A Survey of Ridge Regression for Improvement Over Ordinary Least Squares

R. Singh
Department of Statistics
North Eastern Hill University
Shillong – 793022 (India)

ABSTRACT

Multicollinearity may be possible cause in case of study with two or more explanatory variables. In the presence of multicollinearity, the design matrix becomes nearly singular and hence X and the corresponding $X'X$ are not of full rank. In this situation the ordinary least square (OLS) estimate cannot be obtained. Thus, adequate attention is required to give on the presence of multicollinearity in the data.

In this survey ridge regression only is discussed to solve the problem of multicollinearity. Hoerl and Kennard (HK) proposed first the technique of ridge regression that has become a popular tool for data analysis faced with a high degree of multicollinearity in their data. They have suggested adding a small positive quantity in the diagonal elements of the design matrix, $X'X$ before inverting it. In other words, they propose $\hat{\beta}_R = (X'X + \mathrm{k\ I})^{-1} X'Y$ in place of $\hat{\beta} = (X'X)^{-1} X'Y$, where $\hat{\beta}_R$ and $\hat{\beta}$ are a ridge and an OLS estimates of the parameter vector, β respectively. Though ridge estimate (RE) is biased but it has smaller mean squared error than OLS estimator. A critical appraisal is also discussed on the choice of biasing parameter in addition to its properties, its relation with other estimators and Bayesian and non-Bayesian interpretations.

Key Words: Multicollinearity, Ridge Regression, Singular Value Decomposition, Biasing Parameter, Bayesian and non-Bayesian interpretations, Criticisms.

1. 0 Introduction

When X is not of full rank, the determinant of $X'X$ is zero and one or more of its eigenvalues are zeros. In this situation ordinary least square (OLS) estimate of β and its variance, theoretically, explode. On the contrary, when all columns of X are orthogonal, then $X'X$ = I and the determinant of $X'X$ is unity. The situation of perfect multicollinearity is almost as rare as that of perfect orthogonality. The values between the two extremes are most frequent case, i.e., $0 < |X'X| < 1$ (Cooley and Lohnes, 1971). The departure of $|X'X|$ form unity is called non-orthogonality while its proximity to zero gives rise to multicollinearity. But this distinction has not been maintained in the literature. For convenience, ridge regression literature often ignores the distinction among multicollinearity, non-orthogonality and ill-conditioning.

Multicollinearity occurs when variables are highly correlated (0.90 and above but less then 1), and singularity occurs when the variables are perfectly correlated. In presence of near multicollinearity or multicollinearity, the design matrix becomes nearly singular and hence X is not of full rank. In this situation $\lambda_p \to 0$ and MSE ($\hat{\beta}$) $\to \infty$, where λ_p is the smallest eigenvalue and MSE is the mean squared error. This is the case of multicollinearity and in this situation the OLS estimate cannot be obtained. Farrar and Glauber (1967) suggest first looking at the values of r^{ii} to diagnose multicollinearity, where r^{ij} is the (i, j)th element of inverse of $(X'X)^{-1}$. Marquardt (1970) suggests a rule of thumb that VIF (i) = $r^{ii} > 5$ indicate harmful multicollinearity, where VIF is variance inflation factors. In case of ill conditioned of $X'X$ some of its eigenvlues are close to zero and their reciprocals are very large. The expected squared length of OLS estimators vector is greater than that of true parameter vector. One could refer (Brook and Moore, 1980) for a detailed discussion of this point.

Collecting more data or dropping one or more variables is the traditional solution. But collecting more data may often be expensive or not practicable in numerous situations. Dropping one or more variables from the model to alleviate the problem of multicollinearity may lead to the specification bias and hence the solution may be worse than the disease in certain situations. Our interest will be to squeeze out maximum information from whatever data at ours disposal and this has motivated the researchers to the development of some very ingenious statistical methods, for example, ridge regression, principal component regression and generalized inverse regression. Application of these statistical methods solves the problem of multicollinearity successfully.

In this survey ridge regression only is discussed to solve the problem of multicollinearity. Hoerl and Kennard (HK, 1970) have suggested adding a small positive quantity in the diagonal elements of the design matrix, $X'X$ before inverting it. In other words, they propose $\hat{\beta}_R = (X'X + \mathrm{k\ I})^{-1} X'Y$, where $\hat{\beta}_R$ is a ridge estimate of the parameter vector, β. Though RE is biased but it has smaller MSE than OLS estimator.

1. 1 Some Elementary Results

Consider the multiple linear regression model as

$$Y = X\beta + U \quad \ldots (1)$$

where Y is n x 1 vector of observations on the regressand, X is a full rank n x p matrix of observations on the p regressors, β is a p x 1 vector of coefficients associated with them and U is a n x 1 vector of unobservable disturbances such that

$$E(U) = 0 \text{ and } E(U'U) = \sigma^2 I.$$

The OLS estimate of β, denoted by $\hat{\beta}$, is given by

$$\hat{\beta} = (X'X)^{-1} X'Y \quad \ldots (2)$$

and the dispersion matrix

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \quad \ldots \quad (3)$$

Without any loss of generality in the paper it is assumed, unless otherwise stated, the design matrix $X'X$ in the form of correlation matrix by suitable standardized. The matrix being square symmetric there exists an orthonormal matrix P such that

$$P'(X'X)P = \Lambda \quad \ldots \quad (4)$$

and $$(X'X)^{-1} = P'\Lambda P \quad \ldots \quad (5)$$

where Λ is a diagonal matrix of eigenvalues of $X'X$.

An obvious choice of P is the p x p matrix with its columns the orthonormal eigenvectors of $X'X$ and arranging suitably its column the matrix Λ can be made to display the eigenvalues in the order $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$.

Since P is orthogonal, the model (1) becomes

$$Y = Z\gamma + U \quad \ldots \quad (6)$$

where $Z = XP$ and $\gamma = P'\beta$. Consequently the columns of Z are orthogonal and $Z'Z = \Lambda$. Using the Gauss-Markoff theorem and relations (2) and (3) the OLS estimate of γ, denoted by $\hat{\gamma}$, is written as

$$\hat{\gamma} = P'\hat{\beta} \quad \ldots \quad (7)$$

with dispersion matrix

$$V(\hat{\gamma}) = \sigma^2 \Lambda^{-1} \quad \ldots \quad (8)$$

The elements of $\hat{\gamma}$ vector are observed to be uncorrelated.

The results from (2) to (8) are also obtained by the use of the singular value decomposition (SVD) as discussed in Rao (1991) and Belsley and Klema (1974). In SVD it is always possible to write an n x p matrix X as

$$X = H\Lambda^{\frac{1}{2}}P' \quad \ldots \quad (9)$$

where columns of n x p matrix H are coordinates of the observations along the principal axes of X standardized in the sense that H′H = I, the matrix $\Lambda^{\frac{1}{2}}$ has the square-roots of eigenvalues of $X'X$ along its case $Z = H \Lambda^{\frac{1}{2}}$ and $\gamma = P'\beta$. The OLS estimate of β is now as

$$\hat{\beta} = P\Lambda^{\frac{1}{2}} H'Y \qquad \ldots (10)$$

and OLS estimate of γ as

$$\hat{\gamma} = P'\beta = \Lambda^{\frac{1}{2}} H'Y \qquad \ldots (11)$$

The dispersion matrix of $\hat{\gamma}$ is given by (8). The OLS estimate of β will be now as

$$\hat{\beta} = P\hat{\gamma} \qquad \ldots (12)$$

The above OLS estimators are best linear unbiased estimators (BLUE). But Gauss in 1809 suggested MSE as the most relevant criterion for choice of among estimators (Vinod, 1981). The MSE matrix for an estimator $\hat{\beta}$ of β is defined by

$$\text{MtxMSE}(\hat{\beta}) = E(\hat{\beta} - \beta)'(\hat{\beta} - \beta).$$

The closeness of $\hat{\beta}$ to β is measured in terms of squared Euclidian distance L^2 given by the trace of MtxMSE ($\hat{\beta}$) and given by

$$\text{MSE}(\hat{\beta}) = E(L^2) = E(\hat{\beta} - \beta)'(\hat{\beta} - \beta)$$

$$= \text{tr cov}(\hat{\beta}) + \text{bias}(\hat{\beta})' \text{bias}(\hat{\beta}), \qquad \text{where tr denotes the trace.}$$

1. 2 The Ridge Regression

For ill conditioned $X'X$ the OLS estimators $\hat{\beta}$ is, on the average, longer than the true parameter vector, β. Hoerl (1962) advocated adding, a positive small increment, k to the diagonal elements of $X'X$ before inverting the matrix. This increment is called the

biasing parameter or the charactersing scalar. Thus, this estimate of β is called the ridge estimate (RE) and is given by

$$\hat{\beta}_R = (X'X + \mathrm{k\,I})^{-1} X'Y \qquad \ldots (13)$$

The procedure of ridge regression (RR) actually defines a family of estimators of which OLS estimator is a member for k = 0.

HK (1970) introduce the generalized ridge regression estimator (GRE) as

$$\hat{\beta}_{GR} = [X'X + \mathrm{PDP'}]^{-1} X'Y \qquad \ldots (14)$$

where P is the matrix whose columns are orthonormal characteristic vectors of $X'X$ and D is a diagonal matrix of constants $d_i \geq 0$.

If the constants d_i are all equal and take the value $d_i = k$, the GRE reduces to the ordinary ridge estimator (ORE or RE) $\hat{\beta}_R = (X'X + \mathrm{k\,I})^{-1} X'Y$. The procedure of ORE or RE actually defines a family of estimators of which OLS estimate is a member for k = 0, i.e., with k = 0 the ORE reduces to OLS estimate. The relation between OLS estimate and ORE or RE estimators is as

$$[\mathrm{I} + \mathrm{k}\,(X'X)^{-1}]\,\hat{\beta}_R = \hat{\beta} \qquad \ldots (15)$$

Dealing with optimum solution of many variable equations from response surface methods introduced by Box and Wilson in 1951, a paper by Hoerl in 1959 traced as the genesis of RR. In a subsequent paper Hoerl (1962) applied his method of 'ridge analysis' to multiple regression analysis involving 'poorly-conditioned' data. Perhaps, the next step in the development of RR was a paper by Draper in 1963 that furnished the mathematical proofs lacking in Hoerl's original paper. However, HK (1970 a, b) provided a rigorous statistical basis of RR. The gospel of biased estimation spread in a more practical context through these papers of HK (Bibby and Toutenburg, 1977, p. 1).

HK (1970 a) obtained the RE vector $\hat{\beta}$ by minimizing the squared length of the coefficient vector subject to a fixed residual sum of squares and also, alternatively, by minimizing the residual sum of squares for a fixed squared length of the coefficient vector through the procedures of constrained minimization. Hemmerle in 1975 discussed HK (1970 a)'s GRE procedure as

$$\hat{\beta}_R = (X'X + \mathrm{K})^{-1} X'Y \qquad \ldots \ (16)$$

where K is a diagonal matrix with non-negative diagonal elements $k_1, k_2, \ldots, k_p$.

1. 3 Some Properties of Ridge Regression

The RE is biased and the amount of bias is as

$$\text{Bias } (\hat{\beta}_R) = -\mathrm{k} (X'X + \mathrm{k\ I})^{-1} \beta = (\mathrm{P}\Delta\mathrm{P}' - \mathrm{I})\ \beta \qquad \ldots \ (17)$$

where $\Delta = \text{diag } (\delta_i)$ and being square symmetric $\mathrm{W} = \mathrm{P}\Delta\mathrm{P}'$.

The RE vector with $k \neq 0$ is shorter than the OLS estimate vector and shrinks to zero as k approaches infinity. The RE is a linear transformation of the OLS estimate since

$$\hat{\beta}_R = \mathrm{Wb} = \mathrm{P}\Delta\mathrm{P}'\mathrm{b} \qquad \ldots \ (18)$$

where $\mathrm{W} = [\mathrm{I} + \mathrm{k} (X'X)^{-1}]^{-1}$.

The shrinkage factors, $\delta_i = \lambda_i (\lambda_i + \mathrm{k})^{-1}$ for i = 1, 2, …, p are the eigenvalues of W.

Vinod (1974, 1976 b) discussed 'declining deltas' for strictly positive k and declining λ_i's. The same result is obtained for GRR given in (16) for $\delta_i = \lambda_i (\lambda_i + \mathrm{k}_i)^{-1}$ and RE becomes a special case of it when all k_i are equal. The derivations are greatly simplified with the use of SVD and using canonical reduction model (6), which is equivalent to the original model (1). It can easily be seen that $\hat{\beta}_R = \mathrm{p}\,\hat{\gamma}_R$ just as in OLS estimate case given in (12). Using (18) relation for the shrinkage of $\hat{\gamma}_R$ is as

$$\hat{\gamma}_R = \Delta \hat{\gamma} \qquad \ldots (19)$$

where $\hat{\gamma}$ is the OLS estimate of γ.

HK (1970 a) showed that over 'the improvement region'

$$0 < k < \frac{\sigma^2}{\gamma_0^2} \qquad \ldots (20)$$

where γ_0 is the largest element of vector γ.

RE is superior to OLS estimate in the sense that on an average RE is closer to the parameter than OLS estimate, i.e.,

$$E[L^2(k)] < E[L^2(0)] = \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} \qquad \ldots (21)$$

Though OLS estimate is BLUE, which RE is not, but linearity and unbiasedness are irrelevant for the 'closeness'. Setting k at a value greater than zero HK (1970 a) introduced a biased RE but substantially reduced variance. Stein in 1956 discussed the benefits of requirement of unbiasedness and its mathematical rationale are questionable. Theil (1971, p. 91) wrote that it is not good to give too much importance to this indispensable criterion. Marquardt and Snee (1975) discussed through examples how a high variance unbiased estimator is inferior to a low variance biased estimator.

1. 4 Bayesian and Non-Bayesian Interpretation of Ridge Regression

1. 4. 1 Bayesian Interpretation of Ridge Regression

In model (1) if Y is multivariate normal distribution and the prior distribution of β is assumed as multivariate normal distribution with mean β_o and the dispersion matrix $\sigma_\beta^2 = \sigma^2 \, PK^{-1}P'$ then the mean of posterior distribution of β would also follow the same distribution:

$$\beta^{*} = (X'X + PKP')^{-1} (X'Y + PKP'\beta_o) \quad \ldots (22)$$

Now, when $\beta_o = 0$, (22) becomes equivalent to GRR and when $\beta_o = 0$ and K = kI, k being a scalar, (22) becomes identical to ORE or RE. Some Bayesians feel that this prior is unrealistic, and a prior mean other than the null vector should be used. Such prior knowledge about β_o is available only in rare cases.

Chipman (1964) suggested a method based on the minimum mean square error (MMSE) in face of multicollinearity. Using Bayesian approach Lindley and Smith (1972) discussed in details about RR and GRR. Becon and Hausman (1974) showed that Chipman's MMSE coincides with RE with the biasing parameter $k = \sigma^2/t^2$, i.e., the ratio of error variance to common prior variance. Lindley and Smith worked independent of Chipman's work and within entirely the Bayesian framework and they found a similar result. Both results may be considered superior to HK's. Haitovsky and Wax (1974) suggested a generalization of GRR making shrinkage towards some known β_o instead of null vector. Marquardt and Snee in 1975 state that "In this scaling it is exceedingly rare for the population value of any regression coefficient to be larger than three in a real problem". Lawless and Wang (1976) used the Bayesian framework to reinterpret the biasing parameter suggested by Hoerl, Kennard and Baldwin (1975). The Bayesian technique could provide an important research tool as implicit priors may help avoiding misapplications and over optimism with RR methods (Vinod, 1978). Reviewing the controversy of standardization Draper and Van Nostrand in 1979 conclude that standardization is generally desirable in practical cases. Smith and Campbell clarified the issues involved in 1980.

1. 4. 2 Non-Bayesian Interpretation of Ridge Regression

Marquardt (1970) discussed a class of biased linear estimators employing Generalized inverse (GI) and some theoretical properties shared by GI estimators and RE. He suggests a rule of thumb to ascertain the allowable amount of bias in the two estimators in case of ill-conditioned data. Mayer and Wilke (1973) viewed by examining different classes of biased estimators that RE was as a subclass of the class of linear transformations of the OLS estimator. Allen (1974) showed RE as a special case of data augmentation and developed a procedure to choose predictors for any specific criterion of good prediction. In their elegant paper Goldstein and Smith (1974) used normal regression theory to a suitably transformed linear model and discussed a broad class of shrinkage estimators, which achieve a smaller MSE than OLS estimator for every component of β.

Goldstein and Smith (1974) suggested a further generalization of the shrinkage factor as

$$f(\lambda_i, k) = \lambda_i^m (\lambda_i^m + k)^{-1} \qquad \ldots (23)$$

where m is an integer.

The relevant estimator in terms of original parametric space is of the form

$$\hat{\beta}(m, k) = [(X'X)^m + kI]^{-1} (X'X)^{m-1} X'Y \qquad \ldots (24)$$

This becomes RE for m = 1. HK (1975) regarded this generalized estimator as Sommers' 1975 and showed it as identical to their GRR as in (14). Lowerre (1974) suggested another family of estimators as

$$\hat{\beta}^* = (X'X + C)^+ X'Y \qquad \ldots (25)$$

where C is any symmetric matrix of proper order and $(X'X + C)^+$ is the Moore-Penrose inverse of $(X'X + C)$.

Obenchain (1975) presented another family of RE using SVD and shrinkage factors as

$$\delta_i = \lambda_i \, (\lambda_i + k \, \lambda_i^q)^{-1} \qquad \ldots \ (26) \qquad i = 1, 2, \ldots, p.$$

where k and q are parameters. When q = 0, it becomes the ORE and when q = 1, it is equivalent to Mayer and Wilke (1973) type shrunken estimators.

1. 5 Relation of Ridge Estimator with Other Estimators

Hawkins (1975) used a technique of eigenanalysis as estimator, which is identical with RE. He outlined this technique in 1973. The data matrix D = (Y **:** X) could be used to form another matrix T = D′D which could be diagonalized through an orthogonal matrix A such that ATA′ = diag (λ_i), where λ_i's are eigenvalues of T. It was shown that the regression coefficient of the linear model (1) could be estimated as a linear combination of eigenvectors in A. Now, if by suitably augmenting the data matrix with dummy observations, a matrix T (k) = T + kI is obtained then, a direct application of the result of Hawkins' 1973 would lead to estimators which are identical with RE.

Farebrother (1975) discussing MMSE of β, as given in Theil (1871, p. 125), derived a method of computing it through an estimator,

$$\hat{\beta}_F = (X'X + \frac{\sigma^2}{\beta'\beta} I) \, X'Y \qquad \ldots \ (27)$$

This estimator has close affinity with RE. However, he used the operational value of biasing parameter, k = $\frac{s^2}{\hat{\beta}'\beta}$. He showed analyzing the data used by HK (1970 b) that optimum value of k through his method is much less than HK's and also less than that of Mallows' 1973 who used a weighted MSE criterion for obtaining the optimal value.

Goldstein and Smith (1974) suggested the relevant estimator (23) and it becomes RE when m = 1. Obenchain (1975) presented another family of RE using SVD (25) and it becomes the ORE when q = 0. Stein in 1960 and James and Stein (1961) demonstrated that MSE ($\hat{\beta}$) > MSE (δ^s $\hat{\beta}$), where δ^s is a shrinkage factor based on coefficient of multiple determination. Later Baranchik in 1970 showed that a positive part δ^{s+} = max [0, δ^s] dominates δ^s. The corresponding positive part of Stein-rule estimators of β is a special case of GRE where all δ_i are equal to the uniform shrinkage fraction δ^{s+}.

RR is closely related to Bayesian estimation. Many authors like Leamer in 1978, Zellner in 1971, etc discuss the use of Bayesian method in regression. But two major drawbacks of this method are the data analyst must make an explicit statement about the form of the prior distribution and the statistical theory in this connection is not yet widely understood.

Vinod (1978) explored the relationship between the GRR due to HK (1970 a, b) and the Bhattacharya (BH) estimator of 1966 based on results by Stein and James' 1956 and 1961. Both the GRR and BH estimators are motivated by potential reductions in their MSE ($\hat{\beta}_0$) compared to the MSE for OLS denoted by MSE ($\hat{\beta}$), where $\hat{\beta}_0$ denotes any estimator of β.

Paris in 2001 introduced the maximum entropy leuven (MEL) estimator to combat the multicollinearity problem in regression analysis. MEL estimator makes the use of the information available in the data more efficiently than the OLS does, and it does not need any additional information to be supplied by the users. Wan (2002) considers the balanced loss function as a basis of measuring the performance of feasible generalized ridge and almost unbiased feasible generalized ridge estimators and obtained that both of these

estimators continue to improve over the OLS estimator in the case of ill-conditioned data, even if a relatively heavy weight is given to goodness of fit in the balanced loss function.

Mishra (2004) improves the MEL estimator to the modular MEL (MMEL) estimator and develops by Monte Carlo experiments that MMEL estimators performs significantly better than OLS and MEL estimators. Vago and Kemeny in 2006 compared the effectiveness of logistic RR and Maximum likelihood (ML) regression using clinical data of kidney-transplanted patients. They concluded that the use of RR proved to be more effective than the ML estimation for small samples but the use of RR is not recommended for the large samples.

1. 6 Choice of biasing Parameter

A value of parameter, k is needed for the ridge estimate and this is determined by the data in practice. Suggestion to choose the value of k is the smallest where components of $\hat{\beta}_R$ tend to stabilize in some subjective sense. Lin and Kmenta in 1982 discussed about prior information enabling to determine k, which is always uncertain. In most cases, however, the value of k is not known a priori but is determined on the basis of available sample observations. Under these circumstances the ORR estimator is no longer linear in observations and its properties are unknown. It incorporates no prior information but provides a convenient way for trading bias for a reduction in variance. The ridge trace has been proposed for the selection of the biasing parameter, k that is based on many mechanical rules and a graphical method. Much controversy is arising on the choice of k. In the literatures several authors have discussed for the selection of k. No firm recommendation for the choice of k seems to emerge. Some of the methods are as.

HK (1970 a, b) suggested the ridge trace (RT), which is a two-dimensional plot of $\hat{\beta}_R$ against k in the interval [0, 1]. Theil in 1963 seems to have independently suggested

a similar plot. The biasing parameter in the Bayesian interpretation is k = $\frac{\sigma^2}{\sigma_\beta^2}$, which is the ratio of error variance to the prior variance of β. During discussion HK (1970 a) proved superiority of RE over OLS estimator of a positive k whose range is given by (20). HK (1970 b) recommended standardizing the data before plotting on a RT, so as to retain numerical comparability of regression coefficients. The main theoretical justification for RR given in HK (1970) is their theorem that a strictly positive k exists for which the trace of the MSE matrix satisfies MSE ($\hat{\beta}_R$) < MSE ($\hat{\beta}$).

Minimizing the mean squared error of $\hat{\alpha}_R$ with respect to k Wermuth (1972) suggested the value of k as

$$k = \frac{\hat{\sigma}^2 \sum_{i=1}^{p} \lambda_i (\lambda_i + k)^{-3}}{\sum_{i=1}^{p} \lambda_i \hat{\alpha}_i^2 (\lambda_i + k)^{-3}} \qquad \ldots (28)$$

where $\hat{\alpha} = P' \hat{\beta}$ and $\hat{\alpha}_R = P' \hat{\beta}_R$.

Dempster (1973) developed an empirical Bayes estimator for a prior distribution of α given as α ~ N (0, ω^2 I). He then suggested replacing σ^2 by $\hat{\sigma}^2$ and using the fact of E (χ^2) = p. The suggested value of k is estimated by solving the following equation.

$$p = \frac{\sum_{i=1}^{p} \hat{\alpha}_i^2}{\hat{\sigma}^2 (\frac{1}{k} + \frac{1}{\lambda_i})} \qquad \ldots (29)$$

Sclove (1973) proposed another empirical Bayesian estimator and suggests calculating k by solving the following equation

$$\sum_{i=1}^{p} \frac{\hat{\alpha}_i^2}{\frac{1}{k} + \frac{1}{\lambda_i}} = \frac{p \hat{\sigma}^2 (n-p)}{n-p-2} \qquad \ldots (30)$$

Swindel and Chapman in 1973 suggested that a necessary and sufficient condition for MtxMSE ($\hat{\beta}$) > MtxMSE ($\hat{\beta}_R$) is

$$0 < k < \frac{2}{-\min(0, \zeta)} \qquad \ldots (31)$$

where ζ is the minimum eigenvalue of $(X'X)^{-1} - \frac{\beta\beta'}{\sigma^2}$. In case of positive and minimum eigenvalue (31) suggests that any stochastic k in the open interval (0, +∞) will reduce the MSE of OLS.

Becon and Hausman in 1974 showed that Chipman's MMSE coincides with RE with the biasing parameter k = $\frac{\sigma^2}{t^2}$, i.e., the ratio of error variance to common prior variance. While generalizing the result for weighted Mtx MSE Theobald (1974) showed the difference of the MSE for $\hat{\beta}$ and for $\hat{\beta}_R$ is a positive definite matrix if k is less than $\frac{2\sigma^2}{\hat{\beta}'\hat{\beta}}$. The recommended value of k by Farebrother (1975) was k = $\frac{\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$, which is empirically lesser optimal value of k than that of HK (1970 b) and Mallows (1973). Hoerl, Kennard and Baldwin (1975) suggested an appropriate choice for k for a better estimate of β as.

$$k_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \qquad \ldots (32)$$

where $\hat{\beta}$ is OLS estimate of β, $\hat{\sigma}^2 = \frac{1}{n-p}$ $(Y - X\hat{\beta})'(Y - X\hat{\beta})$ and p is the number of explanatory variables.

Guilkey and Murphy (1975) suggested method of directed ridge estimator (DRE), which in fact is an improvement of the HK's iterative procedure. The increments k_i are

made only to those diagonal elements of X′X whose λ_i are small. A rule of thumb given by the authors is applied to judge the minimum of λ_i. The main advantages of DRE method are relatively precise original OLS estimators and retaining unbiasedness. It is entirely possible that DRE method would have a smaller MSE than RE although Re method would be MSE optimal if true k_i were known.

Marquardt and Snee (1975) define an admissible range of k wherein MSE ($\hat{\beta}_R$) < MSE ($\hat{\beta}$). The term acceptable is used here to avoid confusion with decision theoretic admissibility concepts.

Suggested value of k by Schmidt (1976) was $\frac{\sigma^2}{\max(\gamma_i)}$, which unfortunately gives a negative value of k. The Monte Carlo results by HKB (1975) as well as by Lawless and Wang (1976) showed superiority of RE over OLS estimates in this case, it was later shown that this estimator is minimum minimax under some very restrictive conditions and hence of little practical use (Thisted in 1977).

Thisted (1976) modified the HKB estimator due to over shrinking towards zero and proposed k_T as an estimate.

$$k_T = \frac{(p-2)\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \qquad \ldots (33)$$

In a subsequent paper HK (1976) suggested the sequence of estimates of β and k, which is based on k_{HKB} and an iterative estimation procedure till the achievement of convergence.

$\hat{\beta}$: $k_0 = \frac{\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$

$$\hat{\beta}_R(k_0) \quad : \qquad k_1 = \frac{p\hat{\sigma}^2}{\hat{\beta}_R(k_0)'\,\hat{\beta}_R(k_0)}$$

$$\hat{\beta}_R(k_1) \quad : \qquad k_2 = \frac{p\hat{\sigma}^2}{\hat{\beta}_R(k_1)'\,\hat{\beta}_R(k_1)} \qquad \ldots \ (34)$$

•

•

•

Vinod (1976 b) advocated a modification of HK's ridge trace (RT) method to make the optimal choice of biasing parameter more objective and meaningful. One suggestion of Vinod was to plot the regression estimates against m instead of k.

$$m = p - \sum_{i=1}^{p} \lambda_i (\lambda_i + k_i)^{-1} \qquad \ldots \ (35)$$

This has three main advantages: (i) it can be used for GRR too, (ii) it narrows down the range of choice of k since $0 \leq k \leq \infty$ corresponds to $1 \leq m \leq p$, and (iii) it does not have the unfortunate property of HK's RT where even for completely orthogonal data the RT may appear more stable for larger value of k.

Another suggestion of Vinod is to quantify the concept of stable region of RT through an index of stability of relative magnitude (ISRM) defined as

$$\text{ISRM} = \sum_{i=1}^{p} \left(\frac{p\,\delta_i^2}{\bar{S}\,\lambda_i} - 1\right)^2 \qquad \ldots \ (36)$$

where $\bar{S} = \frac{dm}{dk} = \sum_{i=1}^{p} \lambda_i (\lambda_i + k)^{-2}$

In practice, one may choose certain k values as 0.0001, 0.001, 0.01, 0.02, . . . etc. Each of these implies p values of $\delta_i = \lambda_i (\lambda_i + k)^{-1}$. The ISRM is zero for completely orthogonal system and largest for k = 0 (i.e., in case of OLS). The value of m from this

type of RT may be chosen corresponding to smallest ISRM, which would yield a considerable narrow range of desirable value of m. In addition ISRM is not stochastic where as in HK's RT the estimates being stochastic and consequently their visual inspection leads to a stochastic determination of k. Vinod (1978) corrected the suggested value of k by Schmidt that is inefficient compared to Theobald's estimate.

Ullah et al. (1981) suggested a family of double h-class ridge estimators, of which many earlier methods are special cases. They showed on the basis of MSE criterion the GRR would dominate over OLS estimate for

$$k_i = h_1 \sigma^2 (r_i^2 - h_2 \sigma^2 \lambda_i^{-1})^{-1} \qquad \ldots (37)$$

They further obtained single value k as.

$$k = p\, h_1 \hat{\sigma}^2 (C'C - h_2 \hat{\sigma}^2 \sum \lambda_i^{-1})^{-1}$$ where C is the OLSE of vector Y.

This value of k coincides with those suggested by Hoerl, Kennard & Baldwin (1975) and Farebrother (1975) for ($h_1 = 1, h_2 = 0$) and ($h_1 = \frac{1}{p}$ and $h_2 = 0$) respectively. The performance of a RR estimator is based on a given value of k and it depends on (i) the number and the values of the regression coefficients, (ii) the degree of multicollinearity, and (iii) the value of the variance of the disturbances, σ^2 (Lin and Kmenta, 1982).

Mardikyan and Cetin (2008) developed a goal-programming model by which efficient value of the biasing parameter k is determined based on the minimization of the VIF and the maximization of the determination coefficient R^2 of the model to yield an efficient biasing parameter for RR. They further clarify that it gives the efficient value of k much closer to the values discussed in the literature, by just one attempt.

1. 7 Some Criticisms on Ridge Regression

In RR controversy concentrates on the choice of biasing parameter, k. Method of RR could not produce a single solution to the estimation problem, if a ridge trace is

considered. A number of criticisms of RR method are available in the literature. Some of them are directed towards HK's approach while others are of a general nature.

(i) The restriction imposed on k to lie in [0, 1] is arbitrarily and tantamount to an assumption, in a Bayesian framework, that error variance is less than the variance of the priors (Smith and Campbell, 1980).

(ii) The guidelines to choose the value of k are very vague. There is no sound logic for the stability of correct choice of k.. Vinod (1976) discussed for quantification of the concept of stability.

(iii) The RR method ignores the fundamental fact that a linear transformation of a model does not change the estimates of the model (Smith and Campbell, 1980).

(iv) Proof given by HK that MSE of $\hat{\beta}_R$ is less than that of $\hat{\beta}$ is valid only when k is known but this is hardly the case in practical situations. However, Vinod (1976 c) attempted to show that MSE of HK's GRR is less than that of OLS for positive and stochastic k_i.

(v) In case of ill-conditioned $X'X$ and large MSE some components of vector $\hat{\beta}$ will be too large and others too small. But unless we know which components are too large and which too small, there seems little justification for reducing their absolute values and no reason to believe why a particular component of $\hat{\beta}_R$ should be closer to the true value than corresponding component of $\hat{\beta}$. Nelder in 1972 suggested in this context that arguments based on average MSE of all coefficient estimates are not very convincing. However the DRE method of Guilkey and Murphy (1975) may, perhaps, be less exposed to such criticisms.

(vi) According to Coniffe and Stone (1973) the ill conditioned $X'X$ actually indicates the inadequacy of data or the misspecification of the model as remedial

measures. But the questions arise that in many areas of applied research collection of further data may be impracticable and, secondly, it is not clear how one would decide whether the inadequacies rest with model or with the data. He further pointed out that MSE is not only the criterion for determining a particular estimator but the estimator would also have a tractable distribution for hypothesis testing and the construction of confidence intervals.

(vii) The MSE of RE should not only be compared to that of OLS estimator, but should be compared to those of other biased estimation procedures (McDonald and Galarneau, 1975).

(viii) According to Schmidt (1976) RR could not be useful to econometricians due to an undeveloped theory of hypothesis testing with RR. Many econometricians such as Goldberger, Houthakker and Taylor, and others are interested in numerical values of regression coefficients.

(ix) McDonald and Schwing (1973) used the ridge regression procedure for analyzing mortality rates by various socio-economic (weather and pollution) variables. Vinod (1974) modified the canonical correlation analysis in the light of RR and used for estimating a joint production function.

(x) When the value of k is not given a priori and has to be determined from sample observations, the resulting ORR estimators are no longer linear and can compete with OLS on equal terms of the same prior information (Lin and Kmenta, 1982).

1. 8 Concluding Remarks

Multicollinearity is frequent for both a theoretical problem and problem with a particular sample of data. In case of its presence in the data the design matrix becomes close to singular and hence X as well as $X'X$ is not of full rank. OLS estimate cannot be obtained in this case. Dropping some of the highly correlated variables is one simple

solution to this problem and this strategy usually works well. However, there are situations when the variables are too important to be excluded from the analysis. Dropping one or more variables from the model to alleviate the problem of multicollinearity may lead to the specification bias and hence the solution may be worse than the disease in certain situations. To collect more data is another solution but this may often be expensive or not practicable in numerous situations. One may be interested to squeeze out maximum information from whatever data one has at one's disposal.

At first detection of multicollinearity is required in the data. RR is an alternative estimation method when there is an extremely high degree of multicollinearity present in the data set (Darlington in 1978). RR is more advanced solution of multicollinearity but generally greatly reduces the MSE giving more reliable estimates of β.

The ridge estimator (RE) is a small positive increment made to the diagonal element of the design matrix before inverting it. However, RE is biased, it has smaller mean square error than OLSE and it is compared with other biased estimators.

Ridge regression was originally developed to prevail over the singularity. Anders (2001) suggests that RR is an application of Tikhonov regularization (TR), a method that has been explored in the approximation theory literature for about as long as RR has been used in Statistics. A large number of choices for $k \in (0, 1)$ are possible. Each choice gives a new ridge estimator and this is the reason to denote it with subscript k.

In applied research RR technique must be used with caution. Though various methods for choosing k have been suggested but no firm recommendation for optimal k seems to emerge. Obtaining the value of k for a specific problem remains something an art. Further, underdeveloped theory of hypothesis testing with RR also limits its utility (Schmidt, 1976), though Vinod (1977) suggested development of non-spherical

confidence intervals centred at RE using Stein's (1974) unbiased estimator of MSE of biased estimators. But these still remain problems of further research.

REFERENCES

Allen, D. M. (1974), "*The Relationship between Variable selection and Data Augmentation and a Method for Prediction*", Technom., 16, 125-127.

Anders, B. (2001), "*Ridge regression and Inverse Problems*", Stock. Univ., Sweden

Belsley, D. A., and Klema, V. C. (1974), "*Detecting and Assessing the Problems Caused by Multicollinearity: A Use of the Singular Value Decomposition*", Working Paper No. 66 (Cambridge, Mass: Nation. Bure. of Econo. Research, Dec.).

Bibby, J. and Toutenburg, H. (1977), "*Prediction and Improved Estimation in Linear Models*", Wiley, New York.

Cooley, W. M. and Lohnes, P. R. (1971), "*Multivariate Data Analysis*", Wiley, New York.

Dempster, A. P. (1973), "*Alternatives to Least Squares in Multiple Regression*", in Kabe and Gupta (eds.), Multiv. Statis. Infer., North Holland, New York, 25-40.

Farebrother, R. W. (1975), "*The Minimum Mean Square Error Linear Estimator and Ridge Regression*", Technom., 17 (Feb.), 127-128.

Goldstein, M. and Smith, A. F. M. (1974), "*Ridge Type Estimators for Regression Analysis*", JRSS, B, 36, 284-291.

Greene, W. H. (2002), "*Econometrics Analysis*", Pears. Edu. Asia, Delhi.

Guilkey, D. K. and Murphy, J. L. (1975), "*Directed Ridge Regression Techniques in Cases of Multicollinearity*", JRSS, 70, 352 (Dec.), 769-775.

Haitovsky, Y. and Wax, Y. (1974), "*Generalized Ridge Regression, Least Squares with Stochastic Prior Information and Bayesian Estimators*", presented at NBER-NSF seminar on Bayesian Inference in Econometrics, Ann Arbor, Michigan, cited in Vinod (1978).

Hawkins, D. M. (1975), "*Relation between Ridge Regression and Eigenanalysis of the Augmented Correlation Matrix*", Technom., 17, 477- 480.

Hoerl, A. E. (1962), "*Application of Ridge Analysis to Regression Problems*", Chem. Engg. Proce, 60, 54-59.

Hoerl, A. E. and Kennard, R. W. (1970 a), "*Ridge Regression: Biased Estimation of Nonorthogonal* Problems", Technom., 12 (Feb.), 55-67. (Republished in Technom., 42 (2000), 80-86.

---- (1970 b), "*Ridge Regression: Application to Nonorthogonal Problems*", Technom., 12 (Feb.), 69-82.

Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975), "*Ridge Regression: Some Simulation*", Commn. in Statis., 4, 105-123.

Lin, K. and Kmenta, J. (1982), "*Ridge Regression Under Alternative Loss Criteria*", Rev. of Econ. & Statis., 64, 488 - 494.

James, W. and Stein, C. (1961), "*Estimation with Quadratic Loss*", Proc. Fourth Berkeley Sympos. in Mathem. and Prob. I, 361-379, cited in Malinvaud, E. (1970).

Lawless, J. F. and Wang, P. (1976), "*A Simulation Study of Ridge and Other Regression Estimators*", Commn. in Statis., A, 5 (4), 307-323.

Leamer, E. E. and Chamberlain, G. (1972), "*A Bayesian Interpretation of Pretesting*", JRSS, B, 38, 85-94.

Lindley, D. V. and Smith, A. F. M. (1972), "*Bayes Estimates for the Linear Model*", JRSS, B, 34, 1- 41.

Lowerre, J. M. (1974), "*On the Mean Square Error of Parameter Estimates for Some Biased Estimators*", Technom., 16 (3), 461- 464.

Malinvaud, E. (1970), "*Statistical Methods of Econometrics*", North Holland, Amsterdam.

Mallows, C. L. (1973), "*Some Comments on Cp*", Technom., 15 (Nov.), 661-675.

Mardikyan, S. and Cetin, E. (2008), "*Efficient Choice of Biasing Constant for Ridge Regression*", Int. J. Contemp. Math. Sciences, 3 (11), 527-536.

Marquardt, D. W. (1970), "*Generalized Inverses, Ridge regression: Biased Linear Estimation and Non-linear Estimation*", Technom., 12, 591-612.

Marquardt, D. W. and Snee, R. D. (1975), "*Ridge Regression in Practice*", The Amer. Statist., 29, 3-20.

Mayer, L. S. and Wilke, T. A. (1973), "*On Biased Estimation in Linear Models*",

Technom., 15 (3), 497-508.

McDonald, G. C. and Galarneau, D. I. (1975), "*A Monte Carlo Evaluation of Some Ridge-Type Estimators*", JASA, 70 (June), 407-416.

McDonald, G. C. and Schwing, R. C. (1973), "*Instabilities of Regression Estimates Relating Air Pollution to Mortality*", Technom., 15, 463-481.

Mishra, S. K. (2004), "*Multicollinearity and Maximum Entropy Leuven Estimator*", Economics Bulletin, 3 (25), 1-11.

Obenchain, R. L. (1975), "*Ridge Analysis Following a Preliminary Test of a Shrunken Hypothesis*", Technom., 17 (Nov.), 431-441.

Rao, C. R. (1991), "*Linear Statistical Inference and Its Applications*", Sixth reprint, Wiley Eastern Ltd., New Delhi.

Schmidt, P. (1976), "*Econometrics*", Marcel Dekker, New York.

Smith, G. and Campbell, F. (1980), "*A Critique of Some Ridge Regression Methods*", JASA, 75, 74-103.

Theobald, C. M. (1974), "*Generalizations of Mean Square Error Applied to Ridge Regression*", JRSS, B, 36, 103-106.

Ullah, A, Vinod, H. D. and Kediyala, K. (1981), "*A Family of Improved Shrinkage Factors for the Ordinary Ridge Estimator*", European Econometric Society Meetings, Athens, 1979, in E. G. Charatsis (ed.), Selected Papers on Contemporary Econometric Problems, North Holland, Amsterdam.

Vinod, H. D. (1978), "*A Survey of Ridge Regression and Related Techniques for Improvements over Ordinary Least squares*", Rev. of Econom. and Statis., 60, 121-131.

Vinod, H. D. and Ullah, A. (1981), "*Recent Advances in Regression Methods*", Marcel Dekker, New York.

Wan, A. T. K. (2002), "*On Generalized Regression Estimators under Collinearity and Balanced Loss*", Appl. Math. and Comput., 129, 455-467.

Wermuth, N. (1972), "*An Empirical Comparison of Regression Methods"*, Unpub. Doct. Disser., Harv. Univ., cited in Lin and Kamena (1982).

** **