CUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

MSc THESIS

Selma TOKER

RIDGE REGRESSION ESTIMATOR AND RELATED ESTIMATORS

DEPARTMENT OF STATISTICS

ADANA, 2011

CUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

RIDGE REGRESSION ESTIMATOR AND RELATED ESTIMATORS

Selma TOKER

MSc THESIS

DEPARTMENT OF STATISTICS

We certify that the thesis titled above was reviewed and approved for the award of degree of the Master of Science by the board of jury on 11/08/2011.

.....	.....	.....
Prof. Dr. Selahattin KAÇIRANLAR	Prof. Dr. Sadullah SAKALLIOĞLU	Asst. Prof. Dr. Gülsen KIRAL
SUPERVISOR	MEMBER	MEMBER

This MSc Thesis is written at the Department of Institute of Natural And Applied Sciences of Cukurova University.
Registration Number:

Prof.Dr. İlhami YEĞİNGİL
Director
Institute of Natural and Applied Sciences

This thesis is supported by

- **Cukurova University Scientific Research Projects Unit Project Number: FEF2009YL55**
- **The Scientific and Technological Research Council of Turkey (TUBITAK)-Departments to Support Scientists (BIDEB).**

ABSTRACT

MSc THESIS

RIDGE REGRESSION ESTIMATOR AND RELATED ESTIMATORS

Selma TOKER

CUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF STATISTICS

Supervisor : Prof. Dr. Selahattin KAÇIRANLAR
Year: 2011, Pages: 117
Jury : Prof. Dr. Selahattin KAÇIRANLAR
: Prof. Dr. Sadullah SAKALLIOĞLU
: Asst. Prof. Dr. Gülsen KIRAL

To cope with the multicollinearity problem several biased estimators are proposed as an alternative to the ordinary least squares (OLS) estimator. One of the biased estimators is ridge estimator (RE) of Hoerl and Kennard (1970). In this study, RE and many kinds of estimator related with RE are examined. RE in different regression models are also examined. Besides this, two parameter ridge estimator proposed by Lipovetsky and Conklin (2005) is compared with OLS estimator, RE, Stein type estimator under the mean square error matrix criterion. A real data set is analyzed to evaluate the performance of mentioned estimators.

Keywords: Multicollinearity, Ordinary Least Squares Estimator, Ridge Estimator, Two Parameter Ridge Estimator.

ÖZ

YÜKSEK LİSANS TEZİ

RİDGE REGRESYON TAHMİN EDİCİ VE İLİŞKİLİ TAHMİN EDİCİLER

Selma TOKER

CUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

Danışman : Prof.Dr. Selahattin KAÇIRANLAR
Yıl: 2011, Sayfa: 117
Jüri : Prof. Dr. Selahattin KAÇIRANLAR
: Prof. Dr. Sadullah SAKALLIOĞLU
: Yrd. Doç. Dr. Gülsen KIRAL

Çoklu iç ilişki problemini gidermek için en küçük kareler (EKK) tahmin ediciye alternatif olarak birçok yanlı tahmin edici önerilmiştir. Bunlardan birtanesi Hoerl and Kennard (1970)'ın ridge tahmin edicisidir. Bu çalışmada ridge tahmin edici ve ridge tahmin edici ile ilişkili birçok tahmin edici incelenmiştir. Ayrıca farklı regresyon modellerindeki ridge tahmin ediciler de incelenmiştir. Bunun yanında Lipovetsky ve Conklin (2005) tarafından tanımlanan iki parametreli ridge tahmin edici; EKK tahmin edici, ridge tahmin edici ve Stein tipi tahmin edici ile hata kareler ortalaması matrisi kriteri altında karşılaştırılmıştır. Belirtilen tahmin edicilerin performanslarını değerlendirnmek için gerçek bir veri seti analiz edilmiştir.

Anahtar Kelimeler: Çoklu İç İlişki, En Küçük Kareler Tahmin Edici, Ridge Tahmin Edici, İki Parametreli Ridge Tahmin Edici.

ACKNOWLEDGEMENTS

First, I will always be grateful to my supervisor Prof. Dr. Selahattin KAÇIRANLAR, for his never ending patience and encouragement. Also, I would like to extend my thanks for his invaluable guidance and support throughout this study. His idealism, optimism and vigorous search for perfection have illuminated the long and arduous path to the successful completion of this work. He will always be a role model for me.

Second, I would like to express my special thanks to the members of the dissertation committee. I would like to give most sincere gratitude to my professors and colleagues in my department whom I benefited much from their experiences, worldviews, attitudes and intellects. Besides this, I was lucky to have special friends and I'm deeply indebted to them for their close friendship.

Last but not least, I would like to extend my deepest thanks to my family for their unrequited love, support and patience, they give all the way. They gived me the necessary encouragement to keep going with my study from beginning till the end.

CONTENTS **PAGE**

LIST OF TABLES **PAGES**

LIST OF FIGURES **PAGES**

LIST OF ABBREVIATONS

AURE	: Almost unbiased ridge estimator
BLUE	: Best linear unbiased estimator
GCV	: Generalized cross validation
GRE	: Generalized ridge estimator
iff	: If and only if
LAV	: Least absolute value
MMSE	: Mean square error matrix
MRE	: Modified ridge estimator
MSE	: Mean square error
nnd	: Nonnegative definite
OLS	: Ordinary least squares
PCR	: Principal component regression
pd	: Positive definite
PLS	: Partial least squares
PMSE	: Predicted mean square error
psd	: Positive semi definite
RE	: Ridge estimator
RMRE	: Restricted modified ridge estimator
RLS	: Restricted least squares
RRE	: Restricted ridge estimator
SURE	: Seemingly unrelated regression equations
URE	: Unbiased ridge estimator
WMSE	: Weighted mean square error

1. INTRODUCTION

1.1. Basic Definitions

Definition 1.1. A multiple linear regression model is

$$y = X\beta + \varepsilon \tag{1.1.}$$

where y is a $n \times 1$ vector of observable random variables; X is a $n \times p$ matrix of non-stochastic regressors; β is a $p \times 1$ vector of unknown regression coefficients; ε is a $n \times 1$ vector of disturbances. The following conventional assumptions are often used: the rank of X is p (full rank); the vector of errors is normally distributed with $\varepsilon \sim N(0, \sigma^2 I)$.

Definition 1.2. $\Lambda = diag(\lambda_1, \dots, \lambda_p)$, is a diagonal matrix whose diagonal elements are eigenvalues of $X'X$. T is a $p \times p$ matrix whose elements are eigenvectors of $X'X$ and $T'X'XT = \Lambda,\ TT' = T'T = I_p$

$$Y = Z\alpha + \varepsilon\,,\ \ \varepsilon \sim N(0, \sigma^2 I) \tag{1.2.}$$

is called canonical model where $Z = XT$, $\alpha = T'\beta$ and $Z'Z = T'X'XT = \Lambda$.

Definition 1.3. Let A be a symmetric matrix and $Q = x'Ax$ is a quadratic form. If $Q > 0$ for all $x \neq 0$, A is positive definite (pd). If $Q = 0$ for at least one $x \neq 0$, A is positive semi definite (psd). A is nonnegative definite (nnd) if and only if A is pd or psd.

Definition 1.4. y is a random vector. Let C and d be respectively a matrix and a vector of the relevant size, the vector

$$\hat{\beta} = Cy + d \tag{1.3.}$$

may be called a linear estimator of β. $\hat{\beta}$ is called homogeneous estimator if $d = 0$; otherwise $\hat{\beta}$ is called a heterogeneous estimator.

Definition 1.5. (MSE) Mean square error of an estimator $\hat{\beta}$ is

$$MSE(\hat{\beta}) = E(\hat{\beta}-\beta)'(\hat{\beta}-\beta) \qquad (1.4.)$$
$$= Tr[Var(\hat{\beta})] + [Bias(\hat{\beta})]'[Bias(\hat{\beta})].$$

Definition 1.6. (WMSE) Weighted mean square error of an estimator $\hat{\beta}$ is

$$WMSE(\hat{\beta}) = E(\hat{\beta}-\beta)'W(\hat{\beta}-\beta) \qquad (1.5.)$$

where W is psd weight matrix.

Definition 1.7. (PMSE) Predictive mean square error of an estimator $\hat{\beta}$ is

$$PMSE(\hat{\beta}) = E(y-X\hat{\beta})'(y-X\hat{\beta})$$
$$= E(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta). \qquad (1.6.)$$

Definition 1.8. (MMSE) Mean square error matrix of an estimator $\hat{\beta}$ is

$$MMSE(\hat{\beta}) = E(\hat{\beta}-\beta)(\hat{\beta}-\beta)' \qquad (1.7.)$$
$$= Var(\hat{\beta}) + [Bias(\hat{\beta})][Bias(\hat{\beta})]'.$$

Definition 1.9. (MMSE Criterion) Let $\hat{\beta}_1$ and $\hat{\beta}_2$ be two estimators of β. Then $\hat{\beta}_2$ is called MMSE superior to $\hat{\beta}_1$ if the difference of their MMSE matrices is ndd, that is, if

$$\Delta(\hat{\beta}_1, \hat{\beta}_2) = MMSE(\hat{\beta}_1) - MMSE(\hat{\beta}_2) \geq 0.$$

Definition 1.10. The quadratic loss function for $\hat{\beta}$ is

$$L(\hat{\beta}, A) = (\hat{\beta}-\beta)'A(\hat{\beta}-\beta) \qquad (1.8.)$$

where A is a symmetric, $(p \times p)$ matrix and pd.

The loss depends on the sample, thus the average or expected loss over all possible samples, called risk, is considered.

Definition 1.11. The quadratic risk of an estimator $\hat{\beta}$ of β is defined as

$$R(\hat{\beta}, A) = E[L(\hat{\beta}, A)]$$
$$= E(\hat{\beta} - \beta)' A(\hat{\beta} - \beta). \quad (1.9.)$$

To find an estimator $\hat{\beta}$ that minimizes the quadratic risk function over a class of appropriate functions, a criterion to compare estimators is defined.

Definition 1. 12. (R(A) superiority) An estimator $\hat{\beta}_2$ of β is called R(A) superiority over another estimator $\hat{\beta}_1$ of β if

$$R(\hat{\beta}_1, A) - R(\hat{\beta}_2, A) \geq 0.$$

Definition 1.13. The quadratic risk function is just a scalar-valued version of the MMSE

$$R(\hat{\beta}, A) = Tr\{A[MMSE(\hat{\beta})]\}. \quad (1.10.)$$

One important connection between $R(A)$ superiority and MMSE superiority has been given by Trenkler (1974).

Theorem 1.1: (Trenkler 1974) Consider two estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ of β. The following two statements are then equivalent

$$\Delta(\hat{\beta}_1, \hat{\beta}_2) \geq 0$$
$$R(\hat{\beta}_1, A) - R(\hat{\beta}_2, A) = Tr\{A[\Delta(\hat{\beta}_1, \hat{\beta}_2)]\} \geq 0 \text{ for all } A \geq 0$$

where $\Delta(\hat{\beta}_1 - \hat{\beta}_2) = MMSE(\hat{\beta}_1) - MMSE(\hat{\beta}_2)$.

Definition 1.14. An estimator $\hat{\beta}_1$ is said to dominate an estimator $\hat{\beta}_2$ under the quadratic loss $L(\hat{\beta}, A)$ if

$$R(\hat{\beta}_2, A) - R(\hat{\beta}_1, A) \geq 0 \text{ for all } \beta$$
$$R(\hat{\beta}_2, A) - R(\hat{\beta}_1, A) > 0 \text{ for some } \beta.$$

The concept of dominating defines a relation between two given estimators. To find out which estimator is best in some sense leads to the concept of admissibility.

Definition 1.15. An estimator is said to be admissible with respect to the risk $R(\hat{\beta}, A)$ if no estimator exists which dominates it.

1.2. Historical Survey

Hoerl (1959) explains ridge analysis in terms of finding the maximum values of a quadratic function on n dimensional sphere. Hoerl (1962) explains what ridge estimator (RE) is and how it can be used to solve problems involving ill conditioned data. In the paper of Draper and Smith (1969) ridge trace is included as a procedure for choosing a best subset of predictor variables from a larger set. Hoerl and Kennard (1970a) is the seminal paper on ridge regression most often cited in the literature. An estimation procedure based on adding small, positive quantities to the diagonal of $X'X$ is proposed. The ridge trace, a method for showing in two dimensions the effects of nonorthogonality is introduced. They also show the mean square error (MSE) optimality over ordinary least squares (OLS) estimators. Hoerl and Kennard (1970b) is the exposition of the use of ridge regression methods. Two examples from the literature are used as a base. Recommendations are made for obtaining a better regression equation than that given by OLS estimators. Marquardt (1970) gives some properties of the OLS for the nonfull rank model where the generalized inverse is employed. He also shows theoretical properties shared by generalized inverse estimators, RE and the corresponding nonlinear estimation procedures. Banerjee and

Carr (1971) comment on the Hoerl and Kennard (1970a). An alternative characterization of the form of biased estimator has been presented in this note. Bibby (1972) proposes estimators for the parameters of a general linear model which minimizes the expected value of quadratic loss function. The admissibility of ridge regression estimators is proved. In Hocking (1972) ridge regression is one of the criteria for selecting a subset of predictor variables and it is noted that ridge trace is used to identify variables which do not hold their predicting power. Lindley and Smith (1972) develop an estimator very similar to the simple RE by Bayesian methods. Coniffe and Stone (1973) examine the ridge regression procedure and contains some critical comments on the usefulness of the method. Dempster (1973) explains how various ridge and contraction estimators are special cases of Bayesian estimators. In the paper of Mayer and Wilkie (1973), the ridge estimators are viewed as a subclass of the class of linear transform of the OLS estimator. An alternative class of estimator, contraction estimator, is considered which satisfy the optimality condition proposed by Hoerl and Kennard (1970a). Furthermore, ridge estimators and contraction estimator are derived as minimum norm estimators in the class of linear transform of the OLS estimator. It is shown in Bacon and Hausman (1974) that ridge regression and Chipman's method of estimation have something to offer the econometrician when the data are highly collinear. A class of shrinkage estimators is defined by Goldstein and Smith (1974). Comparision is made with the James Stein estimator and with the generalized inverse estimator proposed by Marquardt (1970). A bayesian formulation of ridge estimators is presented. Conditions for the MSE of ridge type estimators to be smaller than that of the OLS estimators are given in Lowerre (1974). Both the full and the non-full rank case are considered. Results on ridge regression in weighing designs are presented in Sihota and Banerjee (1974). Swindel (1974) gives an elementary geometrical portrayal of the instabilities of OLS regression coefficients including the possibilities of incorrect signs. In the paper of Theobald (1974), using a generalization of MSE a sufficient condition for the RE to have a smaller MSE than the OLS estimator is given. The article of Churchill (1975) reviews briefly the ridge estimation procedure and discuss some of the theoretical reasons as to why ridge regression analysis might be expected to provide improved

estimates of the β's. Ellingsen (1975) applies ridge regression to sequential estimation to improve estimation when the X′X matrix is ill conditioned. The similarity between the minimum mean square error estimator and ridge regression is studied by Farebrother (1975). Guilkey and Murphey (1975) suggest a directed RE that alters only diagonal elements corresponding to relatively small eigenvalues. The method has the advantage of giving estimates that may be less biased than those of other ridge regression methods that alter all diagonal elements. Hoerl, Kennard and Baldwin (1975) give an algorithm for selecting the biasing parameter, k, in ridge regression. By means of simulation it is shown that the algorithm has a number of desirable properties. Hsiang (1975) stresses a Bayesian view of ridge regression. Smith and Goldstein (1975) do not accept the conclusions propounded by Coniffe and Stone (1973) that ridge estimators are unlikely to be of practical use to the researcher. A limited bibliography of papers on the theory and application of ridge regression was presented by Alldredge and Gilb (1976). A general Bayesian formulation of regression problem is considered and the resulting estimators are related to the LS and ridge estimators of the regression coefficients. Gunst and Mason (1976) compare the principal components estimator (PCR), OLS estimator and RE by using generalized mean square error criterion. Subset selection problem and variable analysis in linear regression is examined in Hocking (1976). Hoerl and Kennard (1976) give an iterative method for selecting the biasing parameter. In the simulation study of Lawless and Wang (1976), two ordinary ridge estimators (Hoerl, Kennard and Baldwin (1975) and the other introduced here) have better performance than OLS and other estimators considered. Admissible linear estimators are presented in Rao (1976) and it is indicated that admissible linear estimators are Bayes linear estimators (BLE) or limits of BLE's. Biased estimators like ridge and shrunken estimators are marked as a special cases of BLE's. It is shown that ridge regression is a special case of the class of mixed estimators in the paper of Smith (1976). Swindel (1976) proposes a modified ridge estimator (MRE) depend upon a prior information. Vinod (1976) makes a simulation study which supports Stein's estimators and his fixed point solution over OLS and Farebrother's estimator. Comparisions to the simulation of Hoerl, Kennard and Baldwin (1975) are made. In

the simulation study of Dempster, Schatzoff and Wermuth (1977), estimated regression coefficients and error in these estimates are computed for 160 artificial data sets drawn from 160 normal linear models with respect to factorial designs. Ridge regression established upon stochastic prior information is examined by Fomby and Johnson (1977). Gunst and Mason (1977) use a MSE or criterion to compare five estimators of the coefficients in a linear regression model which are OLS, principal components, ridge regression, latent root and shrunken estimator. In the paper of Kasarda and Shin (1977), a technique of selecting the optimal k value for minimizing the mean square error of estimation is given. Classical F-tests and confidence regions for non-stochastically shrunken ridge regression estimators are discussed in Obenchain (1977). Equality is shown of the g-inverse and Moore-Penrose inverse representation of the BLUE in general linear model in the paper of Pukelsheim (1977). Snee (1977) use ridge regression as one of the estimation methods to designate the validity of regression models. Alam and Hawkes (1978) show that a class of ridge estimators is dominant over OLS estimator if the matrix $X'X$ is properly conditioned. Baldwin and Hoerl (1978) prove that minimum MSE is bounded. Browne and Rock (1978) provide Newton algorithms for ordinary ridge regression to estimate single additive constant while they are giving the additive constants in closed form for the generalized ridge regression. Farebrother (1978a) describes a class of shrinkage estimators members having a MMSE which is less than that of the OLS estimator. Lamotte (1978) improve Bayes linear estimators by finding the estimators having least average total MSE, averaged over parameter points. The paper of Lawless (1978) examines mean square error properties for a number of biased regression estimators and certain types of ridge estimators seem to have good MSE properties. Strawderman (1978) considers the problem of estimating the vector of regression coefficients of a linear model using generalized ridge regression estimators where the ridge constant is chosen on the basis of the data. Swamy, Mehta, Rappoport (1978) formulate an adjusted version of Hoerl-Kennard ridge regression technique in estimating coefficients in economic relationships. Trenkler (1978) defines a biased estimator called iterative estimator. Ridge and James Stein estimators are expressed from the Bayesian point of view in a review

article of Vinod (1978). Wichern and Churchill (1978) consider the performance of mechanical rules, the ridge trace and the OLS procedure for selecting k by using simulation. Coutsourides and Troskie (1979) indicate that ridge regression, principal components and shrunken estimators provide the same central t and F statistics as the OLS estimator. Draper and Nostrand (1979) review the literature of ridge regression and James Stein estimation and make critical comments. Hoerl and Kennard (1979) present a summary of the advances in the theory. Kadiyala (1979) studies the MSE and conditional mean forecasting of operational ridge regression estimators. Mitra and Ling (1979) compare the performance of some ridge estimators, biased estimators and OLS estimator by Monte Carlo experiment. Moulaert (1979) suggests an estimator, conservative estimator which is an alternative to the ordinary RE. Askin and Montgomery (1980) propose a method for combining biased and robust regression techniques. Brown and Zidek (1980) describe a multivariate version of Hoerl-Kennard ridge regression rule. In the paper of Casella (1980), conditions that are necessary and sufficient for minimaxity of a large class of ridge regression estimators are derived. A critique of some ridge regression methods is made by Smith and Campbell (1980). Thisted, Marquardt, Van Nostrand, Lindley, Obenchain, Peele and Ryan, Vinod and Gunst make comments about this critique. Trenkler (1980) compares the iteration estimator with least squares, principal components, simple shrinkage and ridge. The book of Vinod and Ullah (1981) contains introductory material on ridge regression from both the Bayesian and frequentist point of view. They mention how to choose k and how to select the correct ridge solution. Vinod, Ullah and Kadiyala (1981) show ridge estimators of Hoerl and Kennard (1970) as special cases of double f class estimators. Wang (1982) represents ridge regression estimators as special cases of Bayesian estimators. Peele and Ryan (1982) express how ridge estimators are special cases of minimax estimators and show that minimax estimator have smaller MSE than OLS. Price (1982) beholds possibility of comparing the ridge estimators with the OLS but impossibility of comparing parametric functions of ridge estimators with each other. Toutenberg (1982) displays the relationship between ridge, minimax and mixed estimators. He thinks a different prior information from the Bayesians. Ullah, Srivastava and

Chandra (1983) defines a general class of shrinkage estimators of which a Stein and a ridge type estimator is a special case. Farebrother (1984) enlarges the result of Farebrother (1976, 1978) for the purpose of comparing MSE of a generalized RE with a restricted least squares estimator. Kadiyala (1984) studies the jack-knifing method and the properties of jack-knifed RE. Baye and Parker (1984) propose a general r-k class estimator that includes the OLS estimator, the ridge regression estimator and principal component estimator as a special case. Precht and Rao (1985) evaluate the performance of MSE of different kinds of James Stein and ridge type estimators in comparison with the OLS. Necessary and sufficient conditions for the MRE in order to have smaller risk than the ordinary RE of Hoerl and Kennard (1970) is given by Pliskin (1987). Liski (1988) develops a criteria for linear admissible estimators including ridge estimators. He also develops a test procedure for choosing a linear estimator instead of an OLS. Chawla (1988) obtains necessary and sufficient condition for the matrix MSE of a generalized ridge regression estimator to be smaller than OLS. Nyquist (1988), Nomura (1988) also study the properties of jack-knifed RE. Peddada et al. (1989) consider three criteria which are Pitmann nearness, the MSE averaging over a quadratic loss function and the MSE averaging over a Mahalanobis loss function to compare OLS and ridge estimators. Noor and Mehta (1989) study four different type shrinkage estimators which are Hoerl and Kennard (1970), Hemmerle (1975), Obenchain (1975) and the authors. Firinguetti (1989) does some simulation studies on different kinds of ridge regression estimators and concludes that ridge estimators performance improves as the degree of multicollinearity increases. Hoerl and Kennard (1990) assert that obtaining the sum of squares decomposition and finding degrees of freedom for ridge regression as it is done in OLS is wrong. Miller (1990) mentions ridge and James Stein estimators from the standpoint of choosing an appropriate subset of the independent variables in fitting a linear regression model. Srivastava and Giles (1991) obtain an exact unbiased estimator for the MSE of an operational RE. Silvapulle (1991) illustrates that the ordinary RE is sensitive to the outliers in the y variable so he proposes a class of ridge-type M estimators. Chalton and Troskie (1992) carry out a simulation study with ridge regression and fractional rank estimators. Hadzivukovic (1992)

compares the performance of methods of choosing k based on ridge trace and variance inflation factors with methods dependent upon functions of sample estimates. Le Cessie and Van Houwenling (1992) use ridge estimators to estimate the parameters of a logistic regression model. Sarkar (1992) considers that his new proposed estimator combines the restricted least squares (RLS) method and the ridge regression method. Kejian (1993) considers a modification of the RE where a constant multiple d between zero and one of the LS is added to $X'y$. Sundberg (1993) shows the close relationship between first-factor continuum regression and standard ridge regression. Singh et al. (1994) propose a class of shrinkage estimators that includes the James Stein and ridge regression estimators as a special cases. Crouse, Jin and Hanumara (1995) develop an unbiased ridge estimator (URE) using a random vector of empirical prior information. Fule (1995) applies ridge regression to a problem of ecological inference. Sarkar (1996) compares the performance of the OLS, RE and PCR estimator with respect to their MMSE. He obtains necessary and sufficient conditions for the r-k class estimator to have a smaller MMSE than the OLS, RE and the PCR estimator. Kibria (1996) estimates the coefficients of a general linear regression model when the error term follows a Student t distribution. Three estimators are considered the unrestricted RE, the restricted ridge regression estimator and the preliminary test ridge regression estimator. Schipp and Toutenburg (1996) derive approximate minimax estimator based on partial constraints on the structural parameters in a single equation of a simultaneous equation model. The form of these estimators is similar to ridge type estimators. Markiewicz (1996) characterizes the general ridge estimators as linear sufficient and admissible among the set of all linear estimators for a non-singular linear model. Gross (2003) introduces a RE, restricted RE, for the vector of parameters in a linear regression model under linear restrictions. Lipovetsky and Conklin (2005) obtain a generalization of the ridge regression to two parameter model by minimizing model errors, devitions from orthogonality between regressors and errors and deviations from other desired properties of the solution. Lipovetsky (2006) considers another generalization of the one parameter ridge regression to two parameter model and its asymptotic behaviors. Clark and Troskie (2006) assess the suitability of two new

ridge estimators by means of a simulation study in their article. They compare these estimators with well known ridge estimators. Ozkale and Kaciranlar (2007) propose a new two parameter estimator and under restrictions on the parameter values they also define a new restricted two parameter estimator. In the paper of Druilhet and Mom (2008) the geometrical structure of the shrinkage factors of biased estimators are examined. They also compare shrinkage factors of ridge regression, principal component regression and partial least squares (PLS) regression in the orthogonal directions obtained by the signal to noise ratio algorithm. Sakallioglu and Kaciranlar (2008) introduce a new biased estimator for the vector of parameters in a linear regression model and discuss its properties. They show that their new biased estimator is superior to the OLS estimator, RE and the Liu estimator.

1.3. Multicollinearity

In the multiple linear regression model (1.1.) the matrix X is said to have orthogonal regressors when it is such that $X'X = I$ and nonorthogonality of regressors implies that $X'X \neq I$. Multicollinearity (or collinearity) mean linear dependence among the columns of X. A distinction is often made between exact multicollinearity and near multicollinearity defined as follows:

Exact multicollinearity exists when the rank of X is less than p, namely the columns of X denoted by $x_1, x_2, \dots, x_p$ are linearly dependent. In other words, there exist nonzero constants $a_j,\ j = 1 \dots . p$ such that

$$\sum_{j=1}^{p} a_j x_j = 0 \qquad (1.11.)$$

Near multicollinearity exists when there are nonzero constants a_j such that

$$\sum_{j=1}^{p} a_j x_j \approx 0$$

where ≈ denotes approximate equality.

When the matrix product $X'X$ is singular, its inverse does not exist. Exact multicollinearity implies that $X'X$ is singular, i.e. the smallest eigenvalue $\lambda_p = 0$. Near multicollinearity means that $X'X$ is nearly singular and the smallest eigenvalue is close to zero. When eigenvalues are different from unity, we have nonorthogonality. In particular, the smallest eigenvalue $\lambda_p \neq 1$ does not imply that $\lambda_p = 0$ or $\lambda_p \approx 0$. Thus nonorthogonality does not imply singularity or multicollinearity. However, multicollinearity ($\lambda_p = 0$) does imply nonorthogonality ($\lambda_p \neq 1$).

The numbers of observations are sometimes limited, variables do not vary over a very wide range and thus the samples often do not provide enough information. This lack of sufficient information and the corresponding loss of precision of statistical results based on such data leads to what is commonly called the multicollinearity problem. In this section either the methods proposed for detecting the presence of multicollinearity and mitigating the effects of collinearity or the statistical consequences of multicollinearity are discussed.

1.3.1. The Statistical Consequences of Collinearity

The consequences of collinear relationships among explanatory variables in a statistical model may be summarized as follows:

1. When there are one or more exact linear relationships among the explanatory variables the condition of exact multicollinearity exists, then the $X'X$ matrix is singular and the OLS estimator $\hat{\beta} = (X'X)^{-1}X'y$ is not defined. Thus we can not obtain estimates of β using the least squares estimation rule. When the variables exhibit near linear dependencies or collinearity, the following consequences result:

2. When near linear dependencies among explanatory variables exist, some elements of $(X'X)^{-1}$ may be large. The variance-covariance matrix of the OLS estimator is $\mathrm{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$ so this means that the sampling variances, standard errors and covariances of the OLS estimator may be large. Large standard errors for the OLS estimators imply that sampling variability is high, interval estimates are wide

and the information provided by the sample data about the unknown parameters is relatively imprecise. To explain high variance of $\hat{\beta}$ clearly, let us define the MSE which is the trace of variance-covariance matrix of the OLS estimator:

$$MSE(\hat{\beta}) = \sigma^2 Tr(X'X)^{-1}.$$

In terms of the eigenvalues of $X'X$ we have

$$MSE(\hat{\beta}) = \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1} > \sigma^2 \lambda_p^{-1} \quad . \tag{1.12.}$$

For near multicollinearity $\lambda_P \to 0$ and (1.12.) implies that MSE $\to \infty$ which means that $\hat{\beta}$ is subject to very large variance. Often this is revealed by the low values of the usual t ratios whose denominator has the variance inflation factors. High variance of regression coefficients would generally mean that the null hypothesis $H_0: \beta_i = 0$ will be more likely to be accepted.

3. When estimator standard errors are large, it is likely that the usual t-tests will lead to the conclusion that parameter values are not significantly different from zero. This outcome occurs despite possibly high R^2 or F-values indicating significant explanatory power of the model for this combination of variables. The problem is that collinear variables do not provide enough information to estimate their separate effects.

Multicollinearity can result the estimated regression coefficients appearing to have the wrong sign (Farrar and Glauber (1967), Mullet (1976)) i.e. opposite to be a prior expectations of the researcher. Wrong sign may be due to incorrect specification, outliers, serial correlation, etc. Extreme caution is needed in relying heavily on the appearance of wrong signs because wrong sign may not even be wrong.

Visco (1978) and Leamer (1975) have shown by algebraic methods that omitting a variable with relatively low t value cannot correct the wrong sign of a regressor having a higher t value. However, replacing a less significant regressor by

another can indeed remove the wrong sign problem in certain cases. Also omission of two collinear regressors at the same time can change the signs of significant regressors and even make them insignificant.

When important practical decisions are based on the results of regression analysis, the researcher needs a stable or robust result. The concept of stability of regression coefficient values can be rigorously defined by using some classical concepts in perturbation theory. The conventional tests of significance do not provide adequate information about this kind of stability. In fact, Vinod (1981) shows that Student's t values are themselves quite unstable.

4. Estimators may be very sensitive to the addition or deletion of a few observations, or deletion of an apparently insignificant variable.

1.3.2. Identifying Multicollinearity

In this section, methods that have been proposed for detecting the presence and nature of multicollinearity are discussed.

Examination of Correlation Matrix: Sample correlation coefficients between pairs of explanatory variables can indicate linear associations between them. If the regressors x_i and x_j are nearly dependent, then $|r_{ij}|$, off-diagonal elements in $X'X$, will be near unity. However this way does not give any insights into collinear relationships involving three or more variables.

Variance Inflation Factors: Diagonals of $(X'X)^{-1}$ are called variance inflation factors (VIF) by Marquardt (1970). For the standardized data, $X'X$ is the correlation matrix and the (i,j)th element of the inverse $(X'X)^{-1}$ may be denoted by a superscript notation as r^{ij}. Farrar and Glauber (1967) were the first to suggest looking at the values of r^{jj} to diagnose multicollinearity. Marquardt (1970) suggests a rule of thumb that $VIF_j = r^{jj} > 5$ indicates harmful multicollinearity. Theil (1971) shows that

$$r^{jj} = \frac{1}{(1-R_j{}^2)\|x_j\|^2}$$

where $\|x_j\|^2 = x_j'x_j$ and where ${R_j}^2$ represents the squared multiple correlation coefficient when x_j (j-th column of X) is regressed on the remaining $p-1$ regressors. When multicollinearity is present, there is a linear relation among the regressors. Since multiple correlation measures the explanatory strength of such a linear relationship, ${R_j}^2$ will be large (close to 1), the denominator of r^{jj} will be close to zero, and hence r^{jj} will be large.

Eigensystem Analysis of $X'X$: An analysis of the characteristics roots and vectors of $X'X$ can reveal the presence and perhaps the nature of multicollinearity. The condition number of $X'X$ is

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}.$$

There is no serious problem with multicollinearity if the condition number is less than 100. However condition number between 100 and 1000 imply moderate to strong multicollinearity and if κ exceeds 1000, severe multicollinearity is indicated. The condition indices of the $X'X$ matrix are $\kappa_{j=}\frac{\lambda_{max}}{\lambda_j}$, $j = 1,2,\dots p$. Eigensystem analysis can also be used to identify the nature of the near-linear dependencies in data. The $X'X$ matrix may be decomposed as $X'X = T\Lambda T'$ where Λ is a $p \times p$ diagonal matrix whose main diagonal elements are the eigenvalues λ_j $(j = 1,2,\dots,p)$ of $X'X$ and T is a $p \times p$ orthogonal matrix whose columns are the eigenvectors of $X'X$. If the eigenvalue λ_j is close to zero, indicating a near linear dependence in the data, the elements of the associated eigenvector a_j describe the nature of this linear dependence. Specifically the elements of the vector a_j are the coefficients $a_1, a_2 \dots, a_p$ in (1.11.).

Belsley, Kuh and Welsch (1980) propose a similar approach for diagnosing multicollinearity. The $n \times p$ X matrix may be decomposed as $X = UDT'$ where U is $n \times p$, T is $p \times p$, $U'U = I$, $T'T = I$ and D is a $p \times p$ diagonal matrix with nonnegative diagonal elements μ_j , $j = 1,2,\dots,p$. These are called the singular

values of X and $X = UDT'$ is called the singular value decomposition of X. The singular value decomposition is closely related to the concepts of eigenvalues and eigenvectors, since $X'X = (UDT')'UDT' = TD^2T' = T\Lambda T'$. So that the squares of the singular values of X are the eigenvalues of $X'X$ defined earlier and U is a matrix whose columns are the eigenvectors associated with p nonzero eigenvalues of $X'X$. Belsley, Kuh and Welsch (1980) define the condition indicies of the X matrix as $n_j = \frac{\mu_{max}}{\mu_j}$, $j = 1,2,\ldots,p$.

The variance-covariance matrix of $\hat{\beta}$ is $Var(\hat{\beta}) = \sigma^2(X'X)^{-1} = \sigma^2 T\Lambda^{-1}\mathrm{T}'$ and the variance of the j-th regression coefficient is the jth diagonal element of this matrix:

$$Var(\hat{\beta}_j) = \sigma^2 \sum_{i=1}^{p} \frac{{t_{ji}}^2}{{\mu_i}^2} = \sigma^2 \sum_{i=1}^{p} \frac{{t_{ji}}^2}{\lambda_i}.$$

Note also apart from σ^2, the jth diagonal elements of $T\Lambda^{-1}\mathrm{T}'$ is the jth variance inflation factor, so

$$VIF_j = \sum_{i=1}^{p} \frac{{t_{ji}}^2}{{\mu_i}^2} = \sum_{i=1}^{p} \frac{{t_{ji}}^2}{\lambda_i}$$

Clearly one or more small singular values (or small eigenvalues) can dramatically inflate the variance of $\hat{\beta}_j$. Belsley, Kuh and Welsch suggest using variance-decomposition proportions, defined as

$$\pi_{ij} = \frac{{t_{ji}}^2/{\mu_i}^2}{VIF_j}, \quad j = 1,2,\ldots,p$$

as measures of multicollinearity. Condition indices greater than 30 and variance-decomposition proportions greater than 0.5 are recommended guidelines.

The advantages of the spectral and singular value decompositions are that not only can they isolate which explanatory variables are interrelated, but they also deal

with the issues of what constitutes a small characteristic root, or singular value and whether collinearity is harmful from the point of view of a particular application.

1.3.3. Solutions to the Multicollinearity Problem

Several techniques have been proposed for dealing with the problems caused by multicollinearity.

1.3.3.1. Traditional Solutions to the Multicollinearity Problem

Additional Sample Information: The problem of multicollinearity is that the data do not contain enough information about the individual effects of control variables to permit us to estimate the parameters of the statistical model precisely. One solution is to obtain more information and include it in the analysis. One form the new information can take is more, and better, sample data. Unfortunately collecting data is not always possible because of economic constraints or because the process being studied is no longer available for sampling. The new observations may suffer the same collinear relationships and provide little in the way of new, independent information. An alternative to introducing new sample information is to add structure by introducing non-sample information in the form of restrictions on the parameters.

Exact linear Constraints: One effect of using exact parameter restrictions is to reduce the sampling variability of the estimators, a desirable and given multicollinear data. The imposition of binding constraints, even if incorrect restrictions produce biased parameter estimators. Exact linear restrictions may be employed in case of both extreme and near-extreme multicollinearity.

Stochastic Linear Restrictions: Stochastic linear restrictions arise from prior statistical information, usually in the form of previous estimates of parameters that are also included in a current model. For the purposes of this section these linear stochastic restrictions act just like additional observations, if the statistical information is unbiased, and this they may serve to aid in both situations of extreme and near extreme multicollinearity.

Linear Inequality Restrictions: Their use can improve the precision of estimation when near-extreme multicollinearity is present, but cannot, in the extreme multicollinearity situation, make it possible to estimate a function that cannot be estimated because inexact restrictions may not effect the dimensionality of the parameter space.

1.3.3.2. Ad Hoc Solutions to the Multicollinearity Problem

It is noted that conventional procedures for combining sample and non sample information can be used in an attempt to reduce the ill effects of multicollinearity. One of the ad hoc solution is the RE because the information it introduce is specific to sample in hand.

2. RIDGE REGRESSION

2.1. The Need for Alternatives to the Ordinary Least Squares Estimator

Statistician often use the method of least squares to estimate the parameters of a linear regression model. The OLS estimator is obtained by minimizing the sum of squares of the error terms or equivalently the sum of squares of the differences between the observed and the fitted values. Thus

$$F(\beta) = (y - X\beta)'(y - X\beta)$$

is minimized by differentiation and setting the result to zero. Thus normal equation

$$X'X\beta = X'y$$

is solved to obtain the OLS estimator

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{2.1.}$$

The squared distance from $\hat{\beta}$ to the true parameter vector β is

$${L_1}^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta)$$

and the expected squared distance, $E({L_1}^2)$, is

$$E({L_1}^2) = \sigma^2 Tr(X'X)^{-1}.$$

Since the trace of a matrix is also equal to the sum of its eigenvalues, $E({L_1}^2)$ is equal to:

$$E\left(L_1{}^2\right) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j} \qquad (2.2.)$$

where $\lambda_j > 0, j = 1, \dots, p$ are the eigenvalues of $X'X$.

If the $X'X$ matrix is ill-conditioned because of multicollinearity, at least one of the λ_j will be small and (2.2.) implies that the distance from the least squares estimate $\hat{\beta}$ to the true parameter β may be large. Namely the vector $\hat{\beta}$ is generally longer than the vector β. This implies that the method of least squares produces estimated regression coefficients that are too large in absolute value.

OLS estimator is an unbiased estimator so

$$E\left(\hat{\beta}\right) = \beta$$

and its variance-covariance matrix is :

$$Var\left(\hat{\beta}\right) = \sigma^2 (X'X)^{-1}.$$

The Gauss – Markov theorem establishes that the OLS estimator of β is best linear unbiased estimator (BLUE). Namely, $\hat{\beta}$ has the smallest variance among the class of all unbiased estimators. However this theorem does not ensures that this variance will be small.

Multicollinearity results in large variances for the OLS estimators of the regression coefficients implying that confidence intervals on β would be wide and the point estimate $\hat{\beta}$ is very unstable. A biased estimator $\hat{\beta}^*$ which has a smaller variance than the unbiased estimator $\hat{\beta}$ can be found to ease this problem. MMSE of $\hat{\beta}^*$ is

$$MMSE\left(\hat{\beta}^*\right) = E\left(\hat{\beta}^* - \beta\right)\left(\hat{\beta}^* - \beta\right)' = Var\left(\hat{\beta}^*\right) + Bias\left(\hat{\beta}^*\right)Bias\left(\hat{\beta}^*\right)'.$$

Decrease in variance will be more than the increase in bias of $\hat{\beta}^*$ so MMSE of $\hat{\beta}^*$ will be less than the variance of $\hat{\beta}$ implying that $\hat{\beta}^*$ is a more stable estimator of β. Besides this, confidence intervals on β will be much narrower using the biased estimator. One of the procedures of obtaining biased estimators is ridge regression.

2.2. Ridge Estimator

The RE

$$\hat{\beta}_k = \left(X'X + kI_p\right)^{-1} X'y\,, \;\; k \geq 0 \qquad (2.3.)$$

is a biased linear estimator introduced by Hoerl and Kennard (1970a). In the presence of collinearity this estimator can yield better estimates than the OLS estimator $\hat{\beta}$. The use of this estimator is motivated from the fact that the matrix $X'X$ is stabilized when a positive number k is added to each of its main diagonal elements. By increasing k, the ridge of matrix $X'X + kI$ is raised so that $X'X + kI$ resembles more and more a diagonal matrix, which has an optimal condition. The ultimate goal is to find some k which is large enough to reduce the variance compared to the OLS estimator, but which is small enough to produce some acceptable low bias.

Originally Hoerl and Kennard (1970) obtained the RE by finding the point on the ellipsoid centered at the OLS estimator $\hat{\beta}$ that is closest to the origin. This consist of minimizing the distance $\beta'\beta$ of a vector of parameters β from the origin subject to the side condition:

$$\left(\beta - \hat{\beta}\right)' X'X\left(\beta - \hat{\beta}\right) = \Phi_0.$$

The optimization problem is solved using Lagrange multipliers. Differentiate

$$L = \beta'\beta + \frac{1}{k}\left[\left(\beta - \hat{\beta}\right)' X'X\left(\beta - \hat{\beta}\right) - \Phi_0\right]$$

to obtain

$$\beta + \frac{1}{k} X'X\left(\beta - \hat{\beta}\right) = 0. \qquad (2.4.)$$

The RE in (2.3.) is obtained by solving (2.4.) for β.

When the squared length of the regression vector β is fixed at R^2, in order to make the sum of squares minimum, the Lagrange multipliers is used to solve the problem:

$$L = (y - X\beta) + \frac{1}{k}(\beta'\beta - R^2).$$

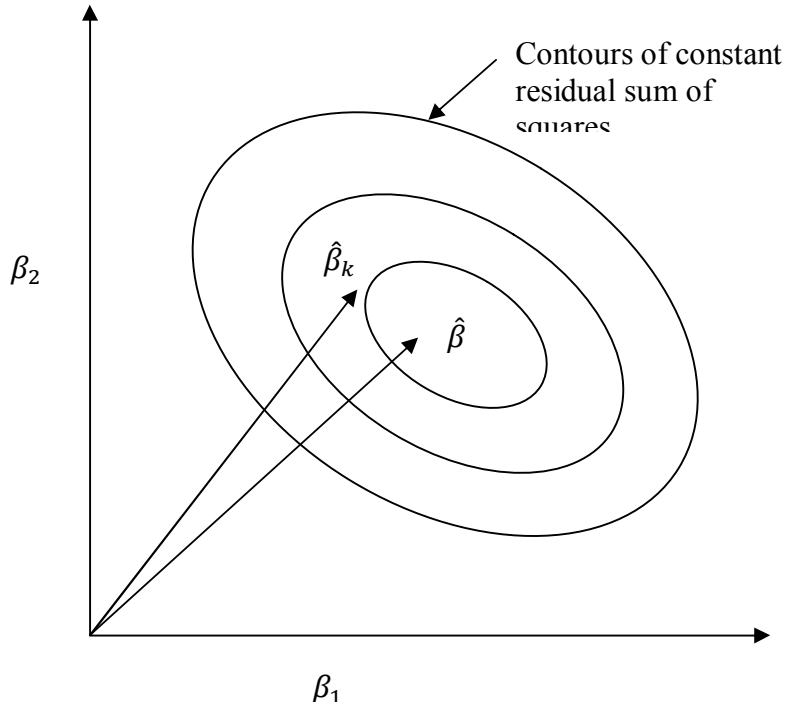

Figure 2.1. A geometrical interpretation of ridge regression (Montgomery (2001))

Figure 2.1. is given to display the geometry of ridge regression in a regression model with two regressors. The residual sum of squares is minimum at the point at

the center of the ellipses. The ridge estimate $\hat{\beta}_k$ is the shortest vector from the origin that produces a residual sum of squares equal to the value represented by the small ellipse where the residual sum of squares is constant at some value greater than the minimum.

2.2.1. Properties of Ridge Estimator

The RE $\hat{\beta}_k$ can be written in the form

$$\hat{\beta}_k = (X'X + kI)^{-1}X'y = WX'y, \quad k \geq 0$$

and also in an alternative form

$$\hat{\beta}_k = [I + k(X'X)^{-1}]^{-1}\hat{\beta} = Z\hat{\beta}, \quad k \geq 0. \tag{2.5.}$$

By writing $Z = WX'X$ the relationship between Z and W is

$$Z = I - k(X'X + kI)^{-1} = I - kW.$$

Since $E(\hat{\beta}_k) = E(Z\hat{\beta}) = Z\beta$, $\hat{\beta}_k$ is a biased estimator and k is the biasing parameter.

Using (2.5.) in the equation the MSE of RE which is just the expected squared distance from $\hat{\beta}_k$ to β is obtained as

$$\begin{aligned} E\left({L_1}^2(k)\right) &= E\left[(\hat{\beta}_k - \beta)'(\hat{\beta}_k - \beta)\right] \\ &= E\left[(\hat{\beta} - \beta)'Z'Z(\hat{\beta} - \beta)\right] + (Z\beta - \beta)'(Z\beta - \beta) \\ &= \sigma^2 Tr(X'X)^{-1}Z'Z + \beta'(Z - I)'(Z - I)\beta \end{aligned}$$

$$= \sigma^2[Tr(X'X + kI)^{-1} - kTr(X'X + kI)^{-2}] + k^2\beta'(X'X + kI)^{-2}\beta$$
$$= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j+k)^2} + k^2\beta'(X'X + kI)^{-2}\beta$$
$$= \gamma_1(k) + \gamma_2(k). \qquad (2.6.)$$

The second term $\gamma_2(k)$ in (2.6.) is the squared distance from $Z\beta$ to β. If $k = 0$, namely Z is equal to I, the second term will be zero. $k = 0$ means that RE is equal to OLS estimator. The first term $\gamma_1(k)$ in (2.6.) is the sum of the variances at the parameter estimates since

$$\hat{\beta}_k = Z\hat{\beta} = Z(X'X)^{-1}X'y$$

and then

$$Var(\hat{\beta}_k) = Z(X'X)^{-1}X'Var(y)X(X'X)^{-1}Z' = \sigma^2 Z(X'X)^{-1}Z'$$
$$Tr\left(Var(\hat{\beta}_k)\right) = \sigma^2 Tr[(X'X + kI)^{-1}X'X(X'X + kI)^{-1}]$$
$$= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2}.$$

The total variance decreases as k increases, while the squared bias increases with k.There are values of k for which the mean square error is less for $\hat{\beta}_k$ than it is for the usual solution $\hat{\beta}$. The function $\gamma_1(k)$ is a monotonic decreasing function of k, while $\gamma_2(k)$ is monotonic increasing.

Theorem 2.1: The total variance $\gamma_1(k)$ is a continuous, monotonically decreasing function of k.

Corollary 2.1: The first derivative with respect to k of the total variance $\gamma_1'(k)$, approaches $-\infty$ as $k \to 0^+$ and $\lambda_p \to 0$.

Both the theorem and the corollary are proved by use of $\gamma_1(k)$ and its derivative:

$$\lim_{k\to 0^+}\left(\frac{d\gamma_1}{dk}\right) = -2\sigma^2\sum\frac{1}{{\lambda_j}^2}.$$

Thus, $\gamma_1(k)$ has a negative derivative which approaches $-2\sigma^2$ as $k \to 0^+$ for an orthogonal $X'X$ and approaches $-\infty$ as $X'X$ becomes ill-conditioned and $\lambda_p \to 0$.

Theorem 2.2: The squared bias $\gamma_2(k)$ is a continuous, monotonically increasing function of k.

Proof: If Λ is the matrix of eigenvalues of $X'X$ and P is the orthogonal transformation such that $X'X = P'\Lambda\mathrm{P}$, then $\gamma_2(k) = k^2\beta'(X'X + kI)^{-2}\beta$ is equal to

$$\gamma_2(k) = k^2\sum_{j=1}^{p}\frac{{\alpha_j}^2}{(\lambda_j+k)^2} \qquad (2.7.)$$

where $\alpha = P\beta$. Since $\lambda_j > 0$ for all j and $k \geq 0$, each element $(\lambda_j + k)$ is positive and there are no singularities in the sum. Clearly, $\gamma_2(0) = 0$. Then $\gamma_2(k)$ is a continuous function for $k \geq 0$. For > 0 , $\gamma_2(k)$ can be written as

$$\gamma_2(k) = \sum_{j=1}^{p}\frac{{\alpha_j}^2}{[1+(\lambda_j/k)]^2} \ . \qquad (2.8.)$$

Since $\lambda_j > 0$ for all j, the functions λ_j/k are clearly monotone decreasing for increasing k and each term of, $\gamma_2(k)$ is monotone increasing. So $\gamma_2(k)$ is monotone increasing.

Corollary 2.2: The squared bias $\gamma_2(k)$ approaches $\beta'\beta$ as an upper limit.

Proof: From (2.8.) $\lim_{k\to\infty}\gamma_2(k) = \sum_{j=1}^{p}{\alpha_j}^2 = \alpha'\alpha = \beta'P'P\beta = \beta'\beta$.

Corollary 2.3: The derivative $\gamma_2'(k)$ approches zero as $k \to 0^+$.

Proof: From (2.7.) it is readily established that

$$\frac{d\gamma_2(k)}{dk} = 2k\sum_{j=1}^{p}\frac{\lambda_j{\alpha_j}^2}{(\lambda_j+k)^3} \ . \qquad (2.9.)$$

Each term in the sum $2k\,\lambda_j{\alpha_j}^2/\left(\lambda_j+k\right)^3$ is a continuous function and the limit of each term as $k \to 0^+$ is zero.

Theorem 2.3: (Existence Theorem) There always exists a $k > 0$ such that

$$E\left[{L_1}^2(k)\right] < E\left[{L_1}^2(0)\right] = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$$

Proof: From (2.7.) and (2.9.)

$$\frac{dE\left[{L_1}^2(k)\right]}{d_k} = \frac{d\gamma_1(k)}{d_k} + \frac{\mathrm{d}\gamma_2(k)}{d_k}$$

$$= -2\sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{\left(\lambda_j+k\right)^3} + 2k \sum_{j=1}^{p} \frac{\lambda_j {\alpha_j}^2}{\left(\lambda_j+k\right)^3} \qquad (2.10.)$$

First note that $\gamma_1(0) = \sigma^2 \sum_{j=1}^{p} \left(\frac{1}{\lambda_j}\right)$ and $\gamma_2(0) = 0$. In Theorem 2.1 and Theorem 2.2 it was established that $\gamma_1(k)$ and $\gamma_2(k)$ are monotonically decreasing and increasing, respectively. Their first derivatives are always non-positive and non-negative, respectively. Thus to prove the theorem, it is only necessary to show that there always exists a $k > 0$ such that $\frac{dE\left[{L_1}^2(k)\right]}{d_k} < 0$. The condition for this is shown by (2.10.) to be:

$$k < \frac{\sigma^2}{\alpha_{max}^2}\,. \qquad (2.11.)$$

2.2.2. The Choice of the Ridge Parameter

There exist several ways to determine k. For determining an appropriate k, two methodical approaches are subjective and objective methods.

2.2.2.1. The Ridge Trace (Subjective Method)

Hoerl and Kennard (1970a) suggested the ridge trace method which considers the p elements of the RE $\hat{\beta}_k$ as functions of k, all of them plotted into one figure. Theil (1963) seems to have independently suggested a similar plot. When data is collinear the regression coefficients appear to change very rapidly as k varies for small values. At some value of k they stabilize; the rate of change slows down to almost zero. This value of k is picked from examining the plot. The user then chooses the value of k at which the functions begin to stabilize. This methods of choosing k is subjective because two people examining the plot might have different opinions as to where the regression coefficients stabilize. The judgment of the ridge trace depends on the user, but also on the range of values k for which the functions are drawn. The impression of stability will be quite different when the ridge trace is considered for different ranges.

Hoerl and Kennard (1970b) recommended standardizing the data before plotting on a ridge trace, so as to retain numerical comparability of regression coefficients. Vinod (1981) points out that the appearance of a ridge trace that does not plot standardized regression coefficients may be dramatically changed by a simple translation of the origin and scale transformation of the variables.

There are also approaches for finding an objective judgment of the ridge trace. For this the function is considered as

$$\varphi(k) = \left(\hat{\beta}_k - \hat{\beta}\right)'\left(\hat{\beta}_k - \hat{\beta}\right) = \left\|\hat{\beta}_k - \hat{\beta}\right\|^2.$$

The derivative $d\varphi(k)/dk$ can be seen as the velocity of change in φ under variation of k. If the maximum of $d\varphi(k)/dk$ is unique and attained for k^* (position of maximum instability of the ridge trace), the first inflection point k^{**} of $d\varphi(k)/dk$ greater than k^* is defined to be the position of the beginning of a stabilize ridge trace.

2.2.2.2. Estimation of k (Objective Method)

The ridge trace, the subjective method, doesn't only depend on the actual value of y, besides this, it depends on the user's judgement. An estimator $\hat{k}$, as a choice of k, can be seen as an objective method in the sense that the choice of k does only depend on y.

For any given parameter values ($\beta \neq 0,\ \sigma^2$) there exists some optimal k_{opt} for which

$$MMSE\left(\beta, \hat{\beta}_{k_{opt}}\right) \leq MMSE\left(\beta, \hat{\beta}_k\right)$$

for any $k \geq 0$. Generally finding a closed form formula for k_{opt} is not possible, however, considering $X'X = I_p$ to be true, k_{opt} can be determined as a function of $\beta \neq 0$ and σ^2 as it is shown in the theorem:

Theorem 2.4: Under the linear regression model with assumptions, the inequality

$$MMSE\left(\beta, \hat{\beta}_{k_{opt}}\right) \leq MMSE\left(\beta, \hat{\beta}_k\right),\ \forall\ \ k \geq 0$$

is true for $k_{opt} = p\sigma^2/\beta'\beta$.

Proof: If $X'X = I_p$ is true than

$$Tr\left[MMSE\left(\beta, \hat{\beta}_{k_{opt}}\right)\right] = \frac{p\sigma^2}{(1+k)^2} + \frac{k^2}{(1+k)^2}\beta'\beta.$$

First, the derivative of this function with respect to k is determined. Then putting the derivative equal to zero and solving for k, $k = p\sigma^2/\beta'\beta$ is obtained.

Based on this theorem, some ad-hoc possibilities for determining an estimator for k can be considered.

Estimator from Hoerl and Kennard (1970a): As it is shown in Theorem 2.3, Hoerl and Kennard (1970a) suggested an estimator for k as in (2.11.). Estimator for k can

be determined for general linear regression problem. It is always possible to reduce the general linear regression problem as defined in the introduction to a canonical form (1.2.).

By defining $(L_1{}^*)^2 = (\hat{\alpha}^* - \alpha)'(\hat{\alpha}^* - \alpha)$, it can be shown that the optimal values for k_i will be $k_i = \frac{\sigma^2}{\alpha_i^2}$. Namely estimators for k are $\hat{k}_i = \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}$.

Estimator from Theobald (1974): Hoerl and Kennard (1970a) has the justification that there is always a positive value of k for which the mean square error is less than the least squares value. Nelder (1972) criticized this for its use of mean square errors of the components of $\hat{\beta}_k$ are put together in an unweighted sum of coefficient mean square errors or to compare the values of

$$E(\tilde{\beta} - \beta)' B(\tilde{\beta} - \beta) \tag{2.12.}$$

where B is nnd matrix, for different estimators $\tilde{\beta}$. Theobald aims to established a relationship between mean square error function in (2.12.) and the second-order moment matrix $E(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'$.

Theorem 2.5: The following conditions are equivalent

(a) $M_1 - M_2$ is non - negative definite (nnd)

(b) $m_1 - m_2 \geq 0$ for all nnd B

where $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are two estimators and $M_j = E(\tilde{\beta}_j - \beta)(\tilde{\beta}_j - \beta)'$, $j = 1,2$ are the second order moment matrices and $m_j = E(\tilde{\beta}_j - \beta)' B(\tilde{\beta}_j - \beta)$, $j = 1,2$ are the overall measure of the mean square error. So using $M(k)$ for $\hat{\beta}_k$ and $M(0)$ for $\hat{\beta}$ the following theorem shows that $\hat{\beta}_k$ and $\hat{\beta}$ satisfy condition (a) of Theorem 2.5

Theorem 2.6: There exists $K > 0$ such that $M(0) - M(k)$ is positive definite whenever $0 < k < K$. In here, $k < \frac{2\sigma^2}{\beta'\beta}$ is a sufficient condition.

That is to say, Theobald (1974) suggested an estimator for k which is $\hat{k}_T = \frac{2\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$.

Estimator from Hoerl, Kennard and Baldwin (1975): Hoerl, Kennard and Baldwin (1975) used the two results in improving an algorithm for selecting k. First, considering $X'X = I_p$ to be true as it is indicated in Theorem 2.4 k is obtained as $k = \frac{p\sigma^2}{\beta'\beta}$. Secondly, in general form optimal values for k_j is $k_j = \frac{\sigma^2}{\alpha_j^2}$. For combining these individual k_i in some way to obtain a single value of k, Hoerl, Kennard and Baldwin (1975) proposed the use of harmonic mean of the k_j values. The arithmetic mean will not be a good choice because very small α_j that have no predicting power would generate a very large k resulting in too much bias. The harmonic mean is calculated using the following formula:

$$k = \frac{1}{\frac{1}{p}\sum_{j=1}^{p}\frac{1}{k_j}} = \frac{1}{\sum_{j=1}^{p}\frac{\alpha_j^2}{\sigma^2}} = \frac{p\sigma^2}{\sum_{j=1}^{p}\alpha_j^2} = \frac{p\sigma^2}{\alpha'\alpha} = \frac{p\sigma^2}{\beta'\beta}.$$

So, an estimator of k for automatic selection of choice is $\hat{k}_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$.

As a result of simulation study, using the biasing parameter $\hat{k}_{HKB}$, ridge estimator produces estimates with smaller MSE than OLS with a probability greater than 0,5.

Estimator from Hoerl, Kennard and Baldwin (1976): The estimator $\hat{k}_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$ which is proposed by Hoerl, Kennard and Baldwin (1975**)** often produces small value of k. In presence of multicollinearity, the squared length $\hat{\beta}'\hat{\beta}$ of $\hat{\beta}$ is often greater than the squared length $\beta'\beta$ of βleading to the use of a very small k value. An iteration of estimation process is proposed to remedy this problem. Considering a sequence of estimates of β and k:

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

$$\hat{\beta}_{k_0} : k_1 = \frac{p\hat{\sigma}^2}{\hat{\beta}'_{k_0}\hat{\beta}_{k_0}}$$

$$\hat{\beta}_{k_1} : k_2 = \frac{p\hat{\sigma}^2}{\hat{\beta}'_{k_1}\hat{\beta}_{k_1}}$$

....

$$\hat{\beta}_{k_t} : k_{t+1} = \frac{p\hat{\sigma}^2}{\hat{\beta}'_{k_t}\hat{\beta}_{k_t}}.$$

A condition must be given in order to end the sequence. The sequence is continued if $\frac{k_{j+1}-k_j}{k_j} > \delta$ otherwise it is terminated and $\hat{\beta}_{k_j}$ is used. $\delta \geq 0$ since $k_{j+1} > k_j$.

An equivalent condition is $\hat{\beta}'_{k_j}\hat{\beta}_{k_j} < \frac{1}{1+\delta}\hat{\beta}'_{k_{j-1}}\hat{\beta}_{k_{j-1}}$. If the condition is not true, the sequence is terminated and $\hat{\beta}_{k_j}$ is used. δ should be choosen providing the estimate to have good properties. When $X'X$ is orthogonal or close to it, the one iteration k estimate will be sufficient. However, as the degree of ill conditioning of $X'X$ increases, the probability of $\hat{\beta}(k)$ being closed to β decreases and further iterations are required. Viz, δ should decreases as the ill-conditioning increases. As a consequences of few simulations $\delta = AT^{-b}, b > 0$ will provide a δ for ending the estimation sequence. T is defined as $T = Tr\,(X'X)^{-1}/p$. Specifically it is suggested to be $\delta = 20T^{-1.30}$.

Simulation study is required to evaluate the properties of the iterative estimation procedure. The simulation procedure, given in Hoerl, Kennard and Baldwin (1975), about the performance of an algorithm is adapted. As a consequences, $\delta -$ criterion makes important decreases in MSE comparing to least squares or a single iteration. Besides this, an error distribution with a smaller standard deviation is produced.

Estimator from McDonald Galarneau (1975): Existence of k values for which the RE have smaller MSE than OLS is showed in Hoerl and Kennard (1970a). The ridge parameter k depends on the unkown coefficient vector β and σ^2. There is no guarantee that a constant value of k produces RE better than OLS for all unkown coefficient vectors. Several rules for selecting k is suggested so as to help researcher in a way to reach an acceptable choice of k. First rule is choosing $k = 0$ which corresponds to least squares estimation. In second and third rule, the choice of k is

made in such a way that the squared length of the corresponding RE equals an estimated squared length of β. An unbiased estimator of $\beta'\beta$ can be obtained as

$$Q = \hat{\beta}'\hat{\beta} - \hat{\sigma}^2 \sum_{j=1}^{p} \lambda_j^{-1}.$$

$\hat{\sigma}^2$ is obtained by $\hat{\sigma}^2 = \phi(0)/(n-p-1)$, where $\phi(0)$ is the residual sum of squares for the OLS estimate. Then the second rule is choosing k such that $\hat{\beta}_k'\hat{\beta}_k = Q$, if $Q > 0$ otherwise choosing $k = 0$. The third rule is choosing k such that $\hat{\beta}_k'\hat{\beta}_k = Q$, if $Q > 0$ otherwise choosing $k = \infty$.

If Q is positive the second and the third rules are same but the second rule defines OLS estimate and the third rule specifies that the zero vector estimates β when Q is negative or zero.

In the simulations described in here the performance of the estimators can be observed as one the factors-variance of random error, the correlations among explanatory variables and the unknown coefficient vector is changed while the remaining two are fixed. Evaluated ridge-type-estimators reduce MSE in some cases and increase in others.

Estimator from Lawless and Wang (1976): Let $(X'X)^{1/2}$ be the uniquely determined symmetric pd matrix such that $(X'X)^{1/2}(X'X)^{1/2} = X'X$ and $(X'X)^{-1/2} = \left[(X'X)^{1/2}\right]^{-1}$. Defining $v = (X'X)^{1/2}\beta$ and $Z = X(X'X)^{-1/2}$ linear regression model $y = X\beta + \varepsilon$ can be written as $y = Zv + \varepsilon$ under which $Z'Z = I_p$ is satisfied. In this model the corresponding estimator $\hat{k}_{HKB}$ is given by

$$\hat{k}_{LW} = \frac{p\hat{\sigma}^2}{\hat{v}'\hat{v}} = \frac{p\hat{\sigma}^2}{\hat{\beta}'X'X\hat{\beta}} .$$

Lawless and Wang (1976) give a different, Bayesian oriented motivation for this estimator.

3. ESTIMATORS RELATED WITH RIDGE ESTIMATOR

3.1. Generalized Ridge Estimator

Hoerl and Kennard (1970a) proposed the general form of the ridge regression. From the canonical model (1.2.) the generalized ridge estimator (GRE) is defined as

$$\hat{\alpha}^* = [Z'Z + K]^{-1}Z'y \qquad (3.1.)$$

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

Optimal values for the $k_i's$ in $\hat{\alpha}^*$ can be considered to be those $k_i's$ that minimize

$$E(L_1^2) = E[(\hat{\alpha}^* - \alpha)'(\hat{\alpha}^* - \alpha)].$$

$E(L_1^2)$ is minimized when

$$k_i = \sigma^2/\alpha_i^2 \qquad i = 1,..,p.$$

An alternative procedure is described by the formula

$$k_{i(j)} = \hat{\sigma}^2/\left(\hat{\alpha}_{i(j)}^*\right)^2 \quad i = 1,..,p$$

where j denotes the j-th iterate. $\hat{\alpha}_{i(0)}^* = \hat{\alpha}_i,\ i = 1,..,p$ where $\hat{\alpha}_i$ is the OLS estimate of α_i. The $k_{i(j)}$ values are used in equation $\hat{\alpha}^*$ to obtain the next $\hat{\alpha}_{i(j+1)}^*$ values for use in $k_{i(j)}$.

Hemmerle (1975) showed that an explicit solution is available for the limiting $\hat{\alpha}_{(j)}$ values so that it is not necessary to iterate in order to obtain these values. This solution will depend upon certain convergence/divergence conditions. Let

$$B=\begin{bmatrix} [Z'y]_1 & 0 & \dots & 0 \\ 0 & [Z'y]_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & [Z'y]_p \end{bmatrix} \text{ and } A_j=\begin{bmatrix} \hat{\alpha}_{1(j)}^{\ *} & 0 & \dots & 0 \\ 0 & \hat{\alpha}_{2(j)}^{\ *} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\alpha}_{p(j)}^{\ *} \end{bmatrix}$$

be in a diagonal form. As a consequence $A_0 = \Lambda^{-1}B$ and the iterative procedure can be described by the matrix formula

$$A_{j+1} = \left[\Lambda + \hat{\sigma}^2 A_j^{-2}\right]^{-1} B = \left[\Lambda + \hat{\sigma}^2 A_j^{-2}\right]^{-1} \Lambda A_0. \quad (3.2.)$$

(3.2.) can be written as

$$A_{j+1} = \left(I + \hat{\sigma}^2 \Lambda^{-1} A_j^{-2}\right)^{-1} A_0 \ . \quad (3.3.)$$

Letting $D = \frac{\Lambda}{\hat{\sigma}^2}$, (3.3.) becomes $A_{j+1} = \left(I + D^{-1} A_j^{-2}\right)^{-1} A_0$ and an expression for A_{j+1}^{-2} is given by

$$A_{j+1}^{-2} = A_0^{-2}\left(I + D^{-1} A_j^{-2}\right)^2. \quad (3.4.)$$

Multiplying both sides of (3.4.) by D^{-1} then yields

$$D^{-1} A_{j+1}^{-2} = D^{-1} A_0^{-2}\left(I + D^{-1} A_j^{-2}\right)^2 \quad (3.5.)$$

so that putting $E_j = D^{-1} A_j^{-2}$ in (3.5.) the iterative procedure is reduced to the simple formula

$$E_{j+1} = E_0\left(I + E_j\right)^2. \quad (3.6.)$$

Then assume that $\hat{\alpha}_i \neq 0$ for all i and that the iterative procedure is convergent such that

$$\lim_{j\to\infty} E_j = E^*. \tag{3.7.}$$

From (3.6.) and (3.7.) the relationship is obtained as $E^* = E_0(I + E^*)^2$. From this relationship e^* is obtained as $e^* = \frac{(1-2e_0)\pm\sqrt{1-4e_0}}{2e_0}$. The matrix formula (3.6.) also consists of p separate iterative expression of the form

$$e_{j+1} = e_0\left(1 + e_j\right)^2 \tag{3.8.}$$

where e_0, e_j and e_{j+1} scalars and the subscript j is used to denote the j-th iterate. A lemma is given that the iterative procedure defined by (3.8.) converges for $e_0 = 1/4$. An explicit solution for all of the optimum GRE $\hat{\alpha}_i^*,\ i = 1, \dots, p$ is

$$\hat{\alpha}_i^* = \begin{cases} 0 & ,for\ e_{i(0)} > 1/4 \\ \frac{[Z'y]_i}{\lambda_i + \lambda_i e_i^*} = \frac{\hat{\alpha}_i}{(1+e_i^*)} & ,for\ 0 < e_{i(0)} \leq 1/4. \end{cases}$$

Finally it was suggested that the iterative estimation may lead to the introduction of too much bias and an unacceptable loss in R^2. Consequently possible ways of constraining the increase in the residual sum of squares is considered.

Hocking, Speed and Lynn (1976) proposed a general class of biased estimators. They define new vectors $\hat{\theta}_i = (\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_i, 0, \dots, 0)'$ where $\hat{\alpha}$ is the OLS estimator of the canonical model (1.2.). The class of estimators of α is defined by $\tilde{\alpha} = \sum_{i=1}^{p} \gamma_i \hat{\theta}_i$ where γ_i are to be determined. An alternative to $\tilde{\alpha}$ is $\tilde{\alpha} = B\hat{\alpha}$. B is diagonal matrix with diagonal components given by $b_i = \sum_{j=i}^{p} \gamma_j$. Note that the vector $\hat{\theta}_i,\ i = 1, \dots p$ are linearly independent and hence p-space.

Two criteria for determining the γ_i are considered as $E\left(L_1^{\,2}\right) = E(\tilde{\alpha} - \alpha)'(\tilde{\alpha} - \alpha)$ and $E\left(L_2^{\,2}\right) = E(\tilde{\alpha} - \alpha)'\Lambda(\tilde{\alpha} - \alpha)$. Minimization of

$E(L_1^2)$ and $E(L_2^2)$ yields $b_i = \frac{\tau_i^2}{1+\tau_i^2}, i = 1,..,p$ where $\tau_i^2 = \frac{\alpha_i^2 \lambda_i}{\sigma^2}$, $i = 1,..,p$. Comparing this result with the GRE, with $k_i = \frac{\sigma^2}{\alpha_i^2}$, it is found that this unconstrained estimator is identical for GRE with relation $b_i = (1 + k_i/\lambda_i)^{-1}$. All of the biased estimators containing ridge and generalized ridge are seen to be special cases of the general class of estimators. Using either the criteria of $E(L_i^2)$ $i = 1,2$, it is found that the unconstrained case leads to the GRE which is thus, in theory, superior to any of the estimators obtained by constraining the α_i.

The estimation of the coefficients b_i is obtained as

$$b_i^* = \begin{cases} 0 & , \quad \text{if} \quad \hat{\tau}_i^2 < 4 \\ 0.5 + \left[0.25 - \left(1/\hat{\tau}_i^2\right)\right]^{1/2} & , \quad \text{if} \quad \hat{\tau}_i^2 \geq 4 . \end{cases}$$

In the paper of Hemmerle (1975), the possibility of constraining the iteration is considered. His recommendation was that a limit be placed on the total loss in R^2 and this recommendation leads to values for b_i given by $\tilde{b}_i = 1 - \sqrt{m}(1 - b_i^*)$ where b_i^* is the limiting value from the iterative procedure. The objection of using $\tilde{b}_i$ is that it forces all of the $\tilde{b}_i$ to be non-zero. By setting some of the b_i^* to zero, the strong influence of a small eigenvalue on variance inflation was eliminated. Using $\tilde{b}_i$ allows the influence of that eigenvalue to return.

In the paper of Hemmerle and Brantle (1978) the aim was to explore and develop methods for selecting the k_i in general ridge regression. They considered constructing estimators of the two criterion $E(L_1^2)$ and $E(L_2^2)$ and then minimizing these estimators using optimization methods. They presented an unbiased estimator for each criterion, which will be minimized by the same $\hat{\alpha}_i^*$ in each case and admitted an explicit closed form solution for the minimizing $\hat{\alpha}_i^*$ values.

$$E(\hat{\alpha}^* - \alpha)'(\hat{\alpha}^* - \alpha) = \sum_{i=1}^{p} \left(\lambda_i \sigma^2 + k_i^2 \alpha_i^2\right)/(\lambda_i + k_i)^2$$

and

$$E(\hat{\alpha}^* - \hat{\alpha})'(\hat{\alpha}^* - \hat{\alpha}) = \sum_{i=1}^{p} k_i^2 (\sigma^2/\lambda_i + \alpha_i^2)/(\lambda_i + k_i)^2$$

results

$$E(\hat{\alpha}^* - \alpha)'(\hat{\alpha}^* - \alpha) = E(\hat{\alpha}^* - \hat{\alpha})'(\hat{\alpha}^* - \hat{\alpha}) + \sigma^2 \sum_{i=1}^{p} (\lambda_i - k_i)/\lambda_i(\lambda_i + k_i).$$

Consequently

$$\hat{L}_1^2 = (\hat{\alpha}^* - \hat{\alpha})'(\hat{\alpha}^* - \hat{\alpha}) + \hat{\sigma}^2 \mathrm{Tr}\{(\Lambda - K)\Lambda^{-1}(\Lambda + K)^{-1}\} \qquad (3.9.)$$

is an unbiased estimator of $E(L_1^2)$. In a similar manner,

$$E(\hat{\alpha}^* - \alpha)'\Lambda(\hat{\alpha}^* - \alpha) = E(\hat{\alpha}^* - \hat{\alpha})'\Lambda(\hat{\alpha}^* - \hat{\alpha}) + \hat{\sigma}^2 \mathrm{Tr}\{(\Lambda - K)(\Lambda + K)^{-1}\}.$$

So that $\hat{L}_2^2 = (\hat{\alpha}^* - \hat{\alpha})'\Lambda(\hat{\alpha}^* - \hat{\alpha}) + \hat{\sigma}^2 \mathrm{Tr}\{(\Lambda - K)(\Lambda + K)^{-1}\}$ is an unbiased estimator of the criterion $E(L_2^2)$. Let $v_i = \lambda_i/(\lambda_i + k_i)$ and $\hat{\alpha}_i^* = \hat{\alpha}_i v_i$, then the ith component, M_i, of (3.9.) may be written as $M_i = \hat{\alpha}_i^2(v_i - 1)^2 + (\hat{\sigma}^2/\lambda_i)(2v_i - 1)$; where $\hat{L}_1^2 = \sum_{i=1}^{p} M_i$ and $\hat{L}_2^2 = \sum_{i=1}^{p} \lambda_i M_i$. With some further algebra we may also represent M_i as $M_i = \hat{\alpha}_i^2 \left\{(v_i - 1 + e_{i(0)})^2 - e_{i(0)}^2\right\}$ where $e_{i(0)} = \hat{\alpha}^2/\lambda_i \hat{\alpha}_i^2$. The problem resolves to minimization of $\tilde{M}_i = \hat{\alpha}_i^2 \left\{(v_i - 1 + e_{i(0)})^2\right\}$ for $i = 1, \ldots, p$, where the v_i are constrained as $0 \le v_i \le 1$ with $v_i = 0$ corresponding to k_i infinite. The solution to this problem is easily seen to be

$$v_i = \begin{cases} 1 - e_{i(0)}, & e_{i(0)} < 1 \\ 0, & e_{i(0)} \ge 1 \end{cases}$$

which corresponds to

$$\alpha_i = \begin{cases} \hat{\alpha}_i\left(1 - e_{i(0)}\right), & e_{i(0)} < 1 \\ 0 \quad , & e_{i(0)} \geq 1 . \end{cases}$$

Lawless (1981) derived mean squared errors and examined for a number of generalized ridge estimators and the estimators are compared in a number of specific regression models. The potential of generalized ridge estimators is discussed. According to the canonical model (1.2.) RE $\hat{\alpha}_k$ is $\hat{\alpha}_{k(i)} = \left(\frac{\lambda_i}{\lambda_i + k}\right)\hat{\alpha}_i \quad i = 1, .., p$ and more generally $\hat{\alpha}_k$ is $\hat{\alpha}_{k(i)} = \left(\frac{\lambda_i}{\lambda_i + k_i}\right)\hat{\alpha}_i \quad i = 1, .., p$. The estimators considered are all of the same basic form with $\hat{\alpha}_{k(i)} = z(t_i)\hat{\alpha}_i$ where $t_i = \sqrt{\lambda_i}\hat{\alpha}_i/\hat{\sigma}$.

Farebrother (1984) proposed the class of statistics specified by

$$\hat{\beta}(k) = (X'X + kR'R)^{-1}\left(X'X\hat{\beta} + kR'R\hat{\beta}_r\right), \; k \geq 0. \qquad (3.10.)$$

In linear regression model (1.1.) suppose that the parameters almost satisfy the constraints $R\beta = r$ where R and r are $m \times p$ and $m \times 1$ matrices of fixed numbers and R has rank m. If r were random,

$$r = R\beta + \eta \quad E(\eta) = 0 \quad E(\eta\eta') = (\sigma^2/k)I_m \qquad (3.11.)$$

and the best linear unbiased estimator of β from the combination of the models (1.1.) and (3.11.) is

$$\hat{\beta}(k) = (X'X + kR'r)^{-1}(X'y + kR'r).$$

But r and β are fixed so that this derivation of $\hat{\beta}(k)$ and the equivalent bayesian derivation are invalid. However (3.10.) is a weighted average of OLS and restricted least squares (RLS) estimator

$$\hat{\beta}_r = \hat{\beta} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1} \qquad (3.12.)$$

with limits $\hat{\beta}(0) = \hat{\beta}$ and $\hat{\beta}(\infty) = \hat{\beta}_r$ since

$$\hat{\beta}_{(k)} = \hat{\beta} - k(X'X + kR'R)^{-1}R'\left(R\hat{\beta} - r\right)$$
$$= \hat{\beta} - k(X'X)^{-1}R'(I_m + kR(X'X)^{-1}R')^{-1}\left(R\hat{\beta} - r\right)$$

and

$$(X'X + kR'R)(X'X)^{-1}R^{-1} = R'(I_m + kR(X'X)^{-1}R').$$

This estimator (3.10.) may be viewed as a GRE, due to the fact that in the particular case where $m = p$, $R = I_p$ and $r = 0$, it simplifies to RE (2.3.).

Chawla (1988) obtained necessary and sufficient conditions such that general ridge estimator of Hoerl and Kennard (1970) supersede the OLS estimator relative to the MMSE. Using canonical model (1.2.), the MMSE of the OLS estimator $\hat{\alpha}$ is $MMSE(\hat{\alpha}) = \sigma^2\Lambda^{-1}$ and the MMSE of the GRE, $\hat{\alpha}^*$ is

$$\text{MMSE}(\hat{\alpha}^*) = \sigma^2(\Lambda + K)^{-1}\Lambda(\Lambda + K)^{-1} + (\Lambda + K)^{-1}K\alpha\alpha'K(\Lambda + K)^{-1}.$$

A necessary and sufficient condition that GRE excels OLS is given by $\alpha'(2K^{-1} + \Lambda^{-1})^{-1}\alpha < \sigma^2$ with a theorem. Chawla (1988) also defined two corollary. In the first one a sufficient condition that GRE is better than the OLS for all K's is given as $\alpha'\Lambda\alpha \leq \sigma^2$. In the second one a sufficient condition $\alpha'K\alpha \leq 2\sigma^2$ is given that GRE excels the OLS. According to the second theorem a necessary and sufficient condition for the $MMSE(\hat{\alpha}) - \text{MMSE}(\hat{\alpha}^*)$ to be pd is $\mu_i > -1$, $i = 1,..,p$; where μ_i' s are the roots of the equation $|\Lambda^{-1} - (\alpha\alpha'/\sigma^2) - 2\mu K^{-1}| = 0$.

Baksalary, Pordzik and Trenkler (1990) briefly discussed some properties of the estimators $\hat{\beta}(k)$ defined by Farebrother (1984). (3.10) can be written as

$$\hat{\beta}(k) = (X'X + kR'R)^{-1}X'y + k(X'X + kR'R)^{-1}R'r. \qquad (3.13.)$$

Firstly (3.13.) can be represented as $Fy + f$ with $F = (\mathrm{X'X} + \mathrm{kR'R})^{-1}\mathrm{X}$ and $f = k(\mathrm{X'X} + \mathrm{kR'R})^{-1}\mathrm{R'r}$. This representation is useful in verifying that all the statistics form of $\hat{\beta}(\mathrm{k})$ are admissible among the class of linear estimators with respect to the MSE criterion.

Secondly, (3.13.) comprises all admissible general ridge estimators of the form

$$(X'X + H)^{-1}X'y + h \tag{3.14.}$$

where H is any nnd matrix and according to the admissibility condition, h is any vector such that $h = (X'X + H)^{-1}Hg$ for some g. This follows by observing that if $H = L'L$ for some L, then substituting $R = \left(1/\sqrt{k}\right)L$ and $r = \left(1/\sqrt{k}\right)Lg$ into (3.13.) yields (3.14.).

3.2. Modified Ridge Estimator

Ridge regression estimate shrinks the OLS estimate toward the zero vector in the sense that the Euclidean distance between $\hat{\beta}_{\mathrm{k}}$ and 0 is smaller than the distance between $\hat{\beta}$ and 0 for $k > 0$. Swindel (1976) proposed shrinking $\hat{\beta}$ toward b, a nonstochastic p dimensional vector, which should be chosen to represent the prior information on β. For expository purposes, b will be referred to as the prior mean.

$$\hat{\beta}(k, b) = (X'X + kI)^{-1}(X'y + kb),\ k \geq 0 \tag{3.15.}$$

is called as modified ridge estimator (MRE). $\hat{\beta}(k, b)$ can be written as

$$\begin{aligned}\hat{\beta}(k, b) &= (X'X + kI)^{-1}X'X(X'X)^{-1}(X'y + kb) \\ &= (I + k(X'X)^{-1})^{-1}\left(\hat{\beta} - \mathrm{b} + \mathrm{b} + k(X'X)^{-1}b\right) \\ &= (I + k(X'X)^{-1})^{-1}\left(\hat{\beta} - \mathrm{b}\right) + b. \end{aligned} \tag{3.16.}$$

If $\hat{\beta} = \text{b}$, $\hat{\beta}(k,b) = b$. If $\hat{\beta} \neq \text{b}$ then $\hat{\beta}(0,b) = \hat{\beta}$ and $\lim_{k\to+\infty} \hat{\beta}(k,b) = b$.

The Euclidean distance between $\hat{\beta}(k,b)$ and b is monotonically decreasing as k increases.

$\hat{\beta}(k,b)$ is closest to b in an equivalence class of estimators of β with the same residual sum of squares. Clearly, $\hat{\beta}(k,b)$ minimizes $(\tilde{\beta}-b)'(\tilde{\beta}-b)$ such that $(y-X\tilde{\beta})'(y-X\tilde{\beta}) = c$. The objective function is

$$\psi = (\tilde{\beta}-b)'(\tilde{\beta}-b) + \tfrac{1}{k}[(y-X\tilde{\beta})'(y-X\tilde{\beta}) - c]$$

where $\frac{1}{k}$ is a Lagrangian multiplier.

Besides this, $\hat{\beta}(k,b)$ has minimum residual sum of squares in an equivalence class of estimators of β which are equal distance from b. $\hat{\beta}(k,b)$ minimizes $(y-X\tilde{\beta})'(y-X\tilde{\beta})$ such that $(\tilde{\beta}-b)'(\tilde{\beta}-b) = c$, with the objective function

$$\psi = (y-X\tilde{\beta})'(y-X\tilde{\beta}) + k[(\tilde{\beta}-b)'(\tilde{\beta}-b) - c].$$

Swindel and Chapman (1973) showed that

$$E(L'\hat{\beta}_k - L'\beta)^2 < E(L'\hat{\beta} - L'\beta)^2, \forall\, L \neq 0$$

if and only if (iff) k is in the never empty open interval

$$K = \begin{cases} (0,+\infty) & , if\ [(X'X)^{-1} - \sigma^{-2}\beta\beta'] \text{ is nnd} \\ \left(0, \dfrac{-2}{min\ root[(X'X)^{-1} - \sigma^{-2}\beta\beta']}\right), & otherwise\ . \end{cases}$$

Thus, $\hat{\beta}_k$ provides strictly smaller mean square error estimators of every nonnull linear combination of the parameters than does $\hat{\beta}$ iff k is in K.

Following Swindel and Champman (1973) a necessary and sufficient condition is given for MRE to yield estimators with smaller MSE than that of OLS estimator. A necessary and sufficient condition for

$$E\left(L'\hat{\beta}(k,b) - L'\beta\right)^2 < E\left(L'\hat{\beta} - L'\beta\right)^2, \forall\, L \neq 0$$

is that k is in the open interval:

$$K = \begin{cases} (0,+\infty) & ,if\,[(X'X)^{-1} - (b-\beta)(b-\beta)'\sigma^{-2}]\ is\ nnd \\ \left(0, \dfrac{-2}{min\ root[(X'X)^{-1} - (b-\beta)(b-\beta)'\sigma^{-2}]}\right), & otherwise\,. \end{cases}$$

Generalizing the formulation by taking b to be realized value of p-dimensional random variable, B, new formula is

$$\hat{\beta}(k,B) = (X'X + kI)^{-1}(X'y + kB)\,, k \geq 0.$$

$L'\hat{\beta}(k,B)$ is the family of estimators $L'\beta$ where $L \neq 0$ is arbitrary.

For defining $\hat{\beta}(k,B)$ to be a good RE of β based on prior information, B, namely a necessary and sufficient condition for

$MMSE\left(L'\hat{\beta}(k,B)\right) < MMSE\left(L'\hat{\beta}\ \right)\ \forall\, L \neq 0$ is that k is in

$$K = \begin{cases} (0,+\infty) & ,if\,[(X'X)^{-1} - \sigma^{-2}MMSE(B)]\ is\ nnd \\ \left(0, \dfrac{-2}{min\ root[(X'X)^{-1} - \sigma^{-2}MMSE(B)]}\right), & otherwise \end{cases}$$

Swindel (1976) did not compare the mean square error properties of his estimator to that of the RE. Pliskin (1987) examines the set of prior means for which Swindel's estimator has lower risk than RE. A necessary and sufficient condition for $MMSE\left(\hat{\beta}_k\right) - MMSE\left(\hat{\beta}(k,b)\right)$ to be psd when both estimators are computed using the same value of k. A comparison of the MMSE of $\hat{\beta}_k$ and $\hat{\beta}(k,b)$ is difficult

because differences between the two estimators can arise from two sources. These are the point toward which the OLS estimate is shrunk and the degree of shrinkage determined by the value of k, the ridge parameter. The analysis will be limited to the case where the same value of k is used to compute $\hat{\beta}_k$ and $\hat{\beta}(k,b)$ in order to focus on the value of the prior information. In addition, k will be assumed to be nonstochastic to make the problem mathematically tractable.

A prior mean b is said to be good if $MMSE(\hat{\beta}_k) - MMSE\left(\hat{\beta}(k,b)\right)$ is psd for all positive values of k when both $\hat{\beta}_k$ and $\hat{\beta}(k,b)$ are computed using the same value of k. A theorem is given for the goodness of b as:

Theorem 3.1: Let $B = \beta\beta' - (\beta - b)(\beta - b)'$. A prior mean b is good iff B is psd.

Kaçıranlar et all. (1998) compared MMSE of MRE and restricted ridge estimator (RRE) introduced by Sarkar (1992). Following the definition of RE such as $\hat{\beta}_k = \left(I_p + k(X'X)^{-1}\right)^{-1} = W\hat{\beta}$, Sarkar (1992) defined RRE as $\hat{\beta}_r(k) = W\hat{\beta}_r$ where $\hat{\beta}_r$ is the RLS estimator (3.12). In this paper a sufficient condition for $MMSE\left(\hat{\beta}(k,b)\right) - MMSE(\hat{\beta}_r(k))$ to be psd whe both estimators are computed using the same value of k and linear restrictions are indeed true. The conditions for MMSE superiority of this estimator when the restrictions are not true is also derived.

Li and Yang (2010a) showed that MRE is a convex combination of the prior information b and the OLS estimator. Let T_k be:

$$\begin{aligned} T_k &= (X'X + kI)^{-1}X'X \\ &= I - k(X'X + kI)^{-1}. \end{aligned}$$

They define the MRE as

$$\begin{aligned} \hat{\beta}(k,b) &= (X'X + kI)^{-1}(X'y + kb) \\ &= (X'X + kI)^{-1}X'y + k(X'X + kI)^{-1}b \\ &= T_k\hat{\beta} + (I - T_k)b \end{aligned} \qquad (3.17.)$$

Following the Linear combination (3.17.) Li and Yang (2010b) introduced a new restricted MRE (RMRE) as:

$$\hat{\beta}_r(k,b) = T_k\hat{\beta}_r + (I - T_k)b. \qquad (3.18.)$$

RMRE is a convex combination of the prior information b and the RLS estimator. They also made comparisons between RMRE, MRE and RRE of Sarkar (1992). Firstly they showed that RMRE is superior to the MRE in the MMSE sense. Secondly it is shown that RMRE is superior to the RRE in the MMSE sense iff $\beta\beta' - (\beta - b)(\beta - b)' \geq 0$.

3.3. r-k Class Estimator

Baye and Parker (1984) proposed a class of estimators which includes OLS, RR and principal components regression (PCR). It is demonstrated that theoretical gains exist from jointly utilizing the PCR and RR. Meanly, r-k class estimators provides an alternative method of dealing with multicollinearity. They also examined a money demand example, the problem of collinearity of interest rates in estimating money demand, to apply the above techniques to illustrate the relationships between various estimators. Instead of using model misspecification by deleting theoretically relevant interest rates, utilizing the RR technique to reduce MSE of PCR is suggested.

In linear regression model (1.1), define $T = (t_1, \ldots\ldots, t_p)$ to be $p \times p$ matrix which puts $X'X$ in a diagonal form. That is

$$T'X'XT = \Lambda = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{pmatrix}; \quad T'T = TT' = I_p.$$

Let $T_r = (t_1, \ldots\ldots, t_r)$ be the remaining eigenvectors of $(X'X)$ after having deleted $(p - r)$ of the columns of T. Thus, ${T_r}'X'XT_r = \Lambda_{\mathrm{r}}$, where Λ_{r} is defined analogous to

Λ above. Also define $\alpha = T'\beta$. So, r-k class of estimator, which combines the RR and PCR techniques into a single estimator, is defined as:

$$\hat{\beta}(r,k) = T_r(T_r{}'X'XT_r + kI_r)^{-1}T_r{}'X'y \qquad (3.19.)$$

where $r \leq p$ and $k \geq 0$.

Some properties of r-k class estimator is given with this theorem :

Theorem 3.2:

1. $E\left(\hat{\beta}(r,k) - \beta\right) = \left[T_r(\Lambda_r + k\mathrm{I}_r)^{-1}\Lambda_r T_r{}' - \mathrm{I}_p\right]\beta.$

2. $Var\left(\hat{\beta}(r,k)\right) = \sigma^2 T_r(\Lambda_r + k\mathrm{I}_r)^{-1}\Lambda_r(\Lambda_r + k\mathrm{I}_r)^{-1}\mathrm{T}_r{}'.$

3. $\hat{\beta}(r,0) = \hat{\beta}_{PCR}$, $\hat{\beta}_{PCR} = T_r(T_r{}'X'XT_r + kI_r)^{-1}T_r{}'X'y.$

4. $\hat{\beta}(p,k) = \hat{\beta}_k(RE), \hat{\beta}(p,0) = \hat{\beta}(OLS).$

5. $MSE\left(\hat{\beta}(r,k)\right) = \sum_{i=1}^{r}\left[(\delta_i - 1)^2\alpha_i{}^2 + \sigma^2\,\delta_i{}^2/\lambda_i\right] + \sum_{i=r+1}^{p}\alpha_i{}^2$

where $\delta_i = \lambda_i/(\lambda_i + k)$.

6. $Plim\hat{\beta}(r,k) = Plim\hat{\beta}_{PCR}$.

So, for appropriate choices of the r and k $\hat{\beta}(r,k)$ assumes the form of RE, PCR and OLS estimators.

The generalization of the existence theorem proved by Hoerl and Kennard (1970) is presented in a theorem as

Theorem 3.3: Let $\sigma^2 > 0$. Then $\exists k > 0$ such that $MSE\left(\hat{\beta}(r,k)\right) < MSE\left(\hat{\beta}_{PCR}\right)$, $\forall\ 0 < r \leq p$.

Theorem 3.3 generalizes the original ridge results to the case $r < p$. Namely there exist a k bounded away from zero such that a PCR estimator may be shrunk to yield an estimator with lower MSE. Also in the RR, the optimal k depends upon the unknown parameters α and σ^2.

Baye and Parker (1984) only compared the r-k class estimator and the PCR estimator, they didn't compare the r-k class estimator with RE and OLS estimator. Nomura and Ohkuba (1985) deal with this comparisons in terms of MSE criterion.

Let $N_r = \{1, \dots, r\}$ and $\overline{N_r} = \{r+1, \dots, p\}$. The MSE is given by $MSE\big(\hat{\beta}(r,k)\big) = E\big(\hat{\beta}(r,k) - \beta\big)' W\big(\hat{\beta}(r,k) - \beta\big)$. When $W = I$ the $MSE\big(\hat{\beta}(r,k)\big)$ represents the MSE of the estimator of the regression coefficients and when $W = X'X$, the $MSE\big(\hat{\beta}(r,k)\big)$ represents the MSE of the predictor of $E[y|X]$. From the definition of W, ${T_r}'WT_r = diag(\mu_1, \dots, \mu_r)$; $\mu_i > 0$ for all $i \in N_r$ where $(\mu_1, \dots, \mu_r) = (1, \dots\dots, 1)$ for $W = I$ and $(\mu_1, \dots, \mu_r) = (\lambda_1, \dots, \lambda_r)$ for $W = X'X$.

According to

$$MSE\big(\hat{\beta}(r,k)\big) = \sum_{i \in N_r} \mu_i \big[(\delta_i - 1)^2 {\alpha_i}^2 + \sigma^2 {\delta_i}^2 / \lambda_i\big] + \sum_{i \in N_r} \mu_i {\alpha_i}^2$$

where $\delta_i = \lambda_i/(\lambda_i + k)$ the MSE of $\hat{\beta}(r,k)$ and with that of RE is compared with a theorem as:

Theorem 3.4: Let $r \in N_1$. Thus,

1. ${\alpha_r}^2 \lambda_r < \sigma^2$ iff

$MSE\big(\hat{\beta}(r,k)\big) > MSE\big(\hat{\beta}(r-1,k)\big)$ for $0 \le k < (\sigma^2 - {\alpha_r}^2 \lambda_r)/2{\alpha_r}^2$ and

$MSE\big(\hat{\beta}(r,k)\big) < MSE\big(\hat{\beta}(r-1,k)\big)$ for $(\sigma^2 - {\alpha_r}^2 \lambda_r)/2{\alpha_r}^2 \le k$.

2. ${\alpha_r}^2 \lambda_r \ge \sigma^2$ iff

$MSE\big(\hat{\beta}(r,k)\big) \le MSE\big(\hat{\beta}(r-1,k)\big)$ for $0 \le k < \infty$.

So, the following corollary is obtained by using Theorem 3.4.

Corollary 3.1: For $r \in N_{p-1}$,

1. If ${\alpha_i}^2 \lambda_i < \sigma^2$ for all $i \in \overline{N_r}$, then there exists a strictly positive K_1 such that

$MSE\big(\hat{\beta}_k\big) > MSE\big(\hat{\beta}(p-1,k)\big) > \cdots > MSE\big(\hat{\beta}(r,k)\big)$ for $0 < k < K_1$.

2. If ${\alpha_i}^2 \lambda_i \le \sigma^2$ for some $i \in \overline{N_r}$, then there exists a nonnegative K_2 such that

$MSE\big(\hat{\beta}_k\big) < MSE\big(\hat{\beta}(p-1,k)\big) < \cdots < MSE\big(\hat{\beta}(r,k)\big)$ for $K_2 < k$

Following the Corollary 3.1 the MSE's of the $\hat{\beta}(r,k)$ and the OLS can be compared by the theorem:

Theorem 3.5: Let $r \in N_p$. Then,

1. If $\sum_{i \in N_r} \mu_i ({\alpha_i}^2 - \sigma^2/\lambda_i) \le 0$ and $\sigma^2/\lambda_i - {\alpha_i}^2 \ge 0$ for all $i \in N_r$ then

$MSE\big(\hat{\beta}(r,k)\big) < MSE\big(\hat{\beta}\big)$ for $0 < k < \infty$.

2. If $\sum_{i\in N_r}\mu_i({\alpha_i}^2-\sigma^2/\lambda_i)\leq 0$ and $\sigma^2/\lambda_i-{\alpha_i}^2<0$ for some $i\in N_r$ then $MSE\left(\hat{\beta}(r,k)\right)<MSE\left(\hat{\beta}\right)$ for $0<k<2\sigma^2/\max_{i\in N_r}({\alpha_i}^2-\sigma^2/\lambda_i)$.

Baye and Parker (1984), Nomura and Ohkuba (1985), Sarkar (1989) compared the performance of r-k class estimator with RE, PCR and OLS estimator by the criterion of MSE. While the MSE is often used to gauge the performance of an estimator, the MMSE is also widely accepted and stronger criterion. Sarkar (1996) derived the conditions under which the r-k class estimator is superior to the RE, PCR and OLS estimators by the criterion of MMSE.

Firstly consider comparing r-k class estimator with the OLS.

$$MMSE\left(\hat{\beta}\right)-MMSE\left(\hat{\beta}(r,k)\right)=\ T\left(S^*(k)\right)^{-1}[\sigma^2\Lambda^*(k)^{-1}-T'\beta\beta'T]\left(S^*(k)\right)^{-1}T'$$

where $S^*(k)=\begin{pmatrix}\frac{1}{k}(\Lambda_r+k\mathrm{I}_r) & 0\\ 0 & I_{p-r}\end{pmatrix}$ and $\Lambda^*(k)^{-1}=\begin{pmatrix}\frac{2}{k}\mathrm{I}_r+{\Lambda_r}^{-1} & 0\\ 0 & {\Lambda_{p-r}}^{-1}\end{pmatrix}$

$MMSE\left(\hat{\beta}\right)-MMSE\left(\hat{\beta}(r,k)\right)$ is nnd matrix iff $\sigma^2\Lambda^*(k)^{-1}-T'\beta\beta'T$ is so $\sigma^2\Lambda^*(k)^{-1}-T'\beta\beta'T$ is nnd iff $\beta'T\Lambda^*(k)T'\beta\leq\sigma^2$ (Rao,1974). By substituting the expression for $\Lambda^*(k)$ the following theorem is defined as:

Theorem 3.6: A necessary and sufficient condition for the r-k class estimator of β to the superior to the OLS estimator by MMSE is

$$\left[\beta'T_r\left(\frac{2}{k}\mathrm{I}_r+{\Lambda_r}^{-1}\right)^{-1}{T_r}'\beta+\beta'T_{p-r}\Lambda_{p-r}{T_{p-r}}'\beta\right]/\sigma^2\leq 1.$$

Secondly to compare r-k class estimator and PCR estimator, differences between the MMSE of estimators are:

$$\begin{aligned}MMSE\left(\hat{\beta}_{PCR}\right)-MMSE\left(\hat{\beta}(r,k)\right)=\ &\sigma^2T_r\left({\Lambda_r}^{-1}-S_r(k)^{-1}\Lambda_rS_r(k)^{-1}\right){T_r}'\\ &+\left(T_r{T_r}'-I_p\right)\beta\beta'\left(T_r{T_r}'-I_p\right)\\ &-\left(T_rS_r(k)^{-1}\Lambda_r{T_r}'-I_p\right)\beta\beta'\left(T_r\Lambda_rS_r(k)^{-1}{T_r}'-I_p\right)\end{aligned}$$

where $S_r(k)=\Lambda_r+k\mathrm{I}_r$. Note that to do this comparison

$$T_r\left(\left({\Lambda_r}^{-1}-S_r(k)^{-1}\right)\Lambda_r S_r(k)^{-1}\right){T_r}' = k^2 T_r S_r(k)^{-1}\left(\frac{2}{k}\mathrm{I}_r+{\Lambda_r}^{-1}\right)S_r(k)^{-1}{T_r}'$$

is a nnd matrix for all $k>0$ and hence obtaining a necessary and sufficient condition for the dominance of r-k class estimator over the PCR estimator becomes difficult. Following the result by Baksalary and Trenkler (1991) the condition simplifies to:

$$\beta' Tr\left(\frac{2}{k}I_r+\Lambda_r^{-1}\right)^{-1}Tr'\beta=0$$

which hold iff $Tr'\beta=0$.

Theorem 3.7: The r-k class estimator of β dominates the PCR estimator by the MMSE criterion iff $Tr'\beta=0$.

To compare the r-k class estimator and RE the MMSE differences are

$$\begin{aligned}MMSE\left(\hat{\beta}_k\right)-MMSE\left(\hat{\beta}(r,k)\right)=\sigma^2 T_{p-r}S_{p-r}(k)^{-1}\Lambda_{p-r}S_{p-r}(k)^{-1}T'_{p-r}\\ +\left(TS(k)^{-1}\Lambda T'-I_p\right)\beta\beta'\left(T\Lambda S(k)^{-1}T'-I_p\right)\\ -(T_rS_r(k)^{-1}\Lambda_r T_r'-I_p)\beta\beta'(T_r)\Lambda_r S_r(k)^{-1}T_r'-I_p\end{aligned}$$

where $S_{p-r}(k)^{-1}$ is the inverse of the matrix $S_{p-r}(k)=\Lambda_{p-r}+kI_{p-r}$.

As in the previous case, Baksalary and Trenkler (1991) is used to obtain the condition. The condition finally reduces to $\beta'\mathrm{T}_{p-r}\Lambda_{p-r}T'_{p-r}\beta=0$, which holds iff $T'_{p-r}\beta=0$. This result is stated in the following theorem.

Theorem 3.8: A necessary and sufficient condition for the r-k class estimator of β to dominate the RE by MMSE criterion is given by $T'_{p-r}\beta=0$.

3.4. Almost Unbiased Ridge Estimator

The Jack-knife procedure is usually applied where an unbiased estimator is not available but there is a reasonable estimator which is biased. An unbiased estimator, namely, the OLS is available but may not be computationally feasible

because of multicollinearity. The RE offers a great computational advantage. So, Singh, Chaubey and Dwivedi (1986) used the Jack-knife procedure to reduce the bias of the RE. The resulting estimator is similar in form as that of the RE, hence both have the same asymptotic properties.

Considering the canonical model (1.2.) the GRE demonstrated with (3.1.) can be written as:

$$\hat{\alpha}^* = (\Lambda + \mathrm{K})^{-1} Z'y = \mathrm{A}^{-1} Z'y. \tag{3.20.}$$

The relationship between $\hat{\alpha}^*$ and $\hat{\beta}$ is;

$$\hat{\beta} = T\hat{\alpha}^* = {A_*}^{-1} X'y$$

where $A_* = X'X + K_*$ and $K_* = TKT'$. Further since $\hat{\alpha} = \Lambda^{-1} Z'y$ (3.20.) may be written as:

$$\hat{\alpha}^* = A^{-1}\Lambda\hat{\alpha} = (I - A^{-1}K)\hat{\alpha}. \tag{3.21.}$$

It is easy to see that

$$Bias(\hat{\alpha}^*) = -A^{-1}K\hat{\alpha}.$$

which is order $(1/n)$ under the condition that $(X'X)^{-1}$ and hence $(Z'Z)^{-1}$ is of order $(1/n)$. $Bias(\hat{\beta}) = TBias(\hat{\alpha}^*) = -TA^{-1}K\alpha = -TA^{-1}KT'\beta$ which is also of order $1/n$.

Hinkley (1977) proposed a Jack-knife form of $\hat{\alpha}^*$ and hence that of $\hat{\beta}$. Let y_{-i} denote the vector y with its i-th coordinate deleted and Z_{-i} denote the matrix Z with its i-th column deleted. $\hat{\alpha}^*_{-i}$ is obtained replacing Z and y with Z_{-i} and y_{-i} in (3.20.). Thus,

$$\hat{\alpha}^*_{-i} = (Z_{-i}{}' Z_{-i} + \mathrm{K})^{-1} Z_{-i}{}' y_{-i} .$$

Let z_i be the i-th column vector of Z and y_i the i-th coordinate of y. Then

$$\hat{\alpha}^*_{-i} = (Z'Z - z_i z_i{}' + K)^{-1}(Z'y - z_i y_i).$$

The above can be further simplified as:

$$\hat{\alpha}^*_{-i} = \hat{\alpha}^* - \frac{A^{-1} z_i u_i}{1 - w_i}$$

where $u_i = (y_i - z_i{}'\hat{\alpha}^*)$ and $w_i = z_i(Z'Z')^{-1} z_i$.

Following Miller (1974) the pseudo-values are defined as:

$$P_i = n\hat{\alpha}^* - (n-1)\hat{\alpha}^*_{-i} \qquad (3.22.)$$

from which the Jack-knife estimator is obtained as:

$$\bar{P} = \frac{1}{n}\sum_{i=1}^{n} P_i = \hat{\alpha}^* + \frac{n-1}{n} A^{-1} \sum_{i=1}^{n} \frac{z_i u_i}{1 - w_i}.$$

This is in line with the Jack-knife method originally developed by Quenouille (1956) and Tukey (1958). But as noted by Hinkley, the pseudo-values in (3.22.) are defined symmetrically with respect to the observations, whereas the model is generally unbalanced. The lack of balance is reflected in the "distances" w_i. Further, since variance of $(\hat{\alpha}^* - \hat{\alpha}^*_{-i})$ is an increasing function of w_i, the pseudo-values are defined as:

$$Q_i = \hat{\alpha}^* + n(1 - w_i)(\hat{\alpha}^* - \hat{\alpha}^*_{-i})$$

and the corresponding Jack-knifed estimator is given by

$$\tilde{\alpha}^* = \bar{Q} = \frac{1}{n}\sum_{i=1}^{n} Q_i = \hat{\alpha}^* + A^{-1}\sum_{i=1}^{n} z_i u_i$$

which can be simplified to yield

$$\tilde{\alpha}^* = (I + A^{-1}K)\hat{\alpha}^* = [I - (A^{-1}K)^2]\hat{\alpha}.$$

It is clear that

$$Bias(\tilde{\alpha}^*) = -(A^{-1}K)^2\alpha$$

is of order $(1/n^2)$. Reduction in bias is clear when $|Bias(\hat{\alpha}^*)|_i$ is compared with $|Bias(\tilde{\alpha}^*)|_i$ where $|.|_i$ denotes the absolute value of the i-th component. It is seen that

$$|Bias(\hat{\alpha}^*)|_i - |Bias(\tilde{\alpha}^*)|_i = \frac{\lambda_i k_i}{(\lambda_i + k_i)^2}|\alpha_i|$$

is positive. This type of bias reduction may not hold between each component of $|Bias(\hat{\beta})|$ and $|Bias(\tilde{\beta})|$ even though $Bias(\tilde{\beta})$ may be of smaller order in magnitude than $Bias(\hat{\beta})$. However,

$$\Delta = \sum\{|Bias(\hat{\beta})|_i - |Bias(\tilde{\beta})|_i\}^2$$

is always positive. This can be easily seen because,

$$\Delta = \beta'\{T[(A^{-1}K)^2 - (A^{-1}K)^4]T'\}\beta.$$

Since the matrix in braces is pd $\Delta > 0$.

Nomura (1988) compared the almost unbiased generalized ridge (AUGR) estimator proposed by Singh, Chaubey and Dwivedi (1986) with the GRE and the OLS estimator in terms of the MSE criterion. Under the assumption that the ridge parameter is nonstochastic.

The MSE of $\hat{\alpha}_i$, the i-th component of the OLS estimator of α is:

$$\text{MSE}(\hat{\alpha}_i) = E(\hat{\alpha}_i - \alpha_i)^2 = \sigma^2/\lambda_i \quad i = 1,..,p. \tag{3.23.}$$

From (3.23.) the MSE of $\hat{\alpha}$ is given by

$$MSE(\hat{\alpha}) = E(\hat{\alpha} - \alpha)'(\hat{\alpha} - \alpha) = \sum_{i=1}^{p} \sigma^2/\lambda_i$$

and the predictive MSE (PMSE) of the $\hat{\alpha}$ is given by

$$PMSE(\hat{\alpha}) = E(\hat{\alpha} - \alpha)'X'X(\hat{\alpha} - \alpha) = p\sigma^2.$$

Similarly the MSE of $\hat{\alpha}_i^*$ from (3.21.) is

$$MSE(\hat{\alpha}_i^*) = \left((\hat{\alpha}_i^*)^2 k_i^2 + \sigma^2\lambda_i\right)/(\lambda_i + k_i)^2, \quad i = 1,..,p.$$

Furthermore by using the following Lemma the MSE of AUGR estimator with that of the RE is compared by a theorem

Lemma 3.1: $MSE(\tilde{\alpha}_i^*)$ is given by

$$\begin{aligned} MSE(\tilde{\alpha}_i^*) &= \sigma^2\lambda_i\,(2k_i + \lambda_i)^2/(\lambda_i + k_i)^4 + k_i^4\alpha_i^4/(\lambda_i + k_i)^4 \\ &= \left(\alpha_i^2 k_i^4 + 4\sigma^2\lambda_i k_i^2 + 4\sigma^2\lambda_i^2 k_i + \sigma^2\lambda_i^3\right)/(\lambda_i + k_i)^4. \end{aligned}$$

Theorem 3.9:

1. $MSE(\tilde{\alpha}_i^*) > MSE(\hat{\alpha}_i^*)$ for $0 < k_i < K1_i$

2. $MSE(\tilde{\alpha}_i^*) < MSE(\hat{\alpha}_i^*)$ for $K1_i < k_i < \infty$ where

$K1_i = \left[3\sigma^2 - \lambda_i\alpha_i^2 + [(3\sigma^2 + \lambda_i\alpha_i^2)^2 + 4\sigma^2\lambda_i\alpha_i^2]^{1/2}\right]/4\alpha_i^2 > 0$

Secondly the MSE of AUGR estimator with that of the OLS estimator is compared by a theorem.

Theorem 3.10:

1. If $\lambda_i {\alpha_i}^2 - \sigma^2 \leq 0$ then

$$MSE(\tilde{\alpha}_i^*) < MSE(\hat{\alpha}_i^*) \text{ for } 0 < k_i < \infty,$$

2. If $\lambda_i {\alpha_i}^2 - \sigma^2 > 0$ then there exists a strictly positive $K2_i$ such that

$$MSE(\tilde{\alpha}_i^*) < MSE(\hat{\alpha}_i^*) \text{ for } 0 < k_i < K2_i$$

and

$$MSE(\tilde{\alpha}_i^*) > MSE(\hat{\alpha}_i^*) \text{ for } K2_i < k_i < \infty,$$

where

$$K2_i = \left[2\sigma^2\lambda_i + \left(2\sigma^4{\lambda_i}^2 + 2\sigma^2{\lambda_i}^3{\alpha_i}^2\right)^{1/2}\right]/(\lambda_i{\alpha_i}^2 - \sigma^2) > 0.$$

3.5. Unbiased Ridge Estimator

Crouse and Hanumara (1995) defined unbiased ridge estimator (URE) with prior information and proposed robust estimate of the ridge parameter k.

Swindel (1976) defined MRE (3.15.) and showed the existence of k value that results in a smaller MSE than the OLS estimator for any fixed prior information b. Such on existence of a k value is also noted when b is a random vector independent of $\hat{\beta}$. Pliskin (1987) compared the MRE and RE using the same value of k. A robust user-oriented non-iterative method for estimating k which corporates a random vector of empirical prior information b will be developed by Crouse and Hanumara (1995). The proposed ridge estimation provides unbiased estimates of β.

Theorem 3.11: Consider the standard linear regression model (1.1.), the OLS, $\hat{\beta}$, is normally distributed $N(\beta, \sigma^2(X'X)^{-1})$. The prior information b is independent of $\hat{\beta}$ and b is normally distributed $N(\beta, V)$. Also assume that V is a full rank covariance matrix and the convex estimator is $\hat{\beta}(C, b) = C\hat{\beta} + (I - C)b$, where C is $p \times p$

matrix but otherwise is unrestricted. Then the optimal C in terms of minimum MSE is $C = V(\sigma^2(X'X)^{-1} + V)^{-1}$.

Corollary 3.2: Suppose that $\hat{\beta}$ is an estimator of β and covariance matrix Σ and b is prior information with mean β and covariance matrix V. Further, assume b is uncorrelated with $\hat{\beta}$ and V and Σ are of full rank. Then the convex estimator $\beta(C, b)$ has minimum MSE for optimal value of $C = V(V + \Sigma)^{-1}$.

Theorem 3.12: Let $\hat{\beta}$ have a distribution with mean β and covariance $\sigma^2(X'X)^{-1}$, denoted by $M(\beta, \sigma^2(X'X)^{-1})^{-1}$, as in the linear model. Similarly, let b be distributed $M(\beta, (\sigma^2/k)I)$ for $k > 0$ and define $\beta(C, b) = C\hat{\beta} + (I - C)b$. Then for the optimal value of C in terms of minimum MSE, $\beta(C, b) = \hat{\beta}(k, b) = (X'X + kI)^{-1}(X'y + kb)$ and $\beta(C, b)$ is an unbiased estimate of β.

The optimal convex estimator stated in Theorem 3.11 can be seen as a Swindel type RE. However empirical prior information in (3.15.) is a fixed vector while b in Theorem 3.11 is a random vector with specified mean and covariance. In fact, the optimal convex estimator $\hat{\beta}(k, b)$ in Theorem 3.11 is not an estimator in the usual sense because it depends on an unknown ridge parameter k that needs to be estimated.

Swindel (1976) didn't proposed a method to estimate k. Crouse and Hanumara (1995) considered the problem of estimating the ridge parameter k. They propose a procedure to estimate $1/k$ rather than k as follows.

It is assumed that empirical prior information b as in Theorem 3.11 is uncorrelated with the OLS $\hat{\beta}$. Note that $\hat{\beta} - b$ has a distribution with mean 0 and covariance $(\sigma^2/k)I + \sigma^2(X'X)^{-1}$, denoted by

$$\mu(0, (\sigma^2/k)I + \sigma^2(X'X)^{-1}) \tag{3.24.}$$

and $E\left[\left(\hat{\beta} - J\right)'\left(\hat{\beta} - J\right)\right] = (p\,\sigma^2/k) + \sigma^2 T_r(X'X)^{-1}$. An unbiased estimate of $1/k$ is

$$\frac{1}{k} = \frac{(\hat{\beta}-b)'(\hat{\beta}-b) - \sigma^2 T_r(X'X)^{-1}}{p\sigma^2}. \tag{3.25.}$$

If σ^2 is known. The more common situation is when σ^2 is unknown. Then σ^2 can be estimated by an unbiased estimator $s^2 = \left(y - x\hat{\beta}\right)'\left(y - x\hat{\beta}\right)/(n-p)$ and an estimate of $1/k$ is

$$\frac{1}{k} = \frac{(\hat{\beta}-b)'(\hat{\beta}-b)-s^2 T_r(X'X)^{-1}}{ps^2} . \qquad (3.26.)$$

There are other estimates of σ^2 which may be used as well. The proposed estimate of $1/k$ in (3.25.) and (3.26.) provides an URE of β under certain conditions on the distribution $\mu(.,.)$ while all of other available ridge estimators are biased. The unbiased estimation results are summarized as follows.

Theorem 3.13: Given that $\hat{\beta} - b$ is distributed as (3.24.) and the distribution $\mu(.,.)$ is symmetric about 0. If σ^2 is known, then the proposed RE, $\hat{\beta}\left(\hat{k}, b\right)$, is unbiased where $\hat{k}$ is given by (3.26.).

Theorem 3.14: Given that $\hat{\beta} - b$ is distributed as $N(0, (\sigma^2/k)I + \sigma^2(X'X)^{-1})$. If σ^2 is unknown, then the proposed RE $\hat{\beta}\left(\hat{k}, b\right)$ is unbiased where $\hat{k}$ is given by (3.26.).

3.6. Restricted Ridge Estimator

The problem of multicollinearity and its statistical consequences on OLS method, such as producing large sampling variances, leads to two estimation procedures the restricted least squares (RLS) method and the method of ridge regression. In the paper of Sarkar (1992) a new estimator by combining the two approaches followed by the RE and RLS estimator is suggested. When the restrictions on the parameters are true, often a reparameterization in terms of linear combinations of the original parameters is obtained. If multicollinearity still remains a problem with reparameterized model, it becomes necessary to modify the usual RLS estimator in the line of RE. Namely a new estimator is suggested by grafting ridge regression philosophy into the RLS estimator.

Consider the standard linear regression model (1.1.). The set of a prior restrictions on the parameters are represented by

$$R\beta = r \tag{3.27.}$$

where r is a $q \times 1$ vector of known elements and R is a $q \times p$ known prior information design matrix that expresses the structure of information on the individual parameters or some linear combinations of these parameters. Both of the case that restrictions are true or not true is considered. The RLS estimator of β is obtained by minimizing $(y - X\beta)'(y - X\beta)$ subject to the restrictions given in (3.27.) can be written as:

$$\hat{\beta}_r = \hat{\beta} + S^{-1}R'(RS^{-1}R')^{-1}\left(r - R\hat{\beta}\right)$$

where $S = X'X$. $E\left(\hat{\beta}_r\right) \neq \beta$ unless (3.27.) holds. The variance covariance matrix of $\hat{\beta}_r$ is $Var\left(\hat{\beta}_r\right) = \sigma^2(S^{-1} - S^{-1}R'(RS^{-1}R')^{-1}RS^{-1})$ and $Var\left(\hat{\beta}\right) - Var\left(\hat{\beta}_r\right)$ is psd matrix and hence it is concluded that the RLS estimator has a smaller sampling variance than the OLS estimator. Further,

$$MSE\left(\hat{\beta}_r\right) = \sigma^2 Tr(A) + \delta'(RS^{-1}R')^{-1}RS^{-2}R'(RS^{-1}R')^{-1}\delta$$

where $A = S^{-1} - S^{-1}R'(RS^{-1}R')^{-1}RS^{-1}$ and $\delta = r - R\beta$.

Since $\hat{\beta}_k = W\hat{\beta}$, $W = (I + kS^{-1})^{-1}$, Sarkar (1992) defines a restricted ridge regression estimator (RRE) as

$$\hat{\beta}_r(k) = W\hat{\beta}_r \text{ , } k \geq 0. \tag{3.28.}$$

Obviously $\hat{\beta}_r(k) = \hat{\beta}_r$ when $k = 0$. The proposed RRE can be obtained by minimizing $\beta'\beta$ subject to a fixed value φ for $\left(\beta - \hat{\beta}_r\right)'X'X\left(\beta - \hat{\beta}_r\right)$.

$E\left(\hat{\beta}_r(k)\right) = W\beta + WS^{-1}R'(RS^{-1}R')^{-1}\delta$. Thus, the RRE is always a biased estimator unless $k = 0$ and $\delta = 0$. Further $Var\left(\hat{\beta}_r(k)\right) = \sigma^2 WAW'$ and the MSE of $\hat{\beta}_r(k)$ is

$$MSE\left(\hat{\beta}_r(k)\right) = \sigma^2 Tr(WAW') + [WS^{-1}\delta^* - kS(k)^{-1}\beta]'[WS^{-1}\delta^* - kS(k)^{-1}\beta].$$

where $\delta^* = R'(RS^{-1}R')^{-1}\delta$. If $\delta = 0$, namely the restrictions are true, then MSE reduces to

$$MSE\left(\hat{\beta}_r(k)\right) = \sigma^2 Tr(WAW') + k^2\beta' S(k)^{-2}\beta.$$

Sarkar (1992) also studied the properties of this estimator and MSE superiority of the proposed estimator over both the RE and RLS estimator when parametric restrictions are indeed true or not true.

Gross (2003) introduced a new estimator which is a generalization of the restricted least squares estimator and is confined to the (affine) subspace which is generated by the restrictions. Gross (2003) claimed that $\hat{\beta}_r(k)$ does not satisfy $R\hat{\beta}_r(k) = r$ for every outcome of y and therefore is not a restricted estimator with respect to $R\beta = r$. Since Sarkar (1992) didn't assume that $R\beta = r$ must always be true when $\hat{\beta}_r$ is used, it is reasonable to allow the estimator $\hat{\beta}_r(k)$ to take values outside the (affine) subspace which is generated by the equation $R\beta = r$. In the paper of Gross (2003) a new ridge-like estimator for β, which is a restricted estimator with respect to a given equation $R\beta = r$.

Consider that there is no prior knowledge about an appropriate vector b in MRE, $\hat{\beta}(k, b)$. Then shrinking towards $b = 0$ results in RE. If linear restrictions $R\beta = r$ are assumed to hold, then it is no longer appropriate to shrink towards the zero vector but towards the shortest vector β_0 which satisfies the restrictions. This vector is given by $\beta_0 = R'(RR')^{-1}r$, but the resulting estimator

$$\hat{\beta}(k,\beta_0) = \left(X'X + kI_p\right)^{-1}(X'y + kR'(RR')^{-1}r)\ , k \geq 0 \qquad (3.29)$$

is of course not a restricted estimator with respect to $R\beta = r$. Gross (2003) proposed

$$\hat{\beta}_r{}^*(k) = \hat{\beta}(k,\beta_0) - S_k^{-1}R'[RS_k^{-1}R']^{-1}\left(R\hat{\beta}(k,\beta_0) - r\right), k \geq 0$$

with $S_k = X'X + kI_p$ that this estimator is restricted with respect to $R\beta = r$.
It is also clear that for $k = 0$, $\hat{\beta}_r{}^*(0) = \hat{\beta}_r$, but it is not immediately seen whether this estimator moves towards $\beta_0 = R'(RR')^{-1}r$ when k becomes larger.

Gross (2003) indicates a theorem that $\lim_{k\to\infty}\hat{\beta}_r{}^*(k) = \beta_0$, where $\beta_0 = R'(RR')^{-1}r$. Therefore this new estimator $\hat{\beta}_r{}^*(k)$ satisfies $R\hat{\beta}_r{}^*(k) = r$ and can be regarded as a movement from the RLS estimator $\hat{\beta}_r$ into the direction of β_0, which is the shortest vector satisfying the restrictions. Gross (2003) also gave necessary and sufficient conditions for the superiority of the now estimator over the RLS estimator.

4. RIDGE ESTIMATORS FROM DIFFERENT POINTS OF VIEW

4.1. . Ridge Estimators in Seemingly Unrelated Regression Equations Model

A multiple linear regression model is

$$y = XB + \varepsilon \tag{4.1.}$$

where y is an $(n \times q)$ matrix of q responses, X is an $(n \times p)$ matrix of regressors, whose elements are treated as fixed constants, $B = (\beta_1, .., \beta_q)$ is a $(p \times q)$ matrix of unknown coefficients and $\varepsilon = (e_1, .., e_q)$ is an $(n \times q)$ error matrix, such that $E(e_j) = 0,\ E(e_j e_l{}') = \sigma_{jl}^2 I_n\ ,\ j, l = 1, \dots q.$

Within each equation the disturbance vector is assumed to be homoscedastic, but the disturbances are permitted to be correlated across equations.

The system of linear regression equations (4.1.) can be rewritten in a form

$$y = Z\beta + e \tag{4.2.}$$

where y is stringed out into a $(qn \times 1)$ vector y, and the error matrix ε is stringed out into a $(qn \times 1)$ vector e. The unknown parameters β form a $(pq \times 1)$ vector and Z is a $q(n \times p)$ block matrix with the original X_j matrices, $j = 1, \dots q$, on the diagonal, $Z = I_q \otimes X_j$, where $\otimes$ denotes Kronecker product of matrices. The disturbance vector e has zero mean and the dispersion matrix

$$E(ee') = \sum \otimes I_n = \Gamma$$

where $\sum = \{\sigma_{ij}^2\}$ is a $(q \times q)$ pd symmetric matrix. In this form, the regression model (4.2.) is called seemingly unrelated regression equations (SURE) model. Even though the equations in model (4.1.) appear to be structurally unrelated (the dependent variable in one equation does not appear as a regressor in other equation),

the fact that the disturbances are correlated across equations constitutes a link among them.

Such a system was first considered by Zellner (1962) and clearly the generalized least squares (GLS) estimator is

$$\hat{\beta}_{GLS} = (Z'\Gamma^{-1}Z)^{-1}Z'\Gamma^{-1}y$$

and it is minimum variance unbiased with

$$Var(\hat{\beta}_{GLS}) = (Z'\Gamma^{-1}Z)^{-1}.$$

However there are numbers of conditions under which the OLS estimator of β is fully efficient and there is no gain in efficiency from the use of the GLS estimator. Brinkley (1982) and Brinkley and Nelson (1988) discussed the problem of collinearity in the context of SURE. Considering a two equation system of SURE, Brinkley (1982) concludes that correlation between a variable in one equation and those in the other will adversely affect the efficiency of $\hat{\beta}_{GLS}$ if this correlation is not already reflected by existing multicollinearity between the variable and the remaining ones in its own equation. However Brinkley and Nelson (1988) point out that even if there is correlation among variables across equations, gains in efficiency can still be made through the use of $\hat{\beta}_{GLS}$ when there is multicollinearity among the variables within an equation. Clearly, the discussion of multicollinearity in a system of SURE is much more complex than in classical linear regression model and multicollinearity can produce $\hat{\beta}_{GLS}$ estimates with larger variance than expected.

Since multicollinearity can have damaging effects on the GLS estimator and under certain conditions the estimator will not produce any gains in efficiency over OLS, it is appropriate to consider biased estimators as means of reducing the variance of parameter estimates. One such possibility is provided by the ridge regression. The ordinary RE of β in the SURE model (4.2.) is defined by Srivastava and Giles (1987) as

$$\hat{\beta}_{RE} = (Z'\Gamma^{-1}Z + kI)^{-1}Z'\Gamma^{-1}y.$$

This estimator is obtained as a solution to a minimization of $\beta'\beta$ for a given level of weighted squared errors $(y - Z\beta)'\Gamma^{-1}(y - Z\beta)$ or as a solution to a constrained regression problem. Considering first the minimization approach, on the basis of prior information, more weight can be given to some variables than others. Having a prior information on β in the form of a nonstochastic vector b might be preferable to minimize the distance to a given point, other than origin. So instead of minimizing $\beta'\beta$, the squared weighted distance $(\beta - b)'\Theta(\beta - b)$ is minimized subject to the constant $\lambda = (y - Z\beta)'\Gamma^{-1}(y - Z\beta)$. The weights given to distance $(\beta - b)$ are the diagonal elements of Θ^{-1}. Differentiating a Lagrangian expression with multiplier $1/k$ and with respect to β leads to the modified ridge estimator in the SURE model which is defined by Füle (1995) as

$$\hat{\beta}_{MR} = (Z'\Gamma^{-1}Z + k\Theta^{-1})^{-1}(Z'\Gamma^{-1}y + k\Theta^{-1}b). \quad (4.3.)$$

In the second approach, both sources of information the basic model (4.2.) and the prior information given in the form of stochastic restrictions $= \beta + v$, with stochastic component v, $E(v) = 0$, $E(vv') = k^{-1}\Theta$ and a constant $k > 0$, are combined into a model

$$y^* = \begin{bmatrix} y \\ b \end{bmatrix} = \begin{bmatrix} Z \\ I \end{bmatrix}\beta + \begin{bmatrix} e \\ v \end{bmatrix} = Z^*\beta + e^*. \quad (4.4.)$$

It is assumed that e and v are independent, and a transformation matrix P is used such that

$$PP' = \Omega^{-1} = \begin{bmatrix} \Gamma & 0 \\ 0 & k^{-1}\Theta \end{bmatrix}^{-1}$$

and $P'\Omega P = I$. Using P, (4.4.) can be transformed into $P'y^* = P'Z^*\beta + P'e^*$ and applying the OLS estimation technique to this transformed model leads to the estimator (4.3.).

If all $X_j = X$, $j = 1,..,q$, $Z = I_q \otimes X$, there will be no gain in efficiency in estimating the equation system (4.1.) simultaneously over estimating each equation separately. Moreover, assuming $\Sigma = diag\{\sigma_{ii}^2\}$, the expression $Z'\Gamma^{-1}Z$ will be equal to $Z'Z\Gamma_0^{-1}$ for Γ_0^{-1} a $(pq \times pq)$ diagonal matrix and also $Z'\Gamma^{-1}y$ can be written as $\Gamma_0^{-1}Z'y$. According to this, the modified ridge estimator (4.3.) can be simplified by replacing $k\Theta^{-1}\Gamma_0$ by a $(pq \times pq)$ diagonal matrix K, to

$$\hat{\beta}_{MGR} = (Z'Z + K)^{-1}(Z'y + Kb) = (Z'Z + K)^{-1}\left(Z'Z\hat{\beta} + Kb\right) \qquad (4.5.)$$

which is called modified generalized ridge estimator by Fule (1995).

The estimator (4.5.) is a form of a generalized ridge estimator allowing different shrinkage parameters connected with each $\hat{\beta} = (Z'Z)^{-1}Z'y$, i.e. different shrinkage within every equation but also be between equations. The estimator (4.5.) can also regarded as a modified ridge estimator, because of shrinking the OLS estimated parameters towards predetermined target values b.

Shrinking the parameter values toward predetermined b values leads to a lost of an important property of the original model. Therefore, further restrictions on the parameter values have to be imposed on the parameter values. The linear equality restrictions on the parameter vector β is

$$R\beta = r \qquad (4.6.)$$

where R is a $(p \times pq)$ matrix and r is a $(p \times 1)$ vector. The restriction (4.6.) imposed on combined model (4.4.). Applying the RLS estimation technique to model (4.4.), namely minimizing $(y^* - Z^*\beta)'\Omega^{-1}(y^* - Z^*\beta)$, subject to restrictions (4.6.), leads to the estimator

$$\hat{\beta}_{MGR}^{(r)} = \hat{\beta}_{MGR} + W^{-1}R'(RW^{-1}R')^{-1}\left(r - R\hat{\beta}_{MGR}\right)$$

where $W = Z'Z + K$.

This restricted modified generalized ridge estimator, defined by Fule (1995), presents a restricted solution for β, providing a correction of the OLS estimated parameter values through the shrinkage procedure.

4.2. Minimax Ridge Estimator

Minimax estimation is based on the idea that the quadratic risk function for the estimate $\hat{\beta}$ is not minimized over the entire parameter space R^P, but only over an area $B(\beta)$ that is restricted by a prior knowledge. If all restrictions define a convex area, this area can often be enclosed in an ellipsoid of the following form

$$B(\beta) = \{\beta: \beta'T\beta \leq k\} \tag{4.7.}$$

with the origin as center point or in

$$B(\beta, \beta_0) = \{\beta: (\beta - \beta_0)'T(\beta - \beta_0) \leq k\}$$

with the center point vector β_0.

Considering the quadratic risk $R_1\left(\hat{\beta}, A\right) = Tr\{A\left[MMSE\left(\hat{\beta}\right)\right]\}$, a class $\{\hat{\beta}\}$ of estimators and a convex region of prior restrictions for β, $B(\beta)$, the criterion of the minimax estimator is given in the following definition.

Definition 4.1. An estimator $b_* \in \{\hat{\beta}\}$ is called a minimax estimator of β if

$$\min_{\{\hat{\beta}\}} \sup_{\beta \epsilon B} R_1\left(\hat{\beta}, A\right) = \sup_{\beta \epsilon B} R_1(b_*, A).$$

Linear Minimax Estimators: In linear homogeneous estimators $\{\hat{\beta} = Cy\}$ the estimation error in $\hat{\beta}$ is expressed as

$$\hat{\beta} - \beta = (CX - I)\beta + C\varepsilon$$

from which $R_1(\hat{\beta}, A)$ is derived as

$$\begin{aligned} R_1(\hat{\beta}, A) &= E[(CX - I)\beta + C\varepsilon]'A[(CX - I)\beta + C\varepsilon] \\ &= \beta'(X'C' - I)A(CX - I)\beta + \sigma^2 Tr\{ACC'\} \\ &= \beta' \mathrm{T}^{1/2} \tilde{A} \mathrm{T}^{1/2} \beta + \sigma^2 Tr(ACC') \end{aligned}$$

where $\tilde{A} = \mathrm{T}^{-1/2}(CX - I)'A(CX - I)\mathrm{T}^{-1/2}$ and $T > 0$ is the matrix of the prior restriction (4.7.).

Theorem 4.1: Let $A: n \times n$ be symmetric with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$. Then

$$\sup_x \frac{\mathrm{x}'\mathrm{Ax}}{\mathrm{x}'\mathrm{x}} = \lambda_1, \quad \inf_x \frac{\mathrm{x}'\mathrm{Ax}}{\mathrm{x}'\mathrm{x}} = \lambda_n.$$

Using Theorem 4.1,

$$\sup_\beta \frac{\beta' \mathrm{T}^{1/2} \tilde{A} \mathrm{T}^{1/2} \beta}{\beta' T \beta} = \lambda_{max}(\tilde{A})$$

and hence

$$\sup_{\beta' T \beta \leq k} R_1(Cy, A) = \sigma^2 Tr(ACC') + k\lambda_{max}(\tilde{A}). \tag{4.8.}$$

Since the matrix $\tilde{A}$ is dependent on the matrix C, the maximum eigenvalue $\lambda_{max}(\tilde{A})$ is dependent on C. But it doesn't have an explicit form that could be used for differentiation. Kuks (1972), Kuks and Olman (1971,1972) suggested iterative solutions. Besides this, Trenkler and Stahlecker (1987) proposed the inequality $\lambda_{max}(\tilde{A}) \leq Tr(\tilde{A})$ to find an upper limit of $R_1(Cy, A)$ that is differentiable with respect to C. So a substitute problem with an explicit solution is found. This solution

can be achieved if the weight matrices are confined to matrices of the form $A = aa'$ of rank 1, so that the $R_1(\hat{\beta}, A)$ risk equals the weaker $R_2(\hat{\beta}, a)$ risk.

Linear Minimax Estimates for Matrices $= aa'$: In the case of $A = aa'$,

$$\tilde{A} = \left[T^{-1/2}(CX - I)'a\right]\left[a'(CX - I)T^{-1/2}\right] = \tilde{a}\tilde{a}'.$$

According to a theorem that $A = aa'$ with a as nonnull vector has all eigenvalues 0 expect one, with $\lambda = a'a$ and the corresponding eigenvector a, $\lambda_{max}(\tilde{A}) = \tilde{a}'\tilde{a}$. Therefore (4.8.) becomes

$$\sup_{\beta'T\beta \le k} R_2(Cy, a) = \sigma^2 a'CC'a + ka'(CX - I)T^{-1}(CX - I)'a.$$

Differentiation with respect to C leads to

$$\frac{1}{2}\frac{\partial}{\partial C}\left\{\sup_{\beta'T\beta \le k} R_2(Cy, a)\right\} = (\sigma^2 I + kXT^{-1}X')C'aa' - kXT^{-1}aa'. \qquad (4.9.)$$

Since a is any fixed vector, (4.9.) equals zero for all matrices aa' if and only if

$$C_*' = k(\sigma^2 I + kXT^{-1}X')^{-1}XT^{-1}. \qquad (4.10.)$$

After transposing and multiplying from the left with $(\sigma^2 T + kX'X)$, (4.10.) becomes

$$(\sigma^2 T + kX'X)C_* = kX'[\sigma^2 I + kXT^{-1}X'][\sigma^2 I + kXT^{-1}X']^{-1} = kX'$$

which leads to the solution

$$C_* = (S + k^{-1}\sigma^2 T)^{-1}X'$$

where $S = X'X$. Using the abbreviation $D_* = (S + k^{-1}\sigma^2 T)$ the theorems are:

Theorem 4.2: (Kuks (1972)) In linear regression model with restriction $\beta' T\beta \leq k$ with $T > 0$ and the risk function $R_2(\hat{\beta}, a)$, the linear minimax estimator is

$$b_* = (X'X + k^{-1}\sigma^2 T)^{-1}X'y$$
$$= D_*^{-1}X'y$$

with $Bias(b_*) = -k^{-1}\sigma^2 D_*^{-1}T\beta$ and $Var(b_*) = \sigma^2 D_*^{-1}SD_*^{-1}$. The minimax risk is

$$\sup_{\beta' T\beta \leq k} R_2(b_*, a) = \sigma^2 a' D_*^{-1} a.$$

Theorem 4.3: If the restriction in Theorem 4.2 is $(\beta - \beta_0)'T(\beta - \beta_0) \leq k$ with center point $\beta_0 \neq 0$, the linear minimax estimator is

$$b_*(\beta_0) = \beta_0 + D_*^{-1}X'(y - X\beta_0).$$

with $Bias\big(b_*(\beta_0)\big) = -k^{-1}\sigma^2 D_*^{-1}T(\beta - \beta_0)$ and $Var\big(b_*(\beta_0)\big) = Var(b_*)$. The minimax risk is

$$\sup_{(\beta-\beta_0)'T(\beta-\beta_0)\leq k} R_2(b_*(\beta_0), a) = \sigma^2 a' D_*^{-1} a\ .$$

A change of the center point of a prior ellipsoid has an influence only on the estimator itself and its bias. The minimax estimator is not operational, because of the unknown σ^2. The smaller the value of k, the stricter is the prior restriction for fixed T. Analogously, the larger the value of k, the smaller is the influence of $\beta' T\beta \leq k$ on the minimax estimator.

$B(\beta) \to IR^P$ as $k \to \infty$ and $\lim_{k\to\infty} b_* \to \hat{\beta} = (X'X)^{-1}X'y$.

Comparison of b_* and $\widehat{\boldsymbol{\beta}}$:

(i) Minimax Risk: The minimax risk of OLS estimator is

$$\sup_{\beta' T\beta \le k} R_2(\hat{\beta}, a) = R_2(\hat{\beta}, a) = \sigma^2 a' S^{-1} a \ .$$

The linear minimax estimator b_* has a smaller minimax risk than the OLS estimator. Therefore,

$$R_2(\hat{\beta}, a) - \sup_{\beta' T\beta \le k} R_2(b_*, a) = \sigma^2 a'(S^{-1} - (k^{-1}\sigma^2 T + S)^{-1})a \ge 0$$

since $S^{-1} - (k^{-1}\sigma^2 T + S)^{-1} \ge 0$.

(ii) MMSE superiority : The MMSE of b_* is

$$\begin{aligned} \text{MMSE}(b_*) &= Var(b_*) + Bias(b_*)Bias(b_*)' \\ &= \sigma^2 D_*^{-1}(S + k^{-2}\sigma^2 T\beta\beta' T')D_*^{-1}. \end{aligned}$$

b_* is superior to $\hat{\beta}$ if

$$\begin{aligned} \Delta(\hat{\beta}, b_*) &= MMSE(\hat{\beta}) - MMSE(b_*) \\ &= \sigma^2 D_*^{-1}[D_* S^{-1} D_* - S - k^{-2}\sigma^2 T\beta\beta' T']D_*^{-1} \ge 0. \end{aligned}$$

$\Delta(\hat{\beta}, b_*) \ge 0$ if and only if

$$\begin{aligned} B &= D_* S^{-1} D_* - S - k^{-2}\sigma^2 T\beta\beta' T' \\ &= k^{-2}\sigma^4 T[\{S^{-1} + 2k\sigma^{-2}T^{-1}\} - \sigma^{-2}\beta\beta']T \ge 0 \\ &= k^{-2}\sigma^4 T C^{1/2}[\mathrm{I} - \sigma^{-2} C^{-1/2}\beta\beta' C^{-1/2}]C^{1/2} T \ge 0 \end{aligned}$$

with $C = S^{-1} + 2k\sigma^{-2}T^{-1}$.

$\mathrm{I} - \sigma^{-2}\mathrm{C}^{-1/2}\beta\beta'\mathrm{C}^{-1/2} \geq 0$ if and only if $\sigma^{-2}\beta'(S^{-1} + 2k\sigma^{-2}T^{-1})^{-1}\beta \leq 1$. Since $(2k\sigma^{-2}T^{-1})^{-1} - (S^{-1} + 2k\sigma^{-2}T^{-1}) \geq 0$, $k^{-1} \leq \frac{2}{\beta'\beta}$ is the sufficient condition for MMSE superiority of the minimax estimator b_* compared to $\hat{\beta}$. This condition corresponds to the condition $k \leq \frac{2\sigma^2}{\beta'\beta}$ for the MMSE superiority of the RE, $\hat{\beta}_k$, compared to $\hat{\beta}$. The linear minimax estimator b_* is a ridge estimate when the ridge parameter is $k^{-1}\sigma^2$. Hence the restriction $\beta'T\beta \leq k$ has a stabilizing effect on the variance. The minimax estimator is operational if σ^2 can be included in the restriction $\beta'T\beta \leq \sigma^2 k = \tilde{k}$:

$$b_* = \left(X'X + \tilde{k}^{-1}T\right)^{-1}X'y.$$

4.3. Bayes Ridge Estimator

Bayesian Analysis of Regression: Since β is any scalar unknown parameter, its values are uncertain and therefore it can be treated as a random variable. Some subjective information or belief about β can be obtained to have information about β. Then sample evidence y about β can be looked for and combined with prior knowledge to obtain a revised information or opinion. Bayesian analysis involves this process. The prior possibilities about the values of β are formulated in terms of a probability density, called a prior probability density. The random sample vector y is expressed in terms of probability of sample data and called a likelihood function. From the definition of conditional probabilities

$$p(y|\beta) = \frac{p(y,\beta)}{p(\beta)},\ p(\beta|y) = \frac{p(y,\beta)}{p(y)} \text{ and thus } p(\beta|y) = \frac{p(\beta)p(y|\beta)}{p(y)}.$$

The unconditional probability of observing the data, $p(y)$, is equal to $\sum_\beta p(\beta)p(y|\beta)$ in discrete case and equal to $\int_\beta p(\beta)p(y|\beta)\,d\beta$ in the continuous case, and therefore it is considered as the reciprocal of the normalizing constant for the $p(\beta|y)$. So,

$$p(\beta|y) = \frac{1}{p(y)} p(\beta)p(y|\beta)$$
$$\propto p(\beta)p(y|\beta)$$
$$\propto p(\beta)L(\beta)$$

where $P(\beta|y)$ is the posterior probability density for the parameter β given the data vector y, $p(\beta)$ is the prior probability density for β and $p(y|\beta)$, viewed as a function of β, is the familiar likelihood function $L(\beta)$.

As a summary by using Bayes theorem an alternative way of combining the prior and sample information is given. For discussion of the Bayesian analysis of the linear regression model $y = X\beta + \varepsilon$ (1.1.) two different kinds of priors are defined as conjugate priors and diffuse priors. A conjugate prior is a prior that, when combined with the likelihood function, provides a posterior distribution that has the same functional form as the prior. When nothing about the parameters was known noninformative or diffuse prior is used.

Conjugate Priors: Since $\varepsilon \sim N(0, \sigma^2 I)$ and $y \sim N(X\beta, \sigma^2 I)$ the density function of y can be written as

$$f(y|\beta, \sigma, X) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right] \qquad (4.11.)$$

for a given data set on y and X this is regarded as the likelihood function for β and σ, namely,

$$L(\beta, \sigma) = L(\beta, \sigma|y) = f(y|\beta, \sigma, X).$$

Then substituting

$$(y - X\beta)'(y - X\beta) = \left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right) + \left(\beta - \hat{\beta}\right)'X'X\left(\beta - \hat{\beta}\right)$$

in (4.11.)

$$L(\beta,\sigma) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp[-\frac{1}{2\sigma^2}\{(y-X\hat{\beta})'(y-X\hat{\beta}) + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\}]$$
$$\propto \sigma^{-n}exp[-\frac{1}{2\sigma^2}\{(y-X\hat{\beta})'(y-X\hat{\beta}) + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\}]$$

where $\propto$ denotes proportionality and permits a simplification by omitting the constants.

Case 1. (σ known): $L(\beta)$ can be written as

$$L(\beta) \propto \exp[-\frac{1}{2\sigma^2}(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})]$$

because $(y-X\hat{\beta})'(y-X\hat{\beta})$ doesn't involve β. This form of likelihood suggest that one may use the normal distribution as a prior for β.

$$\mathrm{p}(\beta) = \frac{1}{(2\pi)^{p/2}|\sigma^2\Omega|^{1/2}}exp[\frac{1}{2\sigma^2}(\beta-\beta_0)'\Omega^{-1}(\beta-\beta_0)]$$
$$\propto exp[-\frac{1}{2\sigma^2}(\beta-\beta_0)'\Omega^{-1}(\beta-\beta_0)]$$

which is multivariate normal with the prior mean vector β_0 and prior variance covariance $\sigma^2\Omega$.

Using Bayes theorem the posterior density is

$$p(\beta|y) \propto p(\beta)L(\beta)$$
$$\propto \exp[-\frac{1}{2\sigma^2}\{(\beta-\hat{\beta})'X'X(\beta-\hat{\beta}) + (\beta-\beta_0)'\Omega^{-1}(\beta-\beta_0)\}].$$

Then,

$$(\beta-\hat{\beta})'X'X(\beta-\hat{\beta}) + (\beta-\beta_0)'\Omega^{-1}(\beta-\beta_0)$$
$$= \beta'(X'X+\Omega^{-1})\beta - 2\beta'(X'X\hat{\beta}+\Omega^{-1}\beta_0) + c \qquad (4.12.)$$
$$= (\beta-b^*)'(\Omega^*)^{-1}(\beta-b^*) + c^*\ ,$$

where $c = b'X'Xb + \beta_0{}'\Omega^{-1}\beta_0$ does not involve β. Further,

$$b^* = (X'X + \Omega^{-1})^{-1}\left(X'X\hat{\beta} + \Omega^{-1}\beta_0\right)$$
$$\Omega^* = (X'X + \Omega^{-1})^{-1} \qquad (4.13.)$$

and $c^* = c - (b^*)'(\Omega^*)^{-1}b^*$ does not contain β.

Using (4.12.) the posterior density is

$$p(\beta|y) \propto \exp[-\tfrac{1}{2\sigma^2}(\beta - b^*)'(\Omega^*)^{-1}(\beta - b^*)]$$

where the normalizing constant is $(2\pi)^{-n/2}|\sigma^2\Omega^*|^{-1/2}$. This density is again a multivariate normal with mean vector b^* and variance covariance matrix Ω^*, namely the posterior density of $\beta \sim N(b^*, \sigma^2\Omega^*)$. Since both the prior and posterior are normal, the prior for β given above is called a conjugate prior.

Diffuse (Improper) Priors: When nothing was known about the parameters β and σ the joint prior of β and σ is considered.

$$p(\beta, \sigma) = p(\beta|\sigma)p(\sigma)$$
$$\propto \frac{1}{\sigma} \qquad (4.14.)$$

where $p(\beta|\sigma)$ is a constant and $p(\sigma)$ is proportional to $1/\sigma$. The prior (4.14.) is referred to as a diffuse prior. Using the diffuse prior the posterior is obtained as

$$p(\beta, \sigma|y) \propto \sigma^{-(n+1)}\exp[-\tfrac{1}{2\sigma^2}\{SSE\left(\hat{\beta}\right) + \left(\beta - \hat{\beta}\right)'X'X\left(\beta - \hat{\beta}\right)\}].$$

Thus given σ the posterior of $\beta \sim N\left(\hat{\beta}, \sigma^2(X'X)^{-1}\right)$. Integrating with respect to σ the marginal posterior distribution of β is

$$p(\beta|y) \propto [\left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right) + \left(\beta - \hat{\beta}\right)'X'X\left(\beta - \hat{\beta}\right)]^{-n/2}.$$

This is multivariate t with mean vector $\hat{\beta}$ and variance covariance matrix $\frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n-p}(X'X)^{-1}$.

Bayesian Regression Estimator: In both case, conjugate and diffuse priors, the marginal posterior distribution of β is multivariate t distribution. Thus the parameter β can be analyzed by using the well known properties of the multivariate t density function. In case of conjugate priors, bayes estimator of β from (4.13.) is

$$b^* = (X'X + \Omega^{-1})^{-1}\left(X'X\hat{\beta} + \Omega^{-1}\beta_0\right) \qquad (4.15.)$$
$$= (X'X + \Omega^{-1})^{-1}X'X\hat{\beta} + (X'X + \Omega^{-1})^{-1}\Omega^{-1}\beta_0.$$

The estimator b^* is a weighted matrix combination of the estimator $\hat{\beta}$, based on data and the prior mean β_0. b^* can also be written from (4.15.) by adding and substracting $X'X\beta_0$ inside the second parentheses as:

$$b^* = (X'X + \Omega^{-1})^{-1}[X'X\left(\hat{\beta} - \beta_0\right) + (X'X + \Omega^{-1})\beta_0]$$
$$= \beta_0 + (X'X + \Omega^{-1})^{-1}X'X\left(\hat{\beta} - \beta_0\right).$$

Bayesian Interpretations of Ridge Methods: Consider the prior distribution of β as a multivariate normal with mean vector zero and variance covariance matrix σ^2K^{-1}, that is $\beta \sim N(\beta_0, \sigma^2K^{-1})$. The posterior density of β is also a multivariate normal with the posterior mean

$$b^* = (X'X + K)^{-1}(X'y + K\beta_0) \qquad (4.16.)$$
$$= (X'X + K)^{-1}X'X\hat{\beta} + K\beta_0.$$

If the prior of β is $N(0, \sigma^2K^{-1})$, then $b^* = (X'X + K)^{-1}X'y$, which is identical with the GRE. The ordinary RE implies that the prior of β is $N(0, \sigma^2k^{-1}I)$. Some Bayesians feel that this prior is unrealistic, and a prior mean other than the null vector should be used. In the absence of specific prior knowledge it is often

conservative to shrink toward zero vector. In cases when such prior knowledge about β_0 is available, the modification of RE or GRE is that one shrinks toward this known prior, as in (4.16.). There is no difficulty in subjectively specifying a non-zero β_0. The problem is that the others may prefer a different β_0 and the solution can be sensitive to this choice, that almost any solution can be obtained.

The bayesian analysis of the linear regression model (1.1.) when the parameter vector β itself has a structure which would influence the choice of a prior. Consider a simple case where this structure on β is given by $E(\beta) = A\beta_0$, where A is a $p \times p_0$ known matrix and β_0 is a $p_0 \times 1$ vector of hyperparameters. Let β have a multivariate normal distribution with known variance covariance matrix Ω_0. Then the structure implies the following normal prior distribution of $\beta \sim N(A\beta_0, \sigma^2\Omega_0)$. According to Lindley and Smith (1972) if the prior of $\beta \sim N(\beta_0, \sigma^2 K^{-1})$ and further, the prior on the hyperparameter of β is a diffuse prior, then the resulting mean of the posterior density of β is the GRE.

4.4. Robust Ridge Regression Estimators

Multicollinearity and nonnormal error distributions often plague researchers to use regression techniques. When the explanatory variables in a regression equation are highly correlated, the OLS coefficient estimates will be imprecise. Adding or deleting variables may change the estimated regression coefficient values, coefficients may have signs that are opposite of what would be expected. Important variables may have statistically insignificant coefficients because of inflated standard errors when traditional hypothesis testing procedures are used.

One of the remedial measures suggested for the multicollinearity problem is ridge regression. The RE of the regression coefficients is biased but has smaller sampling variability. Thus the ridge regression might be useful in providing estimates which are more precise and stable than OLS estimates in the presence of multicollinearity.

The OLS estimator possesses minimum variance among all unbiased estimators when the disturbances of the regression equation are independent, identically distributed, normal random variables.

If the disturbances are not normally distributed, the OLS estimator can be shown to possess minimum variance among all linear unbiased estimators provided the variance of the disturbance distribution is finite. However, OLS estimator may produce extremely poor estimates in the presence of nonnormal disturbance distributions. Koenker (1982) pointed out that the power of conventional tests may be low when disturbances are nonnormal even though the tests may have the correct level of significance asymptotically. It is also noted that the class of linear estimators is very restrictive and may exclude estimators which outperform least squares in cases of nonnormal disturbances (Koenker, 1982). Estimators which are not as strongly affected by nonnormal disturbances are robust regression estimators. Such robust estimators are intended to be more efficient estimators than OLS estimator when disturbances are nonnormal. The term nonnormal disturbances will indicate disturbance distributions which have fatter tails than normal distributions. These fat tailed distributions are more prone to produce outliers. The adverse effect of outliers on OLS fitted regression line is well known and it is protection against the effect of outliers that is being sought through the use of robust estimators. Several different classifications of robust methods exist. Three of the most commonly considered groups are R-estimators, L-estimators and M-estimators.

R-estimators are estimators based on ranking of the residuals from the regression. To obtain R-estimates of the regression coefficients, the function to be minimized can be written

$$\sum_{i=1}^{n} a(R_i)\hat{e}_i$$

where $\hat{e}_i = y_i - X_i\hat{\beta}$ is the regression residual, R_i is the rank of $\hat{e}_i$ and $a(.)$ is some monotone scores function satisfying

$$\sum_{i=1}^{n} a(i) = 0.$$

L-estimators are linear combinations of the order statistics. As measures of location the construction of L-estimators is fairly straightforward. For example the sample quantiles are L-estimators and any linear combination of the sample quantiles would also be an L-estimator. Koenker and Bassett (1978) developed estimators called the regression quantiles that generalize the L-estimators to linear models. They define θth regression quantile as the solution to minimization problem

$$\min_{\beta} \sum_{\{i|y_i \geq x_i'\beta\}} \theta|y_i - x_i'\beta| + \sum_{\{i|y_i < x_i'\beta\}} (1-\theta)|y_i - x_i'\beta| \qquad (4.17.)$$

where $0 < \theta < 1$, y_i is the ith observation on the dependent variable, x_i is the $p \times 1$ vector of independent variable values for the ith observation and β is the $p \times 1$ vector of unknown regression coefficients to be estimated. When $p = 1$ and the x values are identically one, the solution to the problem in the equation (4.17.) reduces to the θth sample quantiles. When $\theta = 1/2$, the minimization problem in equation (4.17.) becomes

$$\min_{\beta} \sum_{i=1}^{n} |y_i - x_i'\beta| \qquad (4.18.)$$

or minimizing the sum of the absolute values of regression residuals. The solution to this problem is called the least absolute value (LAV) estimator. The LAV estimator is one of the class of regression L-estimators. Rather than minimizing the sum of squared residuals as in OLS estimation, the sum of the absolute values of the residuals is minimized. Thus the effect of outliers on the LAV estimates will be less than that on OLS estimates. The LAV estimator is also a special class of the M-estimators.

M-estimators for linear regression are defined as solutions to the following minimization problem

$$\min_{\beta} \sum_{i=1}^{n} p(y_i - x_i'\beta)$$

where p(.) is a strictly convex function of the regression residuals. Alternatively, M-estimators can be defined as the solution to the equations

$$\sum_{i=1}^{n} x_i \, \psi(y_i - x_i'\beta) = 0$$

where the derivative of p(.) is the function $\psi(.)$. The terms y_i, x_i and β are as defined for equation (4.17.). The OLS estimator can be defined as an M–estimator by defining

$$p(y_i - x_i'\beta) = 1/2\,(y_i - x_i'\beta)^2$$

and

$$\psi(y_i - x_i'\beta) = y_i - x_i'\beta.$$

When both outliers and multicollinearity occur a data set, it would be beneficial to combine methods designed to deal with these problems individually. RE and some robust estimation technique can be combined to produce a robust ridge regression estimator and it is hoped that the resulting robust ridge regression estimators will be resistant to multicollinearity problems and less affected by extreme observations. Askin and Montgomery (1980) discuss augmented robust estimators as a way of combining biased and robust regression techniques. The procedure suggested by Askin and Montgomery (1980) is based on the fact that M-estimates can be computed using weighted least squares procedures. The weighted least squares estimator can be written as

$$\hat{\beta}_{WLS} = (X'WX)^{-1}X'Wy$$

where W is a diagonal matrix with diagonal elements W_{ii}. The W_{ii} are weights applied to the observations and are intended to downweight the extreme

observations. The weights can be determined using any M-estimate ψ function. Askin and Montgomery (1980) suggest using weights

$$W_{ii} = \frac{\psi(y_i - x_i'\beta)}{y_i - x_i'\beta}. \tag{4.19.}$$

The weighted least squares estimates can be computed by applying least squares to the transformed observations $\sqrt{W_{ii}}y_i$ and $\sqrt{W_{ii}}x_i$. This produces an estimate equivalent to $\hat{\beta}_{WLS}$ in equation (4.19.).

To compute robust ridge estimates the formula

$$\hat{\beta}_{WRidge} = (X'WX + kI)^{-1}X'Wy$$

can be used. Askin and Montgomery compute the value of k from the untransformed observations, although the transformed observations could be used. The estimator $\hat{\beta}_{WRidge}$ with the biasing parameter k and the weights W_{ii} determined from the data will be referred to as the weighted ridge estimator. Lawrence and Marsh (1984) gave the applications of the weighted ridge estimator using various weighting schemes.

Pfaffenberger and Dielman (1984,1985) suggested another robust ridge estimator which combines properties of the LAV estimator and RE

$$\hat{\beta}_{RLAV} = (X'X + k^*I)^{-1}X'y.$$

The value of k^* is determined from the data using

$$k^* = \frac{ps_{LAV}^2}{\hat{\beta}_{LAV}'\hat{\beta}_{LAV}} \tag{4.20.}$$

where

$$s_{LAV}^2 = \frac{(y - X\hat{\beta}_{LAV})'(y - X\hat{\beta}_{LAV})}{n-p}$$

and $\hat{\beta}_{LAV}$ is the LAV estimator defined as the solution to equation (4.18.). The value of k^* is the estimator of k suggested by Hoerl et al. (1975) with two important exceptions. First, the LAV estimator of β is used rather than the OLS estimator. Second, the estimator of σ^2 used in formula (4.20.) is modified by using the LAV coefficient estimates rather than OLS estimates. These two changes are intended to reduce the effect of outliers on the value chosen for the biasing parameter.

4.5. Continuum Regression and Ridge Regression

For multiple linear regression with non-orthogonal regressors there are several alternatives to the OLS method such as PCR, RR and partial least squares (PLS) regression. Continuum regression (CR), introduced by Stone and Brooks (1990), ties together these more classical regression methods PCR, RR and PLS. Let the regression to be fitted $y = X\beta$, in centered x and y variables. When multicollinearity exists the regression coefficients are likely to be quite unstable or possibly even non-unique and they cannot be interpreted individually. In this sense identification of a true regression will be impossible. However, satisfactory prediction may still be quite feasible. Cross-validation and related validation techniques will play an important role in the construction of a linear predictor for future y-values. Exact collinearities are evidently unavoidable if there are more x variables than observations. However, even if there are sufficiently many observations to guarantee a full rank design matrix, most variability in x is likely to concentrate in a relatively small-dimensional space.

OLS, PCR and PLS: OLS may be characterized as maximizing correlation. The multiple correlation coefficient R is the maximum overall direction vectors c of the correlation $R(t, y)$ between y and a regressor $t = Xc$. The desired direction c of course satisfies OLS estimator $\hat{\beta}$. Regression of y on t, by simple one-dimensional least squares, yields the OLS multiple regression of y on X.

The PCR and PLS methods depend on other maximization criteria for construction of regressors from X. In a standard PCR the first regressor is the first principal component, formed by letting c_1 be the direction of highest variability in x,

that maximizes the variance $Var(t_1)$ for $t_1 = Xc_1$, $|c_1| = 1$. Successive regressors $t_2 = Xc_2$ and etc. are obtained by repeating the procedure on the residuals from the preceding step, and the now principal components will be uncorrelated with the preceding ones. The number of principal components to be used jointly as regressors will typically be optimized by cross-validation. PLS selects regressors t_1, t_2 etc. by maximizing covariance $cov(t, y)$, rather than the variance. Although this fact PLS and PCR are analogous. Typically, PLS requires fewer PLS factors than PCR needs principal components. This is because principal components needs not necessarily be correlated with y, where as all PLS-factors must be. The reason that PCR may provide a useful regression equation at all, is that it avoids direction c with small variation, and those are the ones which may cause the collinearity problems. PLS may be regarded as a compromise between OLS and PCR, $cov^2(t, y)$ being proportional to the product $R^2(t, y)Var(t)$.

Continuum Regression (CR): In continuum regression the construction rule for new regressors encompasses the three constructions discussed which correspond to special values of a control parameter γ selected within a continuum of possible values. The construction rule can be expressed as follows maximize the function.

$$g_\gamma\big(R^2(t, y), Var(t)\big) = R^2(t, y)Var(t)^\gamma. \qquad (4.21.)$$

Over directions c, with $|c| = 1$, where $t = Xc$ and $\gamma \geq 0$. $\gamma = 0$ yield OLS and $\gamma = 1$ yields PLS, PCR is obtained as $\gamma \to \infty$ in the limit. As is the case with PLS and PCR, the construction rule is applied first on the centered x data and next successively regression on the residuals from the regression in each step, such that the regressors t_1, t_2 etc. are uncorrelated. For each set of regressors, y is then regressed on it, using least squares method. Continuum regression is scale-dependent, a property it shares with other shrinkage regressions, such as PLS, PCR and RR.

Continuum Regression and Ridge Regression: Examining more closely the construction rule (4.21.), it is observed that it yields a regressor $t = Xc$ which is proportional to the ridge regression of y on x, for some ridge constant k dependent on

γ. Hence, CR is closely related to ridge regression. First factor (first regressor) continuum regression, CR(1), is interpreted as a an upscaled ridge regression, with $\frac{1}{1-\gamma}$ as the scale factor, i.e $\hat{\beta}_{CR}(\gamma) = \left(\frac{1}{1-\gamma}\right)\hat{\beta}_k$. For any function $g\left(R^2(t,y), Var(t)\right)$ of correlation and variance, satisfying the natural condition of being increasing in each one of its two arguments, the maximizing regressor $t = Xc$ is of the ridge type for some ridge constant k that depends on the particular function g. The OLS is obtained for $k = 0$ and PLS is arrived for $k \to \pm\infty$ while PCR appears in the limit of k approaches the minus of the largest eigenvalue of $X'X$. Ridge regression is regarded as bringing in two shrinkage effects: one two compensate for near collinearity in the regressors and the other to reduce the MSE by replacing variance by squared bias without any connection with collinearity or regression. On the other hand CR(1) shrinks only to compensate for collinearity. This motivates a modification of conventional ridge regression by the scalar compensation factor $\frac{1}{1-\gamma}$, so that it will correspond to the special case of CR. This modification will be preferable to conventional ridge regression for statisticians who do not like shrinkage estimators as a principle.

5. TWO PARAMETER RIDGE ESTIMATOR AND SOME COMPARISONS

5.1. Two Parameter Ridge Estimator

Lipovetsky and Conklin (2005) proposed two parameter ridge regression estimator as an alternative to OLS estimator in the presence of multicollinearity. Lipovetsky (2006) improved the two parameter model and examined the properties of the parameters like efficiency numbers, the residual error variance, variance, generalized cross validation criterion and asymptotic behaviors.

OLS: Consider the multiple linear regression model (1.1.). For the standardized variables $y'y = 1$, $C = X'X$ is the correlation matrix of the regressors and $r = X'y$ is the vector of correlations of the dependent variable with the regressors. The least squares objective for the regression corresponds to minimizing the sum of squared deviations is

$$S^2 = \|\varepsilon\|^2 = (y - X\beta)'(y - X\beta) = 1 - 2\beta'r + \beta'C\beta\,. \tag{5.1.}$$

The normal system equations with the corresponding solution is

$$C\beta = r,\ \hat{\beta} = C^{-1}r. \tag{5.2.}$$

The quality of the model is estimated by the residual sum of squares (5.1.) or by the coefficient of multiple determination defined from (5.1.) as

$$R^2 = 1 - S^2 = 2\beta'r - \beta'C\beta = \beta'(2r - C\beta). \tag{5.3.}$$

If the objective (5.1.) is minimum, R^2 reaches its maximum. Substituting $C\beta = r$, R^2 can be presented in equivalent form

$$R^2 = \beta'r = \beta'C\beta. \tag{5.4.}$$

The items of the scalar product in (5.4.) are called the net effects $R_j{}^2$ used to estimate the contribution of the regressors into the model

$$R^2 = \sum_{j=1}^{p} \beta_j r_j = \sum_{j=1}^{p} R_j^2. \quad (5.5.)$$

There are some properties of solution for the OLS. Relation $C\beta = r$ can be represented as

$$X'\varepsilon = 0 \quad (5.6.)$$

that expresses orthogonality of each regressor to the error vector

$$\varepsilon = y - X\beta = y - \tilde{y} \quad (5.7.)$$

where $\tilde{y}$ denotes the theoretical, predicted by the model values of the dependent variable

$$\tilde{y} = X\beta. \quad (5.8.)$$

Using (5.7.), (5.6.) can be rewritten as

$$X'y = X'\tilde{y} \quad (5.9.)$$

so the vector of pair correlations r of regressors with the empirical y coincides with such a vector of correlations with the theoretical $\tilde{y}$. The relation of orthogonality in (5.6.) is also true for the theoretical $\tilde{y}$,

$$\tilde{y}'\varepsilon = 0. \quad (5.10.)$$

Using $\varepsilon = y - \tilde{y}$ in equation (5.10.)

$$y'\tilde{y} = \tilde{y}'\tilde{y} = \beta' X' X\beta = R^2 \tag{5.11.}$$

is obtained. Projection of the empirical vector y on the vector of errors, taking into account the equation (5.11.), produces the relation

$$y'\varepsilon = y'y - y'\tilde{y} = 1 - \tilde{y}'\tilde{y} = 1 - R^2. \tag{5.12.}$$

Substituting equation (5.8.) in $y = X\beta + \varepsilon$

$$y'y = (\tilde{y} + \varepsilon)'(\tilde{y} + \varepsilon) = \tilde{y}'\tilde{y} + 2\tilde{y}'\varepsilon + \varepsilon'\varepsilon \tag{5.13.}$$

is obtained. Using the property of orthogonality (5.10.), relation (5.11.) for the coefficient of multiple determination and (5.1.) for the residual sum of squares, the equality is obtained as

$$1 = R^2 + S^2 \tag{5.14.}$$

which is a Pythagorean relation between the sum of squares explained (R^2) and non-explained (S^2) by the regression.

This decomposition of the empirical standardized squared norm of the dependent variable is very convenient for estimating the model quality by the total F-statistics defined by the ratio of variances for the regression related and the residual portions in (5.14.).

If any regressors are highly correlated or multicollinear, matrix C becomes ill-conditioned, so its determinant is close to zero, and the inverse matrix produces a solution with highly inflated values of coefficients of regression. The values of these coefficients have signs opposite to corresponding pair correlations of regressor with dependent variable, important variables become statistically insignificant and some of the net effects become negative. The model can be applied for prediction but it becomes practically useless for the analysis and interpretation of the predictors' role

in the model. Overcoming the difficulties of multicollinearity ridge regression is suggested.

Ridge Regression (Ridge-1): For the standardized variables, adding to the least squares objective (5.1.) a penalizing function of the squared norm of the vector of regression coefficients yields a conditional objective

$$S^2 = \|\varepsilon\|^2 + k\|\beta_k\|^2 = 1 - 2\beta_k' r + \beta_k' C \beta_k + k\beta_k' \beta_k \tag{5.15.}$$

where β_k denotes a vector of ridge regression estimates for the coefficients and k is a positive parameter. From maximizing the objective function by vector β_k the system of equations and the solution of this system is

$$(C + kI)\beta_k = r\,,\ \hat{\beta}_k = (C + kI)^{-1} r,\ \ k \geq 0. \tag{5.16.}$$

The ridge regression proved to be very helpful in applied regression modelling, although its results lack the quality of fit of a regular regression, and cannot always be clearly interpreted because one-parameter or Ridge-1 properties are more similar to a non-linear regression. For instance the multiple determination coefficient is

$$\begin{aligned} R_k^2 &= \beta_k'(2r - C\beta_k) \\ &= 2r'(C + kI)^{-1} r - r'(C + kI)^{-1} C (C + kI)^{-1} r. \end{aligned} \tag{5.17.}$$

But

$$r'(C + kI)^{-1} r \neq r'(C + kI)^{-1} C (C + kI)^{-1} r$$

or in other notation

$$\beta_k' r \neq \beta_k' C \beta_k \tag{5.18.}$$

so the coefficient of multiple determination (5.17.) of Ridge-1 cannot be reduced to an expression similar to (5.4.) for the OLS model. With the theoretical vector $\tilde{y}_k = X\beta_k$ from the Ridge-1 solution, the relation (5.18.) can be represented as $y'\tilde{y}_k \neq \tilde{y}_k{}'\tilde{y}_k$, so Ridge-1 does not satisfy the relation (5.4.) and (5.14.) holding for the OLS regression.

The eigenproblem $C\alpha = \lambda\alpha$ for the matrix of correlations among the regressors yields the spectral decomposition $C = Adiag(\lambda_j)A'$, where A is the matrix containing the eigenvectors α_j in its columns and $diag(\lambda_j)$ is a diagonal matrix of the eigenvalues λ_j.

Some other characteristics of Ridge-1 model include effective numbers of parameters, residual error variance and generalized cross-validation (GCV) of prediction quality. Using the spectral decomposition of C, the effective number of parameters is:

$$p^* = Tr((C + kI)^{-1}C) = \sum_{j=1}^{p} \frac{\lambda_j}{\lambda_j + k}\,. \tag{5.19.}$$

The residual error variance is

$$S_{res}^2 = (1 - R_k^2)/(n - p - 1). \tag{5.20.}$$

The variance of the parameter estimates is

$$\begin{aligned} Var(\beta_k)\cdot p &= \|\beta_k - E(\beta_k)\|^2 \\ &= \beta_k'\beta_k - 2\beta_k' E(\beta_k) + E(\beta_k')E(\beta_k). \end{aligned} \tag{5.21.}$$

The generalized cross-validation criterion widely used for choosing k value is defined as

$$GCV = n\frac{\|(I-H)y\|^2}{(Tr(I-H))^2} \tag{5.22.}$$

where the hat matrix H corresponds to $\tilde{y} = X\beta_k = X(C + kI)^{-1}X'y \equiv Hy$. Note that $Tr(H) = Tr(X'X(C + kI)^{-1})$ so it equals the effective number of parameters (5.19.). Then (5.22.) can be represented explicitly as follows

$$GCV = n\frac{\|y-\tilde{y}\|^2}{(n-p^*)^2} = \frac{1}{1-p^*/n} \cdot \frac{\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2}{n-p^*}.$$

Two Parameter Ridge Regression (Ridge-2): Consider a generalization of the regularization (5.15.) with two additional items

$$\begin{aligned} S^2 &= \|y - X\beta_{qk}\|^2 + k_1\|\beta_{qk}\|^2 + k_2\|X'y - \beta_{qk}\|^2 + k_3\|y'(y - X\beta_{qk})\|^2 \\ &= (1 - 2\beta'_{qk}r + \beta'_{qk}C\beta_{qk}) + k_1(\beta'_{qk}\beta_{qk}) + k_2(r'r - 2\beta'_{qk}r + \beta'_{qk}\beta_{qk}) \\ &\quad + k_3(1 - 2\beta'_{qk}r + \beta'_{qk}rr'\beta_{qk}). \end{aligned} \quad (5.23.)$$

The first two items in (5.23.) coincide with those in the Ridge-1 objective in (5.15.). The next item with k_2 pushes the estimates β_{qk} closer to the paired correlations r with the dependent variable, which helps to get a solution with interpretable coefficients. The last item with k_3 expresses the relation $y'\varepsilon = y'y - y'\tilde{y} = 1 - \tilde{y}'\tilde{y} = 1 - R^2$ due to (5.11.), so its minimizing corresponds to obtaining the maximum coefficient of multiple determination. Minimization of (5.23.) yields a matrix equation $C\beta_{qk} + k_1\beta_{qk} + k_2\beta_{qk} + k_3rr'\beta_{qk} = r + k_2r + k_3r$. The scalar product $r'\beta_{qk}$ can be considered as another constant and combined with the parameter k_3, so this item on the left hand side is proportional to vector r and can be transferred to the right hand side of this equation. Then combining constant on each side of this equation, it is easy to reduce it to the following system with corresponding solution

$$(C + kI)\beta_{qk} = qr, \qquad \hat{\beta}_{qk} = q(C + kI)^{-1}r \quad (5.24.)$$

where k and q are two new constant parameters. It is a Ridge-2 model that differs from the Ridge-1 only in the parameter q, so the solution of Ridge-2 is proportional to the Ridge-1 solution.

For a current value of the free ridge parameter k, the value of the second parameter q can be found using a criterion of maximum quality of the regression fit. For such a criterion, firstly coefficient of multiple determination can be considered for the Ridge-2 model

$$R_{qk}^2 = 2q[r'(C+kI)^{-1}r] - q^2[r'(C+kI)^{-1}C(C+kI)^{-1}r] \\ = 2qQ_1 - q^2Q_2 \qquad (5.25.)$$

where Q_1 and Q_2 denote two quadratic forms. The value of q where this concave function reaches its maximum is

$$q = \frac{Q_1}{Q_2} = \frac{r'(C+kI)^{-1}r}{r'(C+kI)^{-1}C(C+kI)^{-1}r}. \qquad (5.26.)$$

So the second parameter in the solution (5.24.) is uniquely defined as a quotient of two quadratic forms dependent on one parameter k. Substituting q from (5.26.) into (5.25.) yields the maximum coefficient of multiple determination as

$$R_{qk}^2 = \frac{[r'(C+kI)^{-1}r]^2}{r'(C+kI)^{-1}C(C+kI)^{-1}r} = r'\beta_{qk} = \beta_{qk}'C\beta_{qk} \qquad (5.27.)$$

which is presented similarly to the OLS regression (5.4.).

Using the theoretical vector $\tilde{y} = X\beta_{qk}$ it is easy to see that the right hand side equality in (5.27.) presents the property $y'\tilde{y} = \tilde{y}'\tilde{y}$ for Ridge-2 similar to (5.11.) for OLS. This equality leads to $\tilde{y}'\varepsilon = 0$, so (5.10.) and all relations (5.11.) - (5.14.) are true for Ridge-2. Conditions of orthogonality of the predictors with errors (5.6.) are not fulfilled exactly but numerical evaluations show that the deviations from orthogonality for Ridge-2 are always close to zero. Instead of orthogonality for each predictor (5.6.) there exists orthogonality of their linear combination with the weights

of the coefficients of Ridge-2. Indeed rewriting $\tilde{y}'\varepsilon = 0$ as $\beta'_{qk}X'e = 0$ shows that the scalar product of the vector of solution β_{qk} with the vector $X'\varepsilon$ of deviations from orthogonality equals zero.

Thus in the Ridge-2 solution the term k serves for regularization of an ill-conditioned matrix and the term q is used for tuning the fit quality. The coefficient of multiple determination for Ridge-2 model with the additional parameter is always bigger than that of Ridge-1 model. Using the term (5.26.) in (5.24.) yields the Ridge-2 solution in the explicit form

$$\beta_{qk} = \frac{r'(C+kI)^{-1}r}{r'(C+kI)^{-1}C(C+kI)^{-1}r}(C+kI)^{-1}r. \qquad (5.28.)$$

Considering the behavior of this solution with the increasing parameter k, namely in the limit of big k, the matrix $C + kI$ gets a dominant diagonal so the inverse matrix $(C + kI)^{-1}$ reduces to the scalar matrix $k^{-1}I$. The terms k^{-2} in the numerator and denominator of (5.28.) are mutually cancelled and the Ridge-2 solution eventually converges to an asymptote independent of k

$$\beta_{qk} = \frac{k^{-2}r'r}{k^{-2}r'Cr}r = \left(\frac{r'r}{r'Cr}\right)r = \gamma r \qquad (5.29.)$$

where the constant γ is defined as the always positive ratio of two quadratic forms. Thus, unlike for the Ridge-1 solution (5.16.) diminishing to zero, the coefficients of the Ridge-2 solution become proportional to the elements of the vector r (5.29.) of the pairwise regression with beta coefficients equal the pair correlations of y with each regressor. The signs of the Ridge-2 coefficients β_{qk} coincide with the signs of the pair correlations r. This guarantees clear interpretability of the eventual Ridge-2 solution and the positive net effect contributions $\beta_{qk(j)}r_j = \gamma r_j{}^2$ (5.4.) of the regressors to the coefficient of multiple determination (5.27.).

Behaviors of the tuning parameter q (5.26.) and of the coefficient of multiple determination (5.27.) with k increasing are as follows, respectively

$$q = \frac{k^{-1}r'r}{k^{-2}r'Cr} = \gamma k,\ R_{qk}^2 = \frac{k^{-2}(r'r)^2}{k^{-2}r'Cr} = \gamma(r'r) \qquad (5.30.)$$

with the same constant γ as in (5.29.). Thus, eventually q is linearly proportional to k, but the coefficient of multiple determination with k increasing goes to a constant independent of k, so the quality of fit is not decreasing. The constant of proportionality in (5.29.) and (5.30.) can be estimated using the squared norm defined by the maximum eigenvalue of the eigenproblem $C\alpha = \lambda\alpha$ so

$$\gamma = \frac{r'r}{r'Cr} \geq \frac{\|r\|^2}{\|r\|^2\|C\|} = \frac{1}{\lambda_{max}}.$$

For increasing parameter k, the Ridge-2 coefficient of multiple determination R_{qk}^2 (5.27.) stays consistently, (5.30.), close to the maximum R^2 value, (5.4.), of the OLS model, while the Ridge-1 coefficient ${R_k}^2$, (5.17.), quickly diminshes to zero. This means that in Ridge-2 modelling, k can be increased without losing the quality of regression until the asymptotic solution (5.29.) is reached, with the interpretable coefficients of multiple regression proportional to the pair correlations of y with the $x's$.

The effective number of parameters p_2^* for Ridge-2 regression is

$$p_2^* = qTr((C + kI)^{-1}C) = \frac{k}{\lambda_{max}} \cdot \frac{Tr(C)}{k} = \frac{p}{\lambda_{max}}.$$

The residual error variance for Ridge-2 can be found as

$$S_{res}^2 = \frac{1 - \frac{r'r}{\lambda_{max}}}{n - \frac{p}{\lambda_{max}} - 1}.$$

The variance of the parameter estimates is

$$Var(\beta_{qk}) \cdot p = \beta'_{qk}\beta_{qk} - 2q\beta'_{qk}(C+kI)^{-1}C\beta_{qk} + q^2\beta'_{qk}(C+kI)^{-2}C^2\beta_{qk}. \quad (5.31.)$$

The tuning parameter q can not only be found from the criterion (5.25.) of the maximum for the coefficient of multiple determination. Consider a criterion of a minimum variance of coefficient estimates, (5.31.), with β_{qk}, (5.24.), when (5.31.) reduces to

$$Var(\beta_{qk}) \cdot p = q^2[r'(C+kI)^{-2}r] - 2q^3[r'(C+kI)^{-3}Cr] + q^4[r'(C+kI)^{-4}C^2r](5.32.)$$
$$= q^2Q_3 - 2q^3Q_4 + q^4Q_5 \quad (5.32.)$$

where Q_3, Q_4 and Q_5 denote the corresponding quadratic forms in the brackets. (5.32.) is minimized by

$$q = \frac{(3Q_4 + \sqrt{9Q_4 - 8Q_3Q_5})}{4Q_5}.$$

The tuning parameter q can also be found by minimization of the GCV criterion (5.22) applied to Ridge-2

$$GCV = n\frac{1 - 2q[r'(C+kI)^{-1}r] + q^2[r'(C+kI)^{-1}C(C+kI)^{-1}r]}{\left(n - qTr((C+kI)^{-1}C)\right)^2}$$
$$= n\frac{1 - 2qQ_1 + q^2Q_2}{(n - qp^*)^2}. \quad (5.33.)$$

Q_1 and Q_2 denote the same quadratic forms as in (5.25.). Minimizing (5.33.) using the parameter q yields the expression

$$q = \frac{\left(Q_1 - \frac{p^*}{n}\right)}{\left(Q_2 - Q_1\frac{p^*}{n}\right)}.$$

The fraction p^*/n is usually small and diminishing with k increase; this expression is actually close to $q = Q_1/Q_2$, that is the same q parameter as in (5.26.). This guarantees the predictive optimum due to GCV criterion (5.33.).

All regular statistical characteristics of a multiple regression can be constructed for the Ridge-2 model as well, including the variance-covariance matrix for the coefficient estimates and their t-statistics. The F-statistic for Ridge-2 regression with the property of regular decomposition (5.14.) can be defined similarly to OLS regression via the coefficient of multiple determination.

5.2. The Comparisons of the Estimators

Biased estimators are revealed and examined to cope with the multicollinearity problem. There are many type of biased estimators and comparison methods to compare these estimators. One of the comparison methods is the MMSE criterion. In this section the OLS estimator, Ridge-1 estimator, Ridge-2 estimator and $\hat{\beta}_s = s\hat{\beta}$, $0 < s < 1$ (Mayer and Wilke (1973), Stein type estimator) are compared. The MMSE for these estimators are

$$MMSE(\hat{\beta}) = \sigma^2 C^{-1}$$
$$MMSE(\hat{\beta}_k) = \sigma^2 G_k + k^2 (C + kI)^{-1} \beta\beta' (C + kI)^{-1}$$
$$MMSE(\hat{\beta}_{qk}) = q^2\sigma^2 G_k + [q(C + kI)^{-1}C - I]\beta\beta' [q(C + kI)^{-1}C - I]'$$
$$MMSE(\hat{\beta}_s) = s^2\sigma^2 C^{-1} + (s - 1)^2 \beta\beta'.$$

where $G_k = (C + kI)^{-1} C (C + kI)^{-1}$.

The estimators can be rewritten by using the spectral decomposition of $C = P\Lambda P'$ where Λ is a diagonal matrix whose diagonal elements are eigenvalues of $X'X$ and P is a $p \times p$ matrix whose elements are the eigenvectors of $X'X$

$$\hat{\beta} = (P\Lambda^{-1}P')r \quad (5.34.)$$
$$\hat{\beta}_k = (P\Lambda P' + kI)^{-1} r \quad (5.35.)$$

$$\hat{\beta}_{qk} = q(P\Lambda P' + kI)^{-1}r \quad (5.36.)$$

$$\hat{\beta}_s = s(P\Lambda^{-1}P')r \ . \quad (5.37.)$$

MMSE of the estimators can be rewritten as

$$MMSE(\hat{\beta}) = \sigma^2 P\Lambda^{-1}P' \quad (5.38.)$$

$$MMSE(\hat{\beta}_k) = \sigma^2 H_k + k^2(P\Lambda P' + kI)^{-1}\beta\beta'(P\Lambda P' + kI)^{-1} \quad (5.39.)$$

$$MMSE(\hat{\beta}_{qk}) = q^2\sigma^2 H_k + [q(\mathrm{P\Lambda P'} + kI)^{-1}\mathrm{P\Lambda P'} - I]\beta\beta'[q(\mathrm{P\Lambda P'} + kI)^{-1}\mathrm{P\Lambda P'} - I]' \quad (5.40.)$$

$$MMSE(\hat{\beta}_s) = s^2\sigma^2\ \mathrm{P\Lambda^{-1}P'} + (\mathrm{s} - 1)^2\ \beta\beta' \quad (5.41.)$$

where $H_k = (\mathrm{P\Lambda P'} + kI)^{-1}\mathrm{P\Lambda P'}(\mathrm{P\Lambda P'} + kI)^{-1}$.

Comparisons of MMSE are done by using the quality of fit $q > 1$ which is obtained from maximizing the multiple correlation coefficient (5.26.). The theorems of Farebrother (1976) and Trenkler and Toutenberg (1990) are used as o reference while doing the comparisons.

Lemma 5.1: (Farebrother (1976)) Let A is pd matrix and a is a column vector. The necessary and sufficient condition for $A - aa' > 0$ is $a'A^{-1}a < 1$.

Lemma 5.2: (Trenkler and Toutenberg (1990)) $\tilde{\beta}_1$ and $\tilde{\beta}_2$ is two estimator of β. Let $\mathrm{D} = Var(\tilde{\beta}_1) - Var(\tilde{\beta}_2)$ be pd and $d_1 = Bias(\tilde{\beta}_1)$ and $d_2 = Bias(\tilde{\beta}_2)$. The necessary and sufficient condition for $MMSE(\tilde{\beta}_1) - MMSE(\tilde{\beta}_2) > 0$ is $d_2'(D + d_1 d_1')^{-1}d_2 < 1$.

5.2.1. The Comparison of OLS estimator and Ridge-2 Estimator

A criterion to compare the OLS estimator and Ridge-2 estimator for a fixed value of q is given in theorem (5.1.).

Theorem 5.1: Let $\mathrm{k} > \mathrm{k}_{1\mathrm{j}}$. Necessary and sufficient condition for

$MMSE(\hat{\beta}) - MMSE(\hat{\beta}_{qk}) > 0$ is $b_2'[(C^{-1} - q^2 G_k)]^{-1}b_2 < \sigma^2$ where

$k_{1j} = \lambda_j(q-1),\ j = 1,..,p$ and $b_2 = Bias(\hat{\beta}_{qk})$.

Proof: The difference of variance covariance matrix estimators (5.34.) and (5.36.) is

$$Var(\hat{\beta}) - Var(\hat{\beta}_{qk}) = \sigma^2(C^{-1} - q^2 G_k)$$
$$= \sigma^2[C^{-1} - q^2(P\Lambda P' + kI)^{-1}P\Lambda P'(P\Lambda P' + kI)^{-1}]$$
$$= \sigma^2\left[P\ diag\left\{\frac{1}{\lambda_j} - \frac{q^2\lambda_j}{(\lambda_j+k)^2}\right\}_{j=1}^{p} P'\right]$$

$(C^{-1} - q^2 G_k)$ is pd if $k^2 + 2\lambda_j k + (1-q^2)\lambda_j^2 > 0$.

The condition of being pd is achieved when $k > k_{1j}$. From Lemma 5.1 the theorem is proved.

5.2.2. The Comparison of Ridge-1 Estimator and Ridge-2 Estimator

Theorems below are given to compare Ridge-1 estimator and Ridge-2 estimator. Let $b_2 = Bias(\hat{\beta}_{qk})$ and $b_3 = Bias(\hat{\beta}_k)$.

Theorem 5.2: If $b_3' G_k^{-1} b_3 < \sigma^2(q^2-1)$ then $MMSE(\hat{\beta}_{qk}) - MMSE(\hat{\beta}_k) > 0$.

Proof: The difference of variance covariance matrix between the estimator (5.35.) and (5.36.) is $(\hat{\beta}_{qk}) - Var(\hat{\beta}_k) = \sigma^2(q^2-1)G_k$. This difference is always pd for $k > 0$ and $q > 1$ and from Lemma 5.1 the proof is terminated.

Theorem 5.3: The necessary and sufficient condition for $MMSE(\hat{\beta}_{qk}) - MMSE(\hat{\beta}_k) > 0$ is $b_3'(\sigma^2(q^2-1)G_k + b_2 b_2')^{-1} b_3 < 1$.

Proof: $Var(\hat{\beta}_{qk}) - Var(\hat{\beta}_k) > 0$ is indicated in the proof of Theorem 5.2 so form Lemma 5.2 the theorem is proved.

5.2.3. The Comparison of Stein Type Estimator and Ridge-2 Estimator

As $\hat{\beta}_s = s\hat{\beta}$ shrinks OLS estimator, Ridge-2 estimator changes the Ridge-1 estimator with a q factor. Because of the similarity in their structure we can compare

these two estimator. Let $b_2 = Bias(\hat{\beta}_{qk})$ and $b_4 = Bias(\hat{\beta}_s)$. For fixed values of q and s the theorems are given as.

Theorem 5.4:

a) Let $k > k_{1j}$. If $b_2'(s^2C^{-1} - q^2G_k)^{-1}b_2 < \sigma^2$ then

$MMSE(\hat{\beta}_s) - MMSE(\hat{\beta}_{qk}) > 0$

b) Let $k < k_{1j}$. If $b_4'(s^2C^{-1} - q^2G_k)^{-1}b_4 < \sigma^2$ then

$MMSE(\hat{\beta}_{qk}) - MMSE(\hat{\beta}_s) > 0$

$k_{1j} = \frac{\lambda_j(q-s)}{s}$ $j = 1, .., p$.

Proof: For the part (a)

$$Var(\hat{\beta}_s) - Var(\hat{\beta}_{qk}) = s^2\sigma^2C^{-1} - q^2\sigma^2G_k$$

$$= \sigma^2(P\Lambda P' + kI)^{-1}\left[Pdiag\left(s^2\frac{(\lambda_j+k)^2}{\lambda_j} - q^2\lambda_j\right)_{j=1}^{p} P'\right](P\Lambda P' + kI)^{-1}$$

If $s^2k^2 + 2s^2\lambda_j k + \lambda_j{}^2(s^2 - q^2)$ is pd then $s^2\sigma^2C^{-1} - q^2\sigma^2G_k$ i is pd. For $0 < s < 1$ and $q > 1$, this condition is ensured in the case of $k > k_{1j}$. In the part (b) of the theorem $q^2\sigma^2G_k - s^2\sigma^2C^{-1}$ is p.d when $k < k_{1j}$. The proof is completed from Lemma 5.1.

Theorem 5.5:

a) Let $k > k_{1j}$. The necessary and sufficient condition for $MMSE(\hat{\beta}_s) - MMSE(\hat{\beta}_{qk}) > 0$ is $b_2'[\sigma^2(s^2C^{-1} - q^2G_k) + b_4b_4']^{-1}b_2 < 1$.

b) Let $k < k_{1j}$. The necessary and sufficient condition for $MMSE(\hat{\beta}_{qk}) - MMSE(\hat{\beta}_s) > 0$ is $b_4'[\sigma^2(s^2C^{-1} - q^2G_k) + b_2b_2']^{-1}b_4 < 1$.

Proof: In the proof of Theorem (5.4) and for the part (a), it is indicated that $Var(\hat{\beta}_s) - Var(\hat{\beta}_{qk})$ is pd and for the part (b) it is indicated that $Var(\hat{\beta}_{qk}) - Var(\hat{\beta}_s)$ is pd. Thus the theorem is proved from Lemma 5.2.

5.2.4. The Comparison of Two Different Ridge-2 Estimator

Liski (1982) compares two estimators $\hat{\beta}_{\lambda_1} = \lambda_1\hat{\beta}$ and $\hat{\beta}_{\lambda_2} = \lambda_2\hat{\beta}$ when $0 < \lambda_2 < \lambda_1$. According to this idea, two Ridge-2 estimator with two different estimate of q can be compared. Let $Bias(\hat{\beta}_{\mathrm{q_1k}}) = \mathrm{b}_5$ and $Bias(\hat{\beta}_{\mathrm{q_2k}}) = \mathrm{b}_6$.

Theorem 5.6:

a) Let $1 < q_2 < \mathrm{q}_1$. If $b_6' G_k^{-1} b_6 < \sigma^2(\mathrm{q_1}^2 - \mathrm{q_2}^2)$ then $MMSE(\hat{\beta}_{\mathrm{q_1k}}) - MMSE(\hat{\beta}_{\mathrm{q_2k}}) > 0$.

b) Let $1 < q_1 < \mathrm{q}_2$. If $b_5' G_k^{-1} b_5 < \sigma^2(\mathrm{q_2}^2 - \mathrm{q_1}^2)$ then $MMSE(\hat{\beta}_{\mathrm{q_2k}}) - MMSE(\hat{\beta}_{\mathrm{q_1k}}) > 0$.

Proof: In the part (a) of the theorem $Var(\hat{\beta}_{\mathrm{q_1k}}) - Var(\hat{\beta}_{\mathrm{q_2k}}) = \sigma^2(\mathrm{q_1}^2 - \mathrm{q_2}^2)G_k$ is pd for $1 < q_2 < \mathrm{q}_1$. In the part (b) of the theorem $Var(\hat{\beta}_{\mathrm{q_2k}}) - Var(\hat{\beta}_{\mathrm{q_1k}}) = \sigma^2(\mathrm{q_2}^2 - \mathrm{q_1}^2)G_k$ is pd for $1 < q_1 < \mathrm{q}_2$. The proof is completed from Lemma 5.1.

Theorem 5.7:

a) Let $1 < q_2 < \mathrm{q}_1$. The necessary and sufficient condition for $MMSE(\hat{\beta}_{\mathrm{q_1k}}) - MMSE(\hat{\beta}_{\mathrm{q_2k}}) > 0$ is $b_6'[\sigma^2(\mathrm{q_1}^2 - \mathrm{q_2}^2)G_k + \mathrm{b}_5 b_5']^{-1} b_6 < 1$.

b) Let $1 < q_1 < \mathrm{q}_2$. The necessary and sufficient condition for $MMSE(\hat{\beta}_{\mathrm{q_2k}}) - MMSE(\hat{\beta}_{\mathrm{q_1k}}) > 0$ is $b_5'[\sigma^2(\mathrm{q_1}^2 - \mathrm{q_2}^2)G_k + \mathrm{b}_6 b_6']^{-1} b_5 < 1$.

Proof: In the part (a) and part (b) of the theorem it is showed that $Var(\hat{\beta}_{\mathrm{q_1k}}) - Var(\hat{\beta}_{\mathrm{q_2k}}) > 0$ and $Var(\hat{\beta}_{\mathrm{q_2k}}) - Var(\hat{\beta}_{\mathrm{q_1k}})$ is pd respectively. Therefore from Lemma 5.2 the theorem is proved.

5.3. A numerical Example for Two Parameter Ridge Estimator

The theoretical results are examined on a data set called Portland Cement (Woods, Steinou and Starke, 1932). For the canonical model $= Z\alpha + \varepsilon$, the estimators are

$$\hat{\alpha} = \Lambda^{-1}Z'y \quad (5.42.)$$

$$\hat{\alpha}_k = (\Lambda + kI)^{-1}Z'y \quad (5.43.)$$

$$\hat{\alpha}_{qk} = q(\Lambda + kI)^{-1}Z'y \quad (5.44.)$$

$$\hat{\alpha}_s = \Lambda^{-1}Z'y\ . \quad (5.45.)$$

The relationship between the estimators from the original model and canonical model is $\hat{\beta} = P\hat{\alpha}, \hat{\beta}_{qk} = P\hat{\alpha}_{qk}$, $\hat{\beta}_s = P\hat{\alpha}_s$ $(\alpha = P'\beta)$ and the MSE values are the same. The following results are obtained by examining the data set.

a) The eigenvalue of $X'X$: 2.2357, 1.5761, 0.1866, 0.0016

b) OLS estimate of α : $(\hat{\alpha})' = (\ 0.6570,\ -0.0083,\ \ 0.3028,\ 0.3880)$

c) OLS estimate of σ^2 : $\hat{\sigma}^2 = 0.00196$

Substituting the $\hat{\alpha}$ and $\hat{\sigma}^2$, values of the estimates for (5.43.)-(5.45.) and MSE are obtained. Numerical results are given in Table 5.1.-5.3. These results are inferred from Table 5.1. - 5.3.

(1) Let $q = 1{,}2285$. The values of $\mathrm{k_{1j}}$, obtained by using Theorem 1, are $\mathrm{k_{1j}} = (0.0003, 0.0426, 0.3600, 0.5107)$. From Table 1, for $q = 1.2285$ and $k = 0.5117$, $\mathrm{MSE}(\hat{\alpha}_{\mathrm{qk}}) = 0.19386$ and it is smaller than $MSE(\hat{\alpha}) = 1.2186$. (2) For $k = 0.01$ and $q = 1.0053$, $\mathrm{MSE}(\hat{\alpha}_{\mathrm{qk}}) = 0.1469$ and $\mathrm{MSE}(\hat{\alpha}_{\mathrm{k}}) = 0.14677$ so $\mathrm{MSE}(\hat{\alpha}_{\mathrm{k}})$ is smaller than $\mathrm{MSE}(\hat{\alpha}_{\mathrm{qk}})$ as it is indicated in Theorem 2. Examine the values in the Table 2. The values of $\mathrm{MSE}(\hat{\alpha}_{\mathrm{k}})$ is smaller than $\mathrm{MSE}(\hat{\alpha}_{\mathrm{qk}})$ for $k < 0.017$ and the values of $\mathrm{MSE}(\hat{\alpha}_{\mathrm{qk}})$ is smaller than $\mathrm{MSE}(\hat{\alpha}_{\mathrm{k}})$ for $k > 0.017$.

(3) Let $s = 0.5$ and $q = 1.2285$. Using Theorem 4, $\mathrm{k_{1j}}$ values are $\mathrm{k_{1j}} = (0.00237, 0.27186, 2.29619, 3.25723)$. As it is indicated in Theorem 4(a) $(k > k_{1j})$, choose $k = 3.2582$ from Table 3. The MSE value of $\hat{\alpha}_{\mathrm{qk}}$, $\mathrm{MSE}(\hat{\alpha}_{\mathrm{qk}}) = 0.33869$, is smaller than the MSE value of $\hat{\alpha}_{\mathrm{s}}$, $\mathrm{MSE}(\hat{\alpha}_{\mathrm{s}}) = 0.4731$. As it is indicated in Theorem 4(b) $(k < k_{1j})$ choose $k = 0.00137$ from Table 3. The MSE value of $\hat{\alpha}_{\mathrm{s}}$, $\mathrm{MSE}(\hat{\alpha}_{\mathrm{s}}) = 0.4731$, is smaller than the MSE value of $\hat{\alpha}_{\mathrm{qk}}$, $\mathrm{MSE}(\hat{\alpha}_{\mathrm{qk}}) = 0.59923$.

(4) As it is indicated in Theorem 6(a) choose $1 < q_2 = 1.0053 < q_1 = 1.2285$ from Table 2. Thus, $\mathrm{MSE}(\hat{\alpha}_{q_2 k}) = 0.1469$ is smaller than $\mathrm{MSE}(\hat{\alpha}_{q_1 k}) = 0.19321$. As it is indicated in Theorem 6(b) choose $1 < q_1 = 1.2285 < q_2 = 1.8992$ from Table 2. Thus, $\mathrm{MSE}(\hat{\alpha}_{q_1 k}) = 0.19321$ is smaller than $\mathrm{MSE}(\hat{\alpha}_{q_2 k}) = 0.2165$.

Table 5.1. $\mathrm{MSE}(\hat{\alpha}_{qk})$ values for different k and q values

q	k	$\mathrm{MSE}(\hat{\alpha}_{qk})$
1.0000	0.0010	0.49627
1.0027	0.0070	0.15402
1.0252	0.0572	0.15595
1.1167	0.2620	0.17913
1.2285	0.5117	0.19386
1.4518	1.0111	0.20715
1.6754	1.5110	0.21319
1.8992	2.0112	0.21662
2.1230	2.5116	0.21883
2.3468	3.0121	0.22036
2.5707	3.5126	0.22149
2.7946	4.0131	0.22236
3.0185	4.5137	0.22305
3.2424	5.0143	0.22361

Table 5.2. $MSE(\hat{\alpha}_{qk})$ and $MSE(\hat{\alpha}_k)$ values for different k and q values

k	q	$MSE(\hat{\alpha}_{qk})$	$MSE(\hat{\alpha}_k)$
0.001	1.0006	0.49679	0.49627
0.005	1.0027	0.17069	0.17040
0.010	1.0053	0.14690	0.14677
0.015	1.0078	0.14574	0.14570
0.016	1.0084	0.14594	0.14592
0.017	1.0089	0.14619	0.14619
0.018	1.0094	0.14647	0.14648
0.025	1.0129	0.14868	0.14881
0.05	1.0252	0.15460	0.15528
0.1	1.0487	0.16212	0.16441
0.5	1.2285	0.19321	0.21472
1	1.4518	0.20687	0.25757
1.5	1.6754	0.21303	0.29312
2	1.8992	0.21650	0.32383
2.5	2.1230	0.21874	0.35049
3	2.3468	0.22029	0.37373
3.5	2.5707	0.22143	0.39408
4	2.7946	0.22231	0.41199
4.5	3.0185	0.22300	0.42785
5	3.2424	0.22356	0.44196

Table 5.3. $s = 0.5$, $MSE(\hat{\alpha}_{qk})$ and $MSE(\hat{\alpha}_k)$ values for different k and q values

q	$k > k_{1j}$	$MSE(\hat{\alpha}_{qk})$	$k < k_{1j}$	$MSE(\hat{\alpha}_{qk})$
1	2.2367	0.33691	0.00062	0.65364
1.0027	2.2487	0.33694	0.00063	0.65228
1.0053	2.2603	0.33697	0.00064	0.65099
1.2285	3.2582	0.33869	0.00137	0.59923
1.4518	4.2569	0.33964	0.00209	0.63686
1.6754	5.2568	0.34025	0.00282	0.74417
1.8992	6.2572	0.34067	0.00354	0.91309
2.1230	7.2579	0.34098	0.00427	1.13960
2.3468	8.2588	0.34121	0.00499	1.14210
2.5707	9.2599	0.34140	0.00573	1.75710
2.7946	1.0261	0.34155	0.00645	2.14570
3.0185	1.1262	0.34167	0.00718	2.58650
3.2424	1.2263	0.34177	0.00791	3.07870

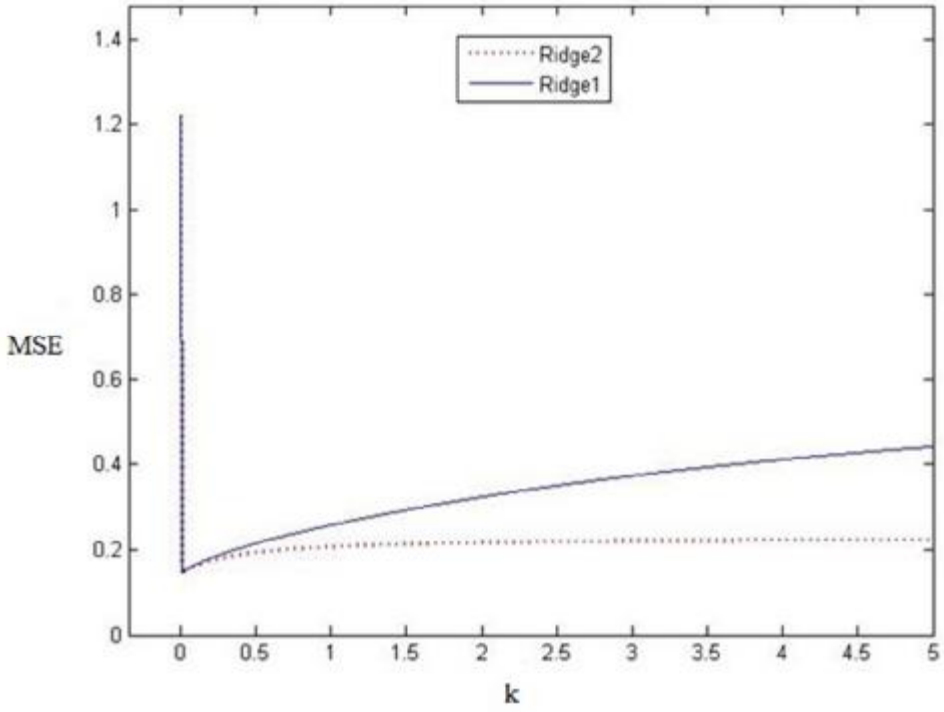

Figure 5.1. MSE values for Ridge-1 and Ridge-2

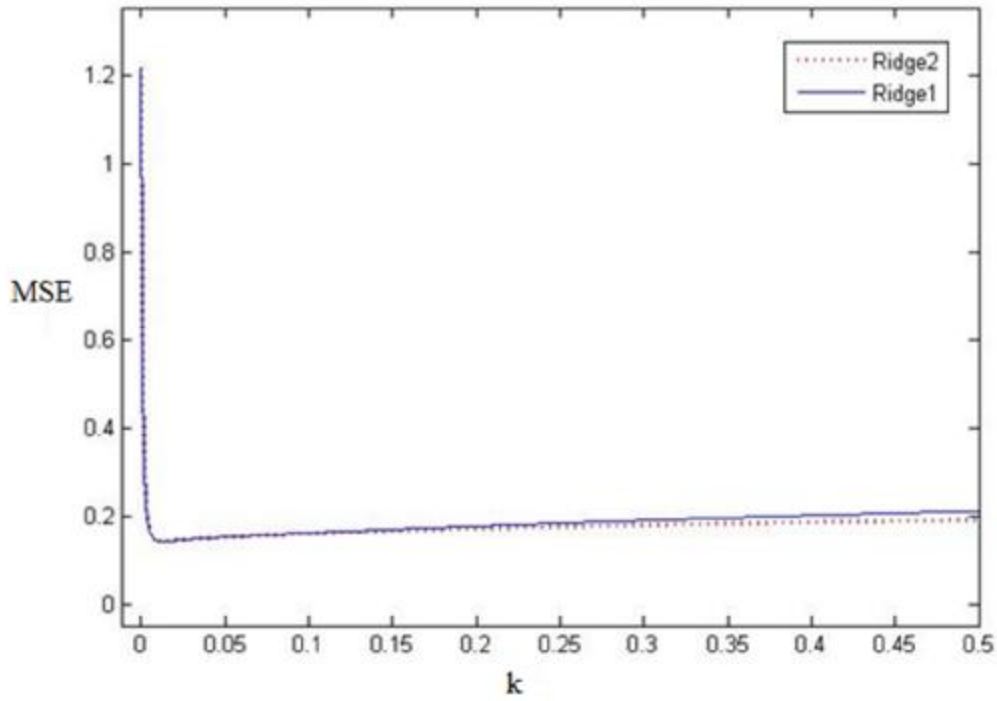

Figure 5.2. MSE values for Ridge-1 and Ridge-2

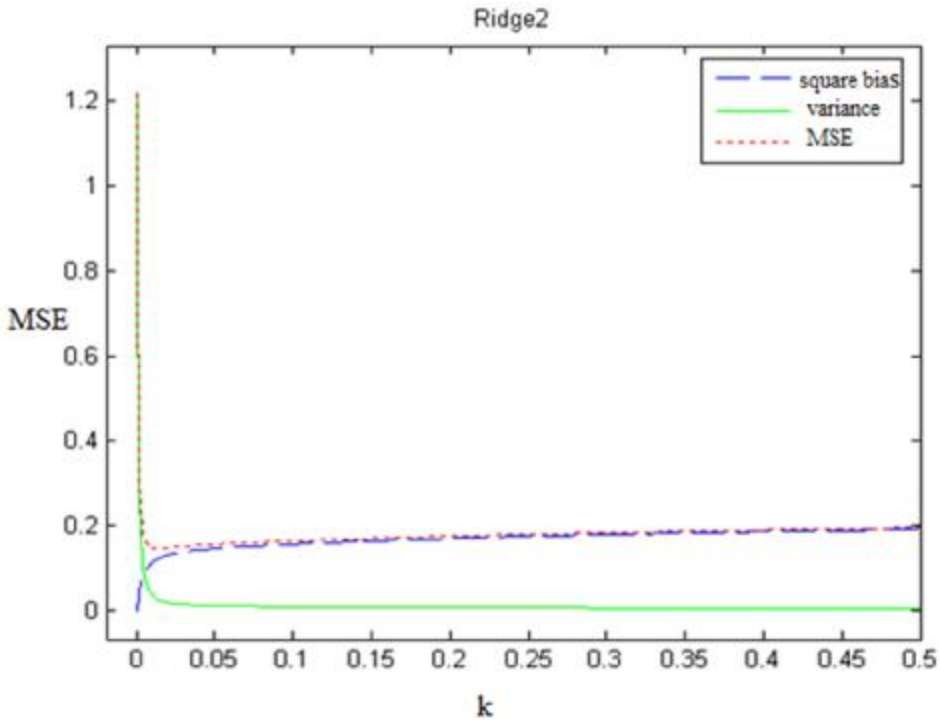

Figure 5.3. MSE, variance and squared bias for Ridge-2

6. CONCLUSIONS

Multicollinearity often produces poor results. The parameter estimates fluctuate wildly with a negligible change in the sample, signs are opposite to the signs of pair correlations and theoretically important variables with insignificant coefficients are produced in the presence of multicollinearity.

Biased regression is an alternative to the OLS regression especially when explanatory variables are highly correlated. One of the biased regression methods is ridge regression. In this study the RE and related estimators, i.e. generalized, modified, r-k class, almost unbiased, unbiased and restricted ridge estimators are examined. Besides this, minimax, bayes, continuum, robust ridge estimators and ridge estimators in SURE model are examined too.

One parameter, ordinary RE is the most widely used technique for overcoming the deficiencies of multicollinearity. However in comparison with OLS regression ordinary RE is a worse quality of fit, doesn't satisfy the relations of orthogonality, so interpretation of its result is more difficult. Therefore, Lipovetsky and Conklin (2005) proposed two parameter ridge estimator which corresponds to a regularized objective that produces a consistent solution not prone to multicollinearity.

In this study, the performance of the two parameter ridge estimator over OLS estimator, ordinary RE and Stein type estimator is compared. The superiority of the two parameter ridge estimator is examined in the mean square error matrix sense. Necessary and sufficient conditions or only necessary conditions for superiority of the estimators over each other are given by the theorems.

A numerical example is done to illustrate the findings. Widely analyzed dataset on Portland cement is used as an example. Furthermore, simulation studies can be done to gauge the performance of the estimators to be compared. Thus, the results which are obtained with a numerical example can be examined by a stronger and trustworthy way.

REFERENCES

ALAM, K. and HAWKES, J. S., 1978. Estimation of Regression Coefficients. Scandinavian Journal of Statistics, 5: 169-172.

ALLDREDGE, J. R. and GILB, N. S., 1976. Ridge Regression: An Annotated Bibliography. International Statistical Review, 44: 355-360.

ASKIN, R. G. and MONTGOMERY, D. C., 1980. Augmented Robust Estimators. Technometrics, 22: 333-341.

BACON, R. W. and HAUSMAN, J. A., 1974. The Relationship Between Ridge Regression and the Minimum Mean Square Error Estimator of Chipman. Oxford Bulletin of Economics and Statistics, 36: 115-124.

BAKSALARY, J. K., and TRENKLER, G., 1991. Nonnegative and Positive Definiteness of Matrices Modified by Two Matrices of Rank One. Linear Algebra Appl., 151: 169-184.

BAKSALARY, J. K., PORDZIK, P. R. and TRENKLER, G.,1990. A Note on Generalized Ridge Estimators. Communication in Statistics - Theory and Methods, 19: 2871-2877.

BALDWIN, K. F. and HOERL, A. E., 1978. Bounds on Minimum Mean Squared Error in Ridge Regression. Communications in Statistics - Theory and Methods, A7(13): 1209-1218.

BANERJEE, K. S. and CARR, R. N., 1971. A Comment on Ridge Regression, Biased Estimation for Nonorthogonal Problems. Technometrics, 13: 895-898.

BAYE, M. R. and PARKER, D. F., 1984. Combining Ridge and Principal Component Regression – A Money Demand Illustration. Communication in Statistics - Theory and Methods, 13: 197-205.

BELSEY, D. A., KUH, E. and WELSCH, R. E., 1980. Regression Diagnostics. John Wiley, New York.

BIBY, J., 1972. Minimum Mean Square Error Estimation, Ridge Regression, and Some Unanswered Questions. Colloquia Mathematica Societatis Janos Bolyai, 9: 107-121.

BROWN, K. G., 1978. On Ridge Estimation in Rank Deficient Models. Communications in Statistics - Theory and Methods, A7(2): 187-192.

BROWN, P. J. and ZIDEK, J. V., 1980. Adaptive Multivariable Ridge Regression the Annals of Statistics, 8: 64-74.

CASELLA, G., 1980. Minimax Ridge Regression Estimation. Annals of Statistics, 8: 1036-1056.

_____, 1985. Condition Numbers and Minimax Ridge Regression Estimators. Journal of The American Statistical Association, 80: 753-758.

CHALTON, D. O. and TROSKIE, C. G., 1992. Identification of Outlying and Influential Data with Biased Estimation: A Simulation Study. Communications in Statistics - Simulation and Computation, 21(2): 607-626.

CHAWLA, J. S., 1988. A Note On General Ridge Estimator. Communication in Statistics - Theory and Methods, 17: 739-744.

_____,1990. A Note On Ridge Regression. Statistics and Probability Letters, 9: 343-345.

CHURCHILL, G.A., 1975. A Regression Estimation Method for Collinear Predictors. Decision Sciences, 6: 670-687.

CLARCK, A.E. and TROSKIE, C. G., 2006. Ridge Regression-A Simulation Study. Communications in Statistics - Simulation and Computation, 35: 605-619.

CONIFFE, D. and STONE, J., 1973. A Critical View of Ridge Regression. The Statistician, 22: 181-187.

COUTSORIDES, D. and TROSKIE, C. G., 1978. F and T Tests for a General Class of Estimators. South African Statistical Journal, 13: 113-119.

CROUSE, R. H., JIN, C. and HANUMARA, R. C., 1995. Unbiased Ridge Estimation with Prior Information and Ridge Trace. Communications in Statistics - Theory and Methods, 24(9): 2341-2354.

DEJONG, S. and FAREBROTHER, R. W., 1994. Extending the Relationship Between Ridge Regression and Continuum Regression. Chemometrics and Intelligent Laboratory Systems, 25: 179-181.

DEMPSTER, A. P., 1973. Alternatives to Least Squares in Multiple Regression. Multivariate Statistical Inference (D.G. Kabe and R.P. Gupta, eds.), North-Holland Publishing Company, 25-40.

DEMPSTER, A. P., SCHATZOFF, M. and WERMUTH, N., 1977. A Simulation Study of Alternatives to Ordinary Least Squares. Journal of the American Statistical Association (with discussion), 72: 77-106.

DRAPER, N. R. and SMITH, H., 1969. Methods for Selecting Variables From a Given Set of Variables for Regression Analysis. Bulletin of the International Statistical Institute, 43: 7-15.

DRAPER, N. R. and VAN NOSTRAND, R. C., 1979. Ridge Regression and James – Stein Estimation: Review and Comments. Technometrics, 21: 451-466.

DRUILHET, P. and MOM, A., 2008. Shrinkage Structure in Biased Regression. Journal of Multivariate Analysis, 2: 232-244.

ELLINGSEN, W. R., 1975. On-line Ridge Regression: Sequential Biased Estimation for Nonorthogonal Problems. Journal of Statistical Computation and Simulation, 3: 249-264.

FAREBROTHER, R. W., 1975. The Minimum Square Error Linear Estimator and Ridge Regression. Technometrics, 17: 127-128.

_____, 1976. Further Results on the Mean Square Error of Ridge Regression. Journal of the Royal Statistical Society Series B-Methodological 38: 248.

_____, 1978a. A Class of Shrinkage Estimators. Journal of the Royal Statistical Society, Series B, 40: 47-49.

_____, 1978b. Partitioned Ridge Regression. Technometrics, 20: 121-122.

_____, 1984. The Restricted Least Squares Estimator and Ridge Regression. Communications in Statistics – Theory and Methods, 13(2): 191-196.

FARRAR, D. E. and GLAUBER, R. R., 1967. Multicollinearity in Regression Analysis: The Problem Revisited. The Review of Economics and Statistics, 49(1): 92-107.

FIRINGUETTI, L., 1989. A Simulation Study of Ridge Regression Estimators with Autocorrelated Errors. Communications in Statistics - Simulation, 18(2): 673-702.

FOMBY, T. B. and JOHNSON, S. R., 1977. MSE Evaluation of Ridge Estimators Based on Stochastic Prior Information. Communications in Statistics - Theory and Methods, A6(13): 1245-1258.

FULE, E., 1995. On Ecological Regression and Ridge Estimation. Communications in Statistics. Simulation and Computation, 24(2): 385-389.

GIBBONS, D. G., 1981. A Simulation Study of Some Ridge Estimators. Journal of The American Statistical Association, 76: 131-139.

GOLDSTEIN, M. and SMITH, A. F. M., 1974. Ridge-type Estimators for Regression Analysis. Journal of the Royal Statistical Society, Series B, 36: 284-291.

GROSS, E., 2003. Restricted Ridge Estimation. Statistics and Probability Letters, 65: 57–64.

GUILKEY, D. K. and MURPHEY, J. L., 1975. Directed Ridge Regression Techniques in Cases of Multicollinearity. Journal of the American Statistical Association, 70: 769-775.

GUNST, R. F. and MASON, R. L., 1976. Generalized Mean Squared Error Properties of Regression Estimators. Communications in Statistics – Theory and Methods, A5(15): 1501-1508.

_____, 1977. Biased Estimation in Regression: An Evolution Using Mean Squared Error. Journal of the American Statistical Association, 72: 616-628.

HADZIVUKOVIC, S., NIKOLIC-DJORIC, E. and COBANOVIC, K., 1992. The Choice of Perturbation Factor in Ridge Regression. Journal of Applied Statistics, 19(2): 223-230.

HEMMERLE, W. J., 1975. An Explicit Solution for Generalized Ridge Regression. Technometrics, 17: 309-314.

HEMMERLE, W. J. and BRANTLE, T. F., 1978. An Explicit and Constrained Generalized Ridge Estimation. Technometrics 20: 109–119.

HEMMERLE, W. J. and CAREY, M. B., 1983. Some Properties of Generalized Ridge Estimators. Communication in Statistics - Simulation and Computation, 12: 239-256.

HINKLEY, D. V., 1977. Jackknife Confidence Limits Using Student-T Approximations. Biometrica 64: 21-28.

HOCKING, R. R., 1972. Criteria for Selection of a Subset Regression: Which One Should Be Used. Technometrics, 14: 967-970.

_____, 1976. The Analysis and Selection of Variables in Linear Regression. Biometrics, 32: 1-49.

HOCKING, R. R., SPEED, F. M. and LYNN, M. J., 1976. A Class of Biased Estimators in Linear Regression. Technometrics 18: 425–437.

HOERL, A.E., 1959. Optimum Solution of Many Variables Equations. Chemical Engineering Progress, 55(11): 69-78.

_____,1962. Application of Ridge Analysis to Regression Problems. Chemical Engineering Progress, 58(3): 54-59.

HOERL, R. W., 1985. Ridge Analysis 25 Years Later. American Statistician, 39: 186-192.

HOERL, A. E. and KENNARD, R. W., 1970a. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12, 1: 55-67.

_____, 1970b. Ridge Regression: Applications to Nonorthogonal Problems. Technometrics, 12, 1: 69-82.

_____, 1976. Ridge Regression: Iterative Estimation of the Biasing Parameter. Communications in Statistics - Theory and Methods, A5(1): 77-78.

_____, 1979. Ridge Regression – Present and Future. Proceedings of the Computer Science and Statistics: 12th Annual Symposium on the Interface, University of Waterloo, 223-227.

_____, 1990. Ridge Regression: Degrees of Freedown in the Analysis of Variance. Communications in Statistics – Simulation and Computation, 19(4): 1485-1495.

HOERL, A. E., KENNARD, R. W., and BALDWIN, K. F., 1975. Ridge Regression: Some Simulation. Communication in Statistics, 4(2): 105-123.

HSIANG, T. C., 1975. A Bayesian View on Ridge Regression. The Statistician, 24: 267-268.

_____, 1970b. Ridge Regression: Applications to Nonorthogonal Problems. Technometrics, 12: 69-82.

HSUAN, F. C., 1981. Ridge Regression From Principal Component Point of View of Communication in Statistics – Theory and Methods, 10: 1981-1995.

JENSEN, D. R. and RAMIREZ, D. E., 2008. Anomalies in the Foundations of Ridge Regression. International Statistical Review, 76: 89-105.

KACIRANLAR, S., SAKALLIOGLU, S. and AKDENIZ, F., 1998. Mean Squared Error Comparisons of the Modified Ridge Regression Estimator and the Restricted Ridge Regression Estimator, Communications in Statistics - Theory and Methods, 27: 131–138.

KADIYALA, K., 1979. Operational Ridge Regression Estimators Under the Prediction Goal. Communications in Statistics - Theory and Methods, A8(14): 1377-1391.

_____, 1984. A Class of Almost Unbiased and Efficient Estimators of Regression Coefficients. Economics Letters, 16: 293-296.

KASARDA, J. D. and SHIN, W. P., 1977. Optimal Bias in Ridge Regression Approaches to Multicollinearity. Sociological Methods and Resarch, 5: 461-470.

KEJIAN, L., 1993. A New Class of Biased Estimate in Linear Regression Communications in Statistics - Theory and Methods, 22(2): 393-402.

KIBRIA, G., 1996. On Preliminary Test Ridge Regression Estimators for Linear Restrictions in a Regression Model with Non-Normal Disturbances. Communications in Statistics - Theory and Methods, 25(10): 2349-2369.

KOENKER, R. W., 1982. Robust Methods in Econometrics. Econometric Rev, 1: 213-290.

KOENKER, R. W. and BASSETT, G. W., 1978. Regression Quantiles. Econometrica, 46: 33-50.

KUBOKAWA, T. and SRIVASTAVA, M. S., 2004. Improved Empirical Bayes Ridge Regression Estimator Under Multicollinearity. Communication in Statistics - Theory and Methods, 33: 1943-1973.

LAMOTTE, L. R., 1978. Bayes Linear Estimators. Technometrics, 20: 281-290.

LAWLESS, J. F., 1978. Ridge and Related Estimation Procedures: Theory and Practice. Communications in Statistics - Theory and Methods. A7(2): 139-164.

LAWLESS, J. F. and WANG, P., 1976. A Simulation Study of Ridge and Other Regression Estimators. Communications in Statistics - Theory and Methods, A5(4).

LAWRENCE, K. D. and MARSH, L. C., 1984. Robust Ridge Estimation Methods for Predicting U. S. Coal Mining Fatalities. Communications in Statistics - Theory and Methods, 13: 139-149.

LEAMER, E. E., 1975. A Result on Sign of Restricted Least Squares Estimates, Journal of Econometrics, 3: 387-390.

LECESSIE, S. and VAN HOUWELINGEN, J. C., 1992. Ridge Estimators in Logistic Regression Applied Statistics, 41(1): 191-201.

LI, Y. L. and YANG, H., 2010a. A New Liu-Type Estimator in Linear Regression Model. Stat. Papers.

_____, 2010b. Two Kinds of Restricted Modified Estimators in Linear Regression Model. Journal of Applied Statistics.

LINDLEY, D.V. and SMITH, A. F. M., 1972. Bayes Estimates for the Linear Model. Journal of the Royal Statistical Society, Series B, 34: 1-72.

LIPOVETSKY, S., 2006. Ridge Regression and Its Convergence to the Eventual Pairwise Model. Mathematical and Computer Modelling, 44: 304-318.

LIPOVETSKY, S. and CONKLIN, W. M., 2005. Ridge Regression in Two Parameter Solution. Applied Stochastic Models in Business and Industry, 21: 525–540.

LISKI, E.P., 1982. A Test of the Mean Square Error Criterion for Shrinkage Estimators. Communication in Statistics - Simulation and Computation. 11(5): 543-562.

_____, 1988. A Test of the Mean Square Error Criterion for Linear Admissible Estimators. Communications in Statistics - Theory and Methods, 17(11): 3743-3756.

LOWERRE, J. M., 1974. On the Mean Square Error of Parameter Estimates for Some Biased Estimators. Technometrics, 16: 461-464.

MARKIEWITCZ, A., 1996. Caracterization of General Ridge Estimators. Statistics and Probability Letters, 27: 145-148.

MARUYAMA, Y. and STRAWDERMAN, W.E., 2005. A New Class of Genaralized Bayes Minimax Ridge Regression Estimators. Annals of Statistics, 33: 1753-1770.

MARQUARDT, D. W., 1970. Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. Technometrics, 12: 591-612.

MAYER, L.W. and WILKIE, T.A., 1973. On Biased Estimation in Linear Models. Technometrics, 15: 497-508.

MILLER, A. J., 1990. Subset Selection in Regression. Chapman and Hall. New York.

_____, 1974a. The Jackknife: a Review. Biometrika 61: 1-15.

_____, 1974b. An Unbalanced Jackknife. Ann. Statist., 2: 880-891.

MITRA, A. and LING, R. F., 1979. A Monte Carlo Comparison of some Ridge and Other Biased Estimators. Journal of Statistical Computation and Simulation, 9: 195 -215.

MOULAERT, F., 1979. A Conservative Ridge Estimator. Economics Letters, 3: 159-164.

MULLET, G. M., 1976. Why Regression Coefficients Have the Wrong Sign, Journal of Quality Technology, 8: 121-126.

NEBEBE, F. and SIM, A.B., 1990. A Relative Performance of Improved Ridge Estimators and An Empirical Bayes Estimator - Some Monte Carlo Results. Communication in Statistics - Theory and Methods, 19: 3469-3495.

NELDER, J. A., 1972. Discussion of "Bayes Estimates for the Linear Model," by Lindley and Smith, Journal of the Royal Statistical Society, Ser. B,34: 18-20.

NOMURA, M., 1988. On The Almost Unbiased Ridge Regression Estimator. Communication in Statistics - Simulation and Computation, 17: 729-743.

NOMURA, M. and OHKUBO, T., 1985. A Note on Combining Ridge and Principal Component Regression. Communication in Statistics - Theory and Methods, 14: 2489-2493.

NOOR, I. and MEHTA, J. S., 1989. A Study of Some Shrinkage Type Estimtors. Communications in Statistics - Theory and Methods, 18(12): 4437-4457.

NYQUIST, H., 1988. Applications of the Jackknife Procedure in Ridge Regression. Computational Statistics and Data Analysis, 177-183.

OBENCHAIN, R. L., 1975. Ridge Analysis Following a Preliminary Test of the Shrunken Hypothesis. Technometrics, 17: 431-435.

_____, 1977. Classical F-Tests and Confidence Regions for Ridge Regression. Technometrics, 19: 429-439.

OZKALE, M. R. and KACIRANLAR, S., 2007. The Restricted and Unrestricted Two - Parameter Estimators. Communication in Statistics - Theory and Methods, 36**:** 2707-2725.

PEDDADA, S. D., NIGAM, A. K. and SAXENA, A. K., 1989. On the Inadmissibility of Ridge Estimator in a Linear Model. Communications in Statistics - Theory and Methods, 18(10): 3571-3585.

PEELE, L. and RYAN, T. P., 1982. Minimax Lineer Regression Estimators with Application To Ridge Regression. Technometrics, 24: 157-159.

PFAFFENBERGER, R. C. and DIELMAN, T. E., 1984. A Modified Ridge Regression Estimator Using the Least Absolute Value Criterion in the Multiple Linear Regression Model. Proceedings of the American Institute for Decision Sciens, Toronto, 791-793.

_____, 1985. A Comparison of Robust Ridge Estimators. Proceedings of the American Statistical Association Business and Economic Statistics Section, LasVegas, Nev., 631-635.

PLISKIN, J. L., 1987. A Ridge-Type Estimator and Good Prior Means. Communications in Statistics - Theory and Methods, 16, 12: 3429-3437.

PRECHT, M. and RAO, P. V. P., 1985. An Evaluation of Biased Estimators of Regression Coefficients. A Simulation Study. Statistiche Hefte, 26: 263-285.

PRICE, J. M. 1982. Comparisons Among Regression Estimators Under the Generalized Mean Square Error Criterion. Communications in Statistics A11(17): 1965-1984.

PUKELSHEIM, F., 1977. Equality of Two Blues and Ridge - Type Estimates. Communications in Statistics - Theory and Methods, A6 (7): 603-610.

QANNARI, E.M. and HANAFI, M., 2005. A Simple Continuum Regression Approach. Journal of Chemometrics**,** 19: 387-392.

OUENOUILLE, M. H., 1956. Notes on Bias in Estimation. Biometrika, 43: 353-360.

RAO, C.R., 1976. The 1975 Wald Memorial Lectures: Estimation of Parameters in a Linear Model. The Annals of Statistics, 4: 1023-1037.

_____, 1974. Linear Statistical Inference and Its Applications.

SAKALLIOGLU, S. and KACIRANLAR, S., 2008. A New Biased Estimator Based on the Ridge Estimation. Stat. Pap., 49: 669–689.

SARKAR, N., 1989. Comparisons Among some Estimators in Misspecified Linear Models with Multicollinearity, Ann. Inst. Statist. Math. 41: 717-724.

_____, 1992. A New Estimator Combining The Ridge Regression and The Restricted Least Squares Methods of Estimation. Communication in Statistics - Theory and Methods, 21:1987–2000.

_____, 1996. Mean Square Error Matrix Comparison of Some Estimators in Linear Regressions with Multicollinearity. Statistics and Probability Letters, 30: 133-138.

SCHIPP, B. and TOUTENBURG, H., 1996. Feasible Minimax Estimators in the Simultaneous Equations Model Under Partial Restrictions. Journal of Statistical Planning and Inference, 50: 241-250.

SIHOTA, S. S. and BANARJEE, K. S., 1974. Biased Estimation in Weighing Designs. Sankhya, 36: 55-66.

SILVAPULLE, M. J., 1991. Robust Ridge Regression Based on an M-Estimator. Australian Journal of Statistics, 33(3): 319-333.

SINGH, B., CHAUBEY, Y. P. and DWIVEDI, T. D., 1986. An Almost Unbiased Ridge Estimator**.** Sankhya - The Indian Journal of Statistics, 48: 342-346.

SINGH, R. K., PANDEY, S. K. and SRIVASTAVA, V. K., 1994. A Generalized Class of Shrinkage Estimators in Linear Regression When Disturbances are not Normal. Communications in Statistics - Theory and Methods, 23(7): 2029-2046.

SMITH, V. K., 1976. A Note on Ridge Regression. Decision Sciences, 7: 562-566.

SMITH, A. F. M. and GOLDSTEIN, M., 1975. Ridge Regression: Some Comments on a Paper of Conniffe and Stone. The Statistician, 24: 61-66.

SMITH, G. and CAMPBELL, F., 1980. A Critique of Some Ridge Regression Methods. Journal of The American Statistical Association, 75: 74-81.

SNEE, R. D., 1977. Validation of Regression Models: Methods and Examples. Technometrics, 19: 415-428.

SRIVASTAVA, V. K. and GILES D. E. A.,1987. Seemingly Unrelates Regression Equations Models. Estimation and Inference, Markel Dekker Inc., New York, 80.

_____, 1991. Unbiased Estimation of the Mean Squared Error of the Feasible Generalised Ridge Regression Estimator. Communications in Statistics - Theory and Methods, 20(8): 2375-2386.

SRIVASTAVA, M. S. and KUBOKAWA, T., 2005. Journal of Multivariate Analysis, 93: 394-416.

STONE, M. and BROOKS, R. J., 1990. Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Square and Principal Components Regression (with discussion). J. R. Statist. Soc. B, 52: 237-269.

STRAWDERMAN, W. E., 1978. Minimax Adaptive Generalized Ridge Regression Estimators. Journal of the American Statistical Association, 73: 623-627.

SUNDBERG, R., 1993. Continuum Regression and Ridge Regression. Journal of The Royal Statistical Society, 55: 653-659.

SWINDEL, B. F., 1974. Instability of Regression Coefficients Illustrated. American Statistician, 28: 63-65.

_____, 1976. Good Ridge Estimators Based on Prior Information. Communications in Statistics - Theory and Methods, A5(11): 1065-1075.

SWINDEL, B. F. and CHAPMAN, D. D., 1973. Good Ridge Estimators. A Paper Presented at the Annual Meeting of the Institute of Mathematical Statistics, The Biometric Society and The American Statistical Association. Abstract in Biometrics, 30: 385-386.

SWAMY, P. A. V. B., MEHTA, J. S. and RAPPOPORT, J. N., 1978. Two Methods of Evaluating Hoerl and Kennard's Ridge Regression. Communications in Statistics - Theory and Methods, A7(12): 1133-1155.

THEIL, H., 1971. Principles of Econometrics. Wiley, New York.

_____, 1963. On the Use of Incomplete Prior Information in Regression Analysis, J. Am. Stat. Assoc. 58: 401–414.

THEOBALD, C. M., 1974. Generalizations of Mean Square Error Applied to Ridge Regression. Journal of the Royal Statistical Society, Series B, 36: 103-105.

TOUTENBURG, H., 1982. Prior Information in Linear Models. John Wiley and Sons, New York.

TRENKLER, G., 1978. An Iteration Estimator for the Linear Model. Compstat 1978, Leiden University, Wien: Physica–Werlag, 125-131.

_____, 1980. Generalized Mean Squared Error Comparisons of Biased Regression Estimators. Communications Statistics - Theory and Methods, A9(12): 1247-1259.

TRENKLER, G. and TOUTENBERG, H., 1990. Mean Squared Error Matrix Comparisons Between Biased Estimators. Statistical Papers, 31: 165-17.

TUKEY, J. W., 1958. Bias and Confidence in not Quite Large Samples (Abstract). Ann. Math. Statist. 29: 614.

ULLAH, A., SRIVASTAVA, V. K. and CHANDRA, R., 1983. Properties of Shrinkage Estimators in Linear Regression When Disturbances are not Normal. Journal of Econometrics, 21: 389-402.

VINOD, H. D., 1976b. Simulation and Extension of a Minimum Mean Square Error Estimator in Comparison with Stein's. Technometrics, 18: 491-496.

_____, 1978a. A Survey of Ridge Regression and Related Techniques for Improvement Over Ordinary Least Squares. The Review of Economics and Statistics, LX, 121-131.

_____, 1978b. Equivariance of Ridge Estimators Through Standardization, A Note. Communications in Statistics – Theory and Methods, A7: 1157-1161.

VINOD, H.D. and ULLAH, A., 1981. Recent Advances in Regression Methods. Marcel Dekker, Inc., New York and Basel, 361.

VINOD, H. D., ULLAH, A. and KADIYALA, K., 1981. Evaluation Mean Squared Error of Certain Ridge Estimators Using Confluent Hypergeometric Function. Sankya: The Indian Journal of Statistics: Series B, 43(3): 360-383.

VISCO, I., 1978. On Obtaining the Ridge Sing of a Coefficient Estimate by Omitting a Variable from the Regression. Journal of Econometrics, 6: 115- 118.

WICHERN, D. W. and CHURCHILL, G. A., 1978. A Comparison of Ridge Estimators. Technometrics, 20: 301-311.

WOODS, H., STEINOUR, H. H., and STARKE, H. R., 1932. Effect of Composition of Portland Cement on Heat Evolved During Hardening. Industrial and Engineering Chemistry, 24: 1207-1214.

ZELLNER, A., 1962. An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. Journal of the American Statistical Association. 57: 348-368.

CURRICULUM VITAE

I was born in Adana in 1985. I graduated from Adana Anatolian High School in 2003. Afterwards, I started my undergraduate education at Çukurova University, Faculty of Science and Letters in 2004. I graduated from the Department of Mathematics and the Department of Statistics with a double major programme in 2008. At the same year I started to the Master of Science programme in the Department of Statistics. I have been working as a research assistant in the Department of Statistics since 2008.