
[1/l3 ],
[1/l4] ]);
s1k5:=array(1..4,1..1,[
[l1/(l1 + l1**(1/2))],
[l2/(l2 + l2**(1/2))],
[l3/(l3 + l3**(1/2))],
[l4/(l4 + l4**(1/2))] ]);
s1k1:=array(1..4,1..1,[
[l1/(l1 + l4)],
[l2/(l2 + l4)],
[l3/(l3 + l4)],
[l4/(l4 + l4)] ]);
s1k2:=array(1..4,1..1,[
[l1/(l1 + l1**(1/2) / l1 )],
[l2/(l2 + l2**(1/2) / l1 )],
[l3/(l3 + l3**(1/2) / l1)],
[l4/(l4 + l4**(1/2) / l1 )] ]);
s1k3:=array(1..4,1..1,[
[l1/(l1 + .25 * l4 )],
[l2/(l2 + .25 * l4 )],
[l3/(l3 + .25 * l4)],
[l4/(l4 +(.25 * l4 ))] ]);lb s1k6:=array(1..4,1..1,[
[l1/(l1 + l4 )],
[l2/(l2 + l4 )],
[l3/(l3 + l4)],
[l4/(l4 +(4 * l4 ))] ]);
s1k4:=array(1..4,1..1,[
```

```
[l1/(l1 + l1**(1/2) * (l1**(1/2) + l4**(1/2) - l1**(1/2) )/ l1 )],
[l2/(l2 + l2**(1/2) * (l1**(1/2) + l4**(1/2) - l2**(1/2) )/ l1 )],
[l3/(l3 + l3**(1/2) * (l1**(1/2) + l4**(1/2) - l3**(1/2) )/ l1 )],
[l4/(l4 + l4**(1/2) * (l1**(1/2) + l4**(1/2) - l4**(1/2) )/ l1 )] ]);
end;
genmat();
#read gen10=l=10, gen100=l=100,gennum=l=1000;
readnum:=proc()
# read numbers into l by 1 vector aa
#and suppress output
read gennum ;
# read gen200 ;
end;

readnum();
#from this file are read the random numbers for aa[i,j]lb # call the proc to read
numbers into aa[l,1] # genl must correspond with l
l:=1000 ;

b:=array(1..l,1..1); c:=array(1..l,1..1);

#proceedure init begin
init:=proc(vv,num)
for v from 1 to l do
vv[v,1]:=num;
od;
end;
# call init(vector,csalar) to make a l by vector of scalarslb
     # this section assumes the matrices b1,s1,s1k1 exist.
for j from 1 to 6 do # number j are for different sigma values
mk0s1:=0;
mk1s1:=0;
mk2s1:=0;
mk3s1:=0;
mk4s1:=0;
mk5s1:=0;
mk6s1:=0;
for i from 1 to 4 do
# number i is for the number of parametersof the k-i-th ridge rule
mk0:=0;
mk1:=0;
mk2:=0;
mk3:=0;
mk4:=0;
mk5:=0;
mk6:=0;

# new section
```

```
s1m:=array(1..1,1..1,[[evalf(s1[i,j])]]);
b1m:=evalf(b1[i,j]);
init(b,b1m);
b1v:=b;
t1:=evalm( aa * s1m ) ; tt:=transpose(t1);
ts:=evalm( tt * t1); mk0:=evalf(ts[1,1]);

#
# s1k1mm:=transpose(s1k1);
s1k1mm:=array(1..1,1..6,[[s1k1[i,1], s1k2[i,1],s1k3[i,1],
s1k4[i,1], s1k5[i,1], s1k6[i,1]] ]); one:=array(1..1,1..6,[[1,1,1,1,1,1]]);
b1vv:=evalm(b1v &* one );
t1:=evalm( ( aa * s1m + b1v ) &* s1k1mm - b1vv); tt:=transpose(t1);
ts:=evalm( tt * t1);
mk1:=evalf(ts[1,1]);
mk2:=evalf(ts[2,2]);
mk3:=evalf(ts[3,3]);
mk4:=evalf(ts[4,4]);
mk5:=evalf(ts[5,5]);
mk6:=evalf(ts[6,6]);


# save the total mse of the i-th parameter with the k-th ridge rule.
mk0s1:=mk0s1 + mk0;
mk1s1:=mk1s1 + mk1;
mk2s1:=mk2s1 + mk2;
mk3s1:=mk3s1 + mk3;

mk4s1:=mk4s1 + mk4;
mk5s1:=mk5s1 + mk5;
mk6s1:=mk6s1 + mk6;
od;
# save the sum of the i-th mse's for the k ridge models and j-th sigma.
fmss1[1,j]:=mk0s1;
fmss1[2,j]:=mk1s1;
fmss1[3,j]:=mk2s1;
fmss1[4,j]:=mk3s1;
fmss1[5,j]:=mk4s1;
fmss1[6,j]:=mk5s1;
fmss1[7,j]:=mk6s1;
od;

% Program simpri prints final simulation results
% and places them in files m110,m210,m310
%program simpri

# the number of simulations is
```

```
l;
# the 1-st and 4-th eigenvalues are
l1;
l4;
# the s1 matrix is
op(s1);
op(s1k0);
op(s1k1);
op(s1k2);
op(s1k3);
op(s1k4);
op(s1k5);
op(s1k6);
mss2:=proc();
fmss2:=array(1..7,1..6);
for j from 1 to 6 do
for i from 1 to 7 do
fmss2[i,j]:= fmss1[i,j] / fmss1[1,j];
od;
od;
end;
mss2();
op(fmss1);
op(fmss2);
```

% program finpri given below takes the results and gives table form

```
m1e:=array(1..7,1..6,[
[625326.8679, 25013.07391, 6253.268679, 1563.317105, 250.1307391, 62.53268679],
[161022.3807, 6448.963130, 1702.026523, 534.9231971, 186.8019881, 45.95174361]
,
[6094.937723, 575.8532760, 554.2386390, 597.8689901, 496.0307056, 215.5002021]
,
[402481.4383, 16059.91512, 4014.137248, 1014.029380, 175.7172826, 43.82166012]
,
[4076.424790, 497.4142485, 545.6494288, 619.1639179, 501.7587064, 191.0993283]
,
[1906.138127, 526.8373482, 612.6921680, 703.4211462, 608.7053657, 361.6419283]
,
[31043.66351, 1425.080085, 646.9335743, 469.9119516, 345.1793275, 84.30588395]
]);

rm1e:=array ( 1 .. 7, 1 .. 6,[

[1.000000000, 1.000000000, 1.000000000, 1.000000000, 1.000000000, 1.000000000]
,
[.2575011389, .2578236946, .2718518956, .3421719083, .7468173995, .7348435829]
,
```

[.00974680289, .0230220914, .0886318288, .3824361597, 1.983085755, 3.446200910]
,
[.6436336882, .6420608350, .6419262395, .6486395989, .7025017526, .7007800619]
,
[.00651887036, .0198861703, .0872582735, .3960577901, 2.005985782, 3.055991004]
,
[.00304822681, .0210624791, .0979795047, .4499542313, 2.433548823, 5.783246281]
,
[.04964389842, .0569734087, .1034552660, .3005864582, 1.379995632, 1.348189056]
]);

m2e:=array(1..7,1..6,[
[33512.01768, 1340.480706, 335.1201768, 83.78004425, 13.40480706, 3.351201768]
,
[15813.44086, 675.8723988, 180.7702152, 63.27904912, 106.0779372, 52.21045123]
,
[9556.213403, 561.3732958, 172.4882063, 99.67060614, 258.4612015, 162.9256304]
,
[25948.81066, 1039.456442, 259.9021641, 66.30642820, 22.98983295, 9.296572744]
,
[8109.315440, 442.5543301, 197.1795526, 148.5092553, 274.3613076, 150.6479100]
,
[3585.418997, 426.5726930, 339.3315113, 333.6004887, 475.0890378, 331.9932614]
,
[12763.91089, 634.3872178, 161.2671853, 57.45315767, 196.8735911, 106.8037172]
]);

rm2e:=array (1 .. 7, 1 .. 6,[

[1.000000000, 1.000000000, 1.000000000, 1.000000000, 1.000000000, 1.000000000]
,
[.4718737323, .5042015120, .5394190733, .7552997815, 7.913425141, 15.57962034]
,
[.2851578050, .4187850622, .5147055243, 1.189670011, 19.28123250, 48.61707581]
,
[.7743135883, .7754355862, .7755491376, .7914346285, 1.715043928, 2.774101170]
,
[.2419823097, .3301459903, .5883846042, 1.772608938, 20.46738206, 44.95339894]
,
[.1069890518, .3182236724, 1.012566640, 3.981860975, 35.44169160, 99.06692715]
,
[.3808756313, .4732535239, .4812219510, .6857618444, 14.68679036, 31.87027359]
]);

m3e:=array(1..7,1..6,[

[260613.7532, 10424.55007, 2606.137532, 651.5343822, 104.2455007, 26.06137532]

```
,
[103082.3977, 4124.123933, 1050.162992, 274.9267871, 47.15435076, 25.59536866]
,
[33809.03589, 1447.410371, 446.4341668, 206.8424743, 104.3721331, 130.8807550]
,
[237979.2583, 9513.727400, 2377.017748, 593.8980623, 94.97559143, 23.76951519]
,
[17552.98684, 877.1581388, 342.5517770, 248.3404327, 162.4984674, 214.4466660]
,
[5962.831776, 613.8941629, 401.2586748, 430.4695093, 360.4415714, 429.5058129]
,
[43974.25225, 1842.705221, 554.1261910, 203.2579423, 83.62891725, 148.1190733]
]);


rm3e:=array(1..7,1..6,[
[1.000000000, 1.000000000, 1.000000000, 1.000000000, 1.000000000, 1.000000000]
,
[.3955370599, .3956164923, .4029576256, .4219681948, .4523394338, .9821188769]
,
[.1297285177, .1388463158, .1713010773, .3174697759, 1.001214752, 5.022020265]
,
[.9131492693, .9126271480, .9120845384, .9115375620, .9110761692, .9120591257]
,
[.06735249627, .0841435009, .1314404067, .3811624367, 1.558805573, 8.228524526]
,
[.02287995819, .0588892718, .1539668072, .6607011404, 3.457622334, 16.48055053]
,
[.1687334291, .1767659236, .2126235412, .3119680985, .8022304722, 5.683471094]
]);


msr:=proc(mm);
for j from 1 to 6 do
for i from 1 to 7 do
mm[i,j]:= round(mm[i,j] / 10);
od;
od;
end;
msr(m1e);
op(m1e);
msr(m2e);
op(m2e);
msr(m3e);
op(m3e);
fmsr:=proc(mm);
for j from 1 to 6 do
for i from 1 to 7 do
```

```
mm[i,j]:= round(mm[i,j] * 100);
od;
od;
end;
fmsr(rm1e);
op(rm1e);
fmsr(rm2e);
op(rm2e);
fmsr(rm3e);
op(rm3e);
```

REFERENCES

Alldredge, J.R., Gilb, N.S.: Ridge Regression: An Annotated Bibliography. International Statistical Review 44, 355-360 (1976).

Askin, R.G., Montgomery, D.C.: Augmented Robust Estimators. Technometrics 22, 333-341 (1980).

Baye, M.R., Parker, D.F.: Combining Ridge And Principle Component Regression : A Money Demand Illustration. Communications Statistics 13(2), 197-205 (1984).

Belsley, D.A., Kuh, E., Welsh, R.E.: Regression Diagnostics: Identifying Influential Data And Sources Of Collinearity. New York: Wiley . 1980.

Bibby, J., Toutenburg, H.: Prediction And Improved Estimation In Linear Models. New York: Wiley. 1977.

Brown, P.J.: Centering And Scaling In Ridge Regression. Technometrics 19, 35-36 (1977).

Brown, P.J., Beattie, B.R.: Improving Estimates Of Economic Parameters By Use Of Ridge Regression With Production Function Applications. American Journal Of Agricultural Economics 57, 21-32 (1975).

Chanda, A.K., Manddala, G.S.: Ridge Estimators For Distributed Lag Models. Communications Statistics 13(2), 217-225 (1984).

Chawla, J.S.: A Note On General Ridge Estimator. Communications Statistics 17(3), 739-744 (1988).

Coniffe, D., Stone, J.: A Critical View Of Ridge Regression. The Statistician 22, 181-187 (1973).

Delaney, N.J., Chatterjee, S.: Use Of The Bootstrap And Cross-Validation In Ridge Regression. Journal Of Business And Economic Statistics 4, 255-262 (1986).

Dempster, A.P., Schatzoff, M., Wermuth, N.: A Simulation Study Of Alternatives To Ordinary Least Squares. Journal Of The American Statistical Association 72, 77-91 (1977).

DiPillo, P.J.: The Application Of Bias To Discriminant Analysis. Communications Statistics A5(9), 843-854 (1976).

Draper, N.R., Herzberg, A.M.: A Ridge-Regression Sidelight. The American Statistician 41, 282-283 (1987).

Dwivedi, T.D., Srivastava, V.K., Hall, R.L.: Finite Sample Properties of Ridge Estimators. Technometrics 22, 205-212 (1980).

Farebrother, R.W.: The Restricted Least Squares Estimator And Ridge Regression. Communications Statistics 13(2), 191-196 (1984).

Gapinski, J.H.: Residents, Recreationists, Ridge Regression, And Revenues:

The Five R's Of Florida's Moter-Fuel Tax. Communications Statistics 13(2), 163-171 (1984).

Gibbons, D.G.: A Simulation Study Of Some Ridge Estimators. Journal Of The American Statistical Association 76, 131-139 (1981).

Gorman, J.W., Toman, R.J.: Selection Of Variables For Fitting Equations To Data. Technometrics 8, 27-51 (1966).

Gunst, R.F., Mason, R.L.: Regression Analysis And Its Applications. New York: Marcel Dekker. 1980.

Hald, A.: Statistical Theory With Engineering Applications. New York: Wiley. (1952).

Hemmerle, W.J.: An Explicit Solution For Generalized Ridge Regression. Technometrics 17, 309-314 (1975).

Hemmerle, J.W., Brantle, T.F.: Explicit And Constrained Generalized Ridge Estimation. Technometrics 20, 109-119 (1978).

Hinkley, D.V.: Jackknifing In Unbalanced Situations. Technometrics 19, 285-292 (1977).

Hocking, R.R.: Developments In Linear Regression Methodology: 1959-1982. Technometrics 25, 219-230 (1983).

Hoerl, A.E., Kennard, R.W.: Ridge Regression: Applications To Nonorthogonal Problems. Technometrics 12, 69-82 (1970b).

Hoerl, A.E., Kennard, R.W.: Ridge Regression: Biased Estimation For Nonorthogonal Problems. Technometrics 12, 55-67 (1970a).

Hoerl, A.E., Kennard, R.W.: Ridge Regression: Iterative Estimation Of The Biasing Parameter. Communications Statistics A5(1), 77-88 (1976).

Hoerl, A.E., Kennard, R.W., Baldwin, K.F.: Ridge Regression: Some Simulations. Communications Statistics A4(2), 105-123 (1975).

Hoerl, R.W., Schuenemeyer, J.H., Hoerl, A.E.: A Simulation Of Biased Estimation And Subset Selection Regression Techniques. Technometrics 28, 369-380 (1986).

Hoerl, R.W.: The Effect Of Ridge Regression On The Intercept. The American Statistician 40, 329-340 (1986).

Hosmane, B.: Some Conditions For Optimality Of The Almost Unbiased Estimators Of Regression Coefficients. Communications Statistics 16(6), 1725-1731 (1987).

Kadane, J.B.: Comparison Of K-Class Estimators When The Disturbances Are Small. Econometrica 39, 723-737 (1971).

Kadiyala, K.: A Class Of Almost Unbiased And Efficient Estimators Of Re-

gression Coefficients. Economics Letters 16, 293-296 (1984).

Lawless, J.F., Wang, P.: A Simulation Study Of Ridge And Other Regression Estimators. Communications Statistics A5(4), 307-323 (1976).

Lawrence, K.D., Marsh, L.C.: Robust Ridge Estimation Methods For Prediction U. S. Coal Mining Fatalities. Communications Statistics 13(2), 139-149 (1984).

Lee, Tze-San : A Simulation Study Of Deterministic Ridge Estimators. Journal Of Statistical Computation And Simulation 24, 171-183 (1986).

Lee, Tze-San : Optimum Ridge Parameter Selection. Applied Statistics 36, 112-117 (1987).

Lee, Tze-San : Selecting The Optimum K In Ridge Regression. Communications Statistics 14(7), 1589-1604 (1985).

Lindley, D.V., Smith, A.F.M.: Bayes Estimates For The Linear Model. Journal Of The Royal Statistical Society, Series B 34, 1-18 (1972).

Liu, Y., Bee, R.H.: Ridge Regression: An Application In Criminology. Communications Statistics 13(2), 263-271 (1984).

Marquardt, D.W.: Generalized Inverses, Ridge Regression, Biased Linear Estimation, And Nonlinear Estimation. Technometrics 12, 591-612 (1970).

Marquardt, D.W., Snee, R.D.: Ridge Regression In Practice. The American Statistician 29, 3-20 (1975).

Mason, R.L., Gunst, R.F.: Outlier-Induced Collinearities. Technometrics 27, 401-407 (1985).

McDonald, G.C., Galarneau, D.I.: A Monte Carlo Evaluation Of Some Ridge-Type Estimators. Journal Of The American Statistical Association 70, 407-416 (1975).

McDonald, G.C., Schwing, R.C.: Instability Of Regression Estimates Relating Air Pollution To Mortality. Technometrics 15, 463-481 (1973).

Miller, T.I., Tracy, R.L.: Further Results Concerning Ridge Regression. Communications Statistics 13(2), 251-262 (1984).

Myers, R.H.: Classical And Modern Regression With Applications. Boston: Duxbury Press. 1986.

Ohtani, K.: On Small Sample Properties Of The Almost Unbiased Generalized Ridge Estimator. Communications Statistics 15(5), 1571-1578 (1986).

Rao, C.R.: Linear Statistical Inference And Its Applications (Second edition). New York: Wiley. 1973.

Schaefer, R.L., Roi, L.D., Wolfe, R.A.: A Ridge Logistic Estimator. Communications Statistics 13(1), 99-113 (1984).

Searle, S.R.: Linear Models. New York: Wiley. 1971.

Seber, G.A.F.: Linear Regression Analysis. New York: Wiley. 1977.

Singh, B., Chaubey, Y.P.: On Some Improved Ridge Estimators. Statistische Hefte 28, 53-67 (1987).

Singh, B., Chaubey, Y.P., Dwivedi, T.D.: An Almost Unbiased Ridge Estimator. Sankhya 48, 342-346 (1986).

Smith, G., Campbell, F.: A Critique Of Some Ridge Regression Methods. Journal Of The American Statistical Association 75, 74-81 (1980).

Srivastava, V.K., Giles, D.E.A.: Exact Finite-Sample Properties Of A Pre-Test Estimator In Ridge Regression. Australian Journal Of Statistics, (1984).

Theil, H.: Principles Of Econometrics. New York: Wiley. 1971.

Theobald, C.M.: Generalizations Of Mean Square Error Applied To Ridge Regression. Journal Of The Royal Statistical Society B36, 103-106 (1974).

Ullah, A., Vinod, H.D.: Improvement Ranges For Shrinkage Estimators With Stochastic Target. Communications Statistics 13(2),207-215 (1984).

Vinod, H.D.: Application Of New Ridge Regression Methods To A Study Of Bell System Scale Economics. Journal Of The American Statistical Association 71, 835-841 (1976a).

Vinod, H.D.: Canonical Ridge And Econometrics Of Joint Production. Journal Of Econometrics 4, 147-166 (1976b).

Vinod, H.D.: Estimating The Largest Acceptable K And A Confidence Interval For Ridge Regression Parameters. Presented At The Econometric Society European Meeting. Vienna, (1977).

Vinod, H.D.: A Survey Of Ridge Regression And Related Techniques For Improvement Over Ordinary Least Squares. Review Of Economics And Statistics 60, 121-131 (1978).

Vinod, H.D.: Letter To The Editor. Technometrics 21, 138 (1979).

Vinod, H.D., Ullah, A.: Recent Advances In Regression Methods. New York: Marcel Dekker. 1981.

Vinod, H.D., Ullah, A., Kadiyala, K.: Evaluation Of The Mean Squared Error Of Certain Generalized Ridge Estimators Using Confluent Hypergeometric Functions. Sankhya 43, 360-383 (1981).

Whichern, D.W., Churchill, G.A.: A Comparison Of Ridge Estimators. Technometrics 20, 301-310 (1978).