University of Cape Town
Department of Statistical Sciences

Comparison of Ridge and other shrinkage estimation techniques

by
Bokang C. Vumbukani

A thesis prepared under supervision of
Prof. Christien Thiart

in fulfilment of the requirements for the degree of
Master of Science in Statistical Sciences

May 2006

Acknowledgements

****

I am thankful to my supervisor, Prof Christien Thiart for guidance, brilliant ideas, and dedication to the work throughout the research period. I express my appreciation to Prof Troskie for valuable suggestions and interest in the work.

My gratitude to

- Prof Theo Stewart for guidance through Latex,
- Dominique for comments and assistance with R and
- Gerard for an incredible support and encouragement.

I am grateful to Lesotho Government and NRF for financial support.

****

Abstract

Shrinkage estimation is an increasingly popular class of biased parameter estimation techniques, vital when the columns of the matrix of independent variables X exhibit dependencies or near dependencies. These dependencies often lead to serious problems in least squares estimation; inflated variances and mean squared errors of estimates, unstable coefficients, imprecision and improper estimation. Shrinkage methods allow for a little bias and at the same time introduce smaller mean squared error and variances for the biased estimators, compared to those of unbiased estimators. However, shrinkage methods are based on the shrinkage factor, of which estimation depends on the unknown values, often computed from the OLS solution. We argue that the instability of OLS estimates may have an adverse effect on performance of shrinkage estimators.

Hence, a new method for estimating the shrinkage factors is proposed and applied on ridge and generalized ridge regression. We propose that the new shrinkage factors should be based on the principal components instead of the unstable OLS estimates. We use the total mean squared errors of estimates to compare efficiencies of the ridge and generalized ridge estimators associated with the new method to the well known estimators, namely, the Stein estimator (James and Stein, 1961), ridge estimators (Hoerl et al., 1975; Lawless and Wang, 1976; Brown, 1993; Kibria, 2003), generalized ridge estimators (Hoerl and Kennard, 1970a; Troskie and Chalton, 1996), Liu estimators (Liu, 1993), the generalized Liu estimators (Liu,1993) and principal component estimators deleting two smallest roots (Kendall, 1957). The goal is to try to find the most efficient estimator and to determine whether or not the estimators associated with the proposed procedure are better than the existing estimators.

The principal components estimator deleting the smallest root shows an outstanding superiority over the rest of the shrinkage estimators. Further, the new estimators based on the principal components estimator deleting one root are superior and an improvement over the existing biased estimators.

Contents

List of Tables

Acronyms

ANOVA	Analysis of variance
BLUE	Best Linear Unbiased Estimator
df	Degrees of freedom
mci	Multicollinearity index
MSE	Mean Squared Error
OLS	Ordinary Least Squares
OLSE	The Ordinary Least Squares Estimator
PC	Principal Components
PCdel1	Principal Component Regression with one root deleted
PCdel2	Principal Component Regression with two roots deleted
RE	Relative Efficiency
SSE	Sum of Squared Error
SSR	Sum of Squared Residuals
SVD	Singular Value Decomposition
TMSE	Total Mean Squared Error
VIF	Variance Inflation Factor

Notation	**Description**
c	The Stein shrinkage factor
C	The condition number
$Cov[.,.]$	Covariance of variables
C_p	Mallows C_p
d	The Liu constant
df	Degrees of freedom
d_{sh}	The general shrinkage factor
$E[.]$	Expectation of a scalar, vector or matrix
$f_x(x)$	Density function of a random variable X
F	F- statistic
H	Hat matrix
H_0	Null hypothesis
H_1	Alternative hypothesis
k	The ridge constant
n	Number of data observations
p	The number of variables in the model
r	Coefficient of correlation

$r(X)$	Rank of matrix X
R^2	Coefficient of Determination
R_i^2	Coefficient of Determination when X_i is regressed on other independent variables
U	The left singular vectors of X
u_i	The i^{th} column of U
u_{ji}	The j^{th} row and i^{th} column element of U
V	The right singular vectors of X
$Var[.]$	Variance of a scalar, vector or matrix
v_i	The i^{th} column of V
v_{ji}	The j^{th} row and i^{th} column element of V
X	A matrix of independent variables
X'	Transpose of matrix X
X^{-1}	Inverse of matrix X
X_i	The i^{th} column of X
x_{ij}	The i^{th} row and j^{th} column entry of matrix X
$\|\| x \|\|$	Norm of vector x
$\sqrt{x'x}$	Length of vector x
Y	A vector of response observations
$\hat{Y}$	A vector of predicted responses
$\hat{Y}_i$	The i^{th} predicted response
Z	An $n \times n$ matrix of principal components
Z_a	An $n \times (p-m)$ matrix of principal components corresponding the largest eigenvalues.
Z_b	An $n \times m$ matrix of principal components corresponding to the smallest eigenvalues.
β	A vector of true coefficients
β_T	[10 0.4 0.5 0.25 0.3 4.5]′ (True coefficients for the simulation study)
$\hat{\beta}$	The OLS estimator
$\hat{\beta}_{-i}$	A vector of OLS coefficients estimated without the i^{th} observation
$\hat{\beta}^G$	The Garotte estimator
$\hat{\beta}_{GL}$	A vector of shrinkage coefficients
$\hat{\beta}_{kd}$	The modified Liu estimator
$\hat{\beta}_L$	The Liu estimator
$\hat{\beta}_{pc}$	The principal components estimator

$\hat{\beta}_R$	The ridge estimator
$\hat{\beta}_s$	The Stein estimator
$\hat{\beta}_{sh}$	The shrinkage estimator
$\tilde{\beta}$	Any vector of coefficients
α	$V'\beta$
$\hat{\alpha}$	A vector of orthogonal least squares coefficients
$\hat{\alpha}_a$	A vector of coefficients corresponding to the retained singular values in principal components regression
$\hat{\alpha}_b$	A vector of coefficients corresponding to the eliminated singular values in principal components regression
$\hat{\alpha}_{GR}$	A vector of generalized ridge coefficients
λ_i	The i^{th} eigenvalue of $X'X$
λ_{max}	The largest eigenvalue of $X'X$
λ_p	The smallest eigenvalue of $X'X$
$\sqrt{\lambda_i}$	The i^{th} singular value of X
ε	A vector of error terms
σ_x^2	Variance of a random variable X
σ_{xi}	The standard deviation for the i^{th} independent variable
σ_ε^2	Error variance
Δ	A diagonal matrix of singular values of X

Chapter 1

Introduction

For a long time, regression analysis has been used as the main statistical technique for fitting equations to data. The technique is widely used in social, biological and physical sciences (Allison, 1999) to portray the relationship between the variable of interest (dependent) and one or more other variables (independent variables).

Regression is usually used in prediction or causal analysis to

- develop the functional form for making predictions about the response variable, based on the independent or explanatory variables and/or
- to determine whether or not the independent variables influence the dependent variable.

Through regression analysis, it is possible to combine more than one variables to produce optimal predictions of the response variable and to determine the magnitude of the unique contribution of each independent variable.

Least squares estimation is the most frequently used statistical procedure, favoured for being unbiased and producing the estimates that have minimum variance. However, sometimes least squares estimation is plagued by existence of dependencies among the independent variables and tend to be imprecise and completely unreliable. It is in such conditions when shrinkage estimation becomes a necessity.

This study strives towards identification of the most stable and reliable shrinkage estimation technique(s), required to curb the problems attributed to dependencies or near dependencies of independent variables. We view some of the biased estimation methods from a shrinkage point of view, hoping to get an insight into why they can be expected to perform well when the data are collinear.

The specific objectives of this study are the following:

- to propose a new method for estimation of the shrinkage factors and assess from the simulation study whether or not the new method improves on the traditional method of estimating shrinkage factors. We hope to observe a great improvement since the new method is based on a considerably stable procedure.
- to bring together 24 biased estimators into a common framework of shrinkage estimation. Our primary aim is to identify the most effective and robust estimator when there exists extreme collinearity among the independent variables. We consider the following estimators
    * the Stein estimator (James and Stein, 1961),
    * 14 ridge estimators (Hoerl et al., 1975; Lawless and Wang, 1976; Brown, 1993; Kibria, 2003),
    * 4 generalized ridge estimators (Hoerl and Kennard, 1970a; Troskie and Chalton, 1996),
    * 2 Liu estimators (Liu, 1993),
    * the generalized Liu estimator (Liu,1993) and
    * 2 estimators from principal components regression (Kendall, 1957)

In the remaining sections of this chapter, we review the theory behind linear regression, Least squares estimation and the major problems that may be encountered in the process. In chapter 2, the research problem is defined; collinearity is considered a broad problem from which the research problem originates. The general issues around collinearity are reviewed and the remedy for collinearity is identified as the main research problem.

Chapter 3 provides the literature review on shrinkage estimation and some of its special cases; we review the theory behind ridge and generalized ridge regression, Stein estimation, Liu and generalized Liu estimation and principal components regression. Estimation of the shrinkage factors is discussed in chapter 4; we look into what is traditionally being done and propose a new method through which shrinkage factors may be estimated. The simulation study is introduced in chapter 5; past simulation studies are reviewed and the manner in which our simulation study contributes to the global issue or the general topic of shrinkage estimation is specified. The design of the simulation and the simulation program are presented. Chapter 6 presents the simulation results. Finally, an evaluation of the objectives, the conclusions, recommendations as well as further research area are provided in chapter 7.

Full program details and some of the important definitions are provided in the appendix in the following sequence

- Appendix A → The R simulation program
- Appendix B → The general theory behind distributions considered in the study
- Appendix C → The estimators and estimation methods considered for simulation.
- Appendix D → Past simulation studies.

1.1 The linear model

The standard form of a linear regression model is denoted by the following equation:

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

where

Y = an $(n \times 1)$ vector of observed responses

n = the number of observations

X = an $(n \times p)$ full column-rank matrix of covariates, also known as a matrix of fixed array of independent numbers. If the model has a constant, it will be stated explicitly and the constant will be represented by a column of ones in the first column of X. That is $x_{i1} = 1$ for i=1, . . . n and x_{ij} = the i^{th}row element and j^{th} column element of X.

p = the number of parameters in the model.

β = a $(p \times 1)$ vector of unknown regression coefficients

ϵ = an $(n \times 1)$ a vector of uncorrelated random error terms.

The random error terms account for all the variables (measurable or otherwise) that are not included in the regression model, with the following properties.

$E[\epsilon] \;\; = 0$

$E[\epsilon\epsilon'] = \sigma^2 I$, assuming homogeneity.

1.2 The orthogonal form of the linear model

The linear model (1.1) can be reduced to an orthogonal form (canonical form) by usage of the singular value decomposition (SVD) of X. The SVD of X is defined as follows:

1.2.1 Singular Value Decomposition (SVD) of X

Let

$X_{n\times p}$ = a matrix of order $(n \times p)$ and rank r(X) = s, for $s \leq p \leq n$,

$U_{n\times n}$ and $V_{p\times p}$ = orthogonal matrices such that

$$U'_{n\times n} \quad X_{n\times p} \quad V_{p\times p} = \Delta_{n\times p} \tag{1.2}$$

where

Δ = a diagonal $n \times p$ matrix whose first s diagonal elements are square roots of eigenvalues of $X'X$, also known as the singular values of matrix X.

$$\mathbf{\Delta} = \begin{bmatrix} D_s & 0 \\ 0 & 0 \end{bmatrix}$$

$$D_s = Diag(\sqrt{\lambda}_1, \ldots, \sqrt{\lambda}_s) \quad for \quad s = r(X); \qquad \sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \ldots\ldots\sqrt{\lambda_s} > 0$$

$\sqrt{\lambda}_i$= the i^{th} singular value of X and
λ_i = the i^{th} eigenvalue of $X'X$
U = $(n \times n)$ left singular vectors of X or eigenvectors of $X'X$
V = $(p \times p)$ right singular vectors of X or eigenvectors of XX'

Both U and V are orthogonal, imply that $U'U = I_n$ and $V'V = I_p$ where $(')$ indicates transpose of the corresponding matrix. In addition, V is said to be orthonormal, meaning that $V' = V^{-1}$, where V^{-1} is the inverse of V.

Hence, in the light of the above and using (1.2), X can be expressed as

$$X_{n\times p} = U_{n\times n}\Delta_{n\times p}V'_{p\times p} \tag{1.3}$$

If the assumption is that all columns of X are independent (X is of full column rank), then

$$X_{n\times p} = U_{n\times p}\Delta_{p\times p}V'_{p\times p} = \sum_{i=1}^{p} \sqrt{\lambda}_i u_i v'_i \tag{1.4}$$

U $= [u_1 \quad u_2 \quad \ldots u_p]$ where u_i is the i^{th} $(n \times 1)$ column of U
V $= [v_1 \quad v_2 \quad \ldots v_p]$ where v_i is the i^{th} $(p \times 1)$ column of V
u_{ji} = the entry in the j^{th} row and the i^{th} column of U
v_{ji} = the entry in the j^{th} row and the i^{th} column of V
$\Delta = D_p$

1.2.2 The orthogonal model

The orthogonal or canonical form of the linear model (1.1) is defined by the following

$$\begin{aligned} Y &= \underbrace{XV}\ \ \underbrace{V'\beta} + \epsilon \\ &= Z \quad\ \ \alpha \quad + \epsilon \end{aligned} \tag{1.5}$$

Where

$Z = X_{n\times p}V_{p\times p}$ and $\alpha = V'_{p\times p}\beta_{p\times 1}$

From (1.4) and (1.5), we derive the following equations for later use in the study.

- $$X = U_{n\times p}\Delta_{p\times p}V'_{p\times p} = \sum_{i=1}^{p} \sqrt{\lambda_i} u_i v_i' \tag{1.6}$$

- $$Z = U\Delta \tag{1.7}$$

- $$\begin{aligned} X'X &= V\Delta U'U\Delta V' \\ &= V\Delta^2 V' = \sum_{i=1}^{p} \lambda_i v_i v_i' \end{aligned} \tag{1.8}$$

- $$Z'Z = \Delta^2 \tag{1.9}$$

- $$(Z'Z)^{-1} = \Delta^{-2} \tag{1.10}$$

- $$\begin{aligned} (X'X)^{-1} &= (V\Delta^2 V')^{-1} \\ &= (V')^{-1}\Delta^{-2}V^{-1} \\ &= V\Delta^{-2}V' \quad \ldots (since \quad V' = V^{-1}) \end{aligned} \tag{1.11}$$

- $$Trace(X'X) = \sum_{i=1}^{p} \lambda_i \tag{1.12}$$

- $$Trace(Z'Z) = \sum_{i=1}^{p} \lambda_i \tag{1.13}$$

- $$Trace(X'X)^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i} \tag{1.14}$$

- $$Trace(Z'Z)^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i} \tag{1.15}$$

1.3 The Assumptions underlying linear models

The following assumptions are adhered to linear regression models.

- The response variable is linearly related to the independent variables hence linear regression estimators are based on a linear equation.
- The observations on the dependent variable are from populations of random variables with the expectation equal to

  $$E[Y \mid X] = X\beta$$

- The independent variables are known, uncorrelated constants, measured without error.
- Random error terms have zero mean, a common variance and are pairwise independent.
- There are no dependencies among the error terms and the independent variables.
- Y_i and x_{ij} are paired observations, both measured on every observational unit.
- For purposes of making significance tests, the dependent variable and the error terms are assumed to be normally distributed.
- X is a fixed, full column rank matrix (orthogonal); $r(X) = p$.

Violation of some of the listed assumptions sometimes leads to poor estimation.

1.4 Bias, Variance, Mean Squared error and Total Mean Squared error

Let $\tilde{\beta}$ denote any estimator of β

1.4.1 Bias

The bias of $\tilde{\beta}$, is defined by the following

$$Bias[\tilde{\beta}] = E[\tilde{\beta}] - \beta \tag{1.16}$$

where

β= the true parameter vector.

Positive (negative) values of $Bias[\tilde{\beta}]$ imply that the estimates of $\tilde{\beta}$ are too much (little) in favour of what is being estimated. For this reason, unbiased estimators are mostly required.

1.4.2 Variance

The variance plays a vital role in regression analysis; it is one of the measures of precision of estimates, hence the basis for assessment of reliability of estimates. By definition

$$Var[\tilde{\beta}] = E\Big[(\tilde{\beta} - E[\tilde{\beta}])(\tilde{\beta} - E[\tilde{\beta}])'\Big] \tag{1.17}$$

Small values of $Var[\tilde{\beta}]$ imply high precision of estimates and vice versa; therefore, estimators whose variances are minimum are mostly desirable.

1.4.3 Mean Squared Error

Mean Squared error (MSE) represents the squared distance between the estimate and the actual parameter. Like the variance, the MSE is vital in assessing the quality of an estimator; the smaller it is, the closer the estimates are to the true values. A good estimator may be characterized by a relatively small MSE (McDonald and Galarneau, 1975).

The MSE can be decomposed into a sum of the variance and the squared bias of the estimator.

By definition,

$$MSE(\tilde{\beta}) = E\Big[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'\Big]$$

$$= E\Big[((\tilde{\beta} - E[\tilde{\beta}]) + (E[\tilde{\beta}] - \beta))((\tilde{\beta} - E[\tilde{\beta}]) + (E[\tilde{\beta}] - \beta))'\Big]$$

$$= \underbrace{E\Big[(\tilde{\beta} - E[\tilde{\beta}])(\tilde{\beta} - E[\tilde{\beta}])'\Big]}_{Var[\tilde{\beta}]} + \underbrace{(E[\tilde{\beta}] - \beta)(E[\tilde{\beta}] - \beta)'}_{(E[\tilde{\beta}]-\beta)(E[\tilde{\beta}]-\beta)'} + \underbrace{2E[(\tilde{\beta} - E[\tilde{\beta}])(E[\tilde{\beta}] - \beta)]}_{0} \tag{1.18}$$

1.4.4 Total Mean Squared Error (TMSE)

The TMSE is another measure of precision of estimates, defined by

$$TMSE[\tilde{\beta}] = Trace(MSE[\tilde{\beta}]) \tag{1.19}$$

A trace of a matrix is defined as a sum of the diagonal elements of the matrix under consideration. If A is an $(n \times n)$ matrix with eigenvalues $\lambda_1, \ldots \lambda_n$, then $Trace[A] = \sum_{i=1}^{n} \lambda_i$.

The interpretation of $TMSE[\tilde{\beta}]$ is similar to that of $Var[\tilde{\beta}]$ and $MSE[\tilde{\beta}]$; the smaller it is, the better the estimator and the more precise the estimates are. Importantly, high variances of estimators may be balanced with the bias; there exists a trade off between the variance and the bias of estimators.

1.5 Ordinary Least Squares (OLS) Estimation

In regression analysis, there always exists some error, of which the magnitude varies per estimation method employed and to a large extend, estimators that are unbiased and have the minimum variance are mostly favoured. The ordinary least squares estimator (OLSE) is one such estimator, known to be the best fitting linear unbiased estimator (BLUE) in the sense of minimum variance.

In a class of linear unbiased estimators, the least squares estimator has the least variance. This is a critical factor and the most desirable property because minimal variance implies closeness to the true parameter, thus accuracy. The least squares regression procedure employs the criterion that the solution must yield the smallest sum of squared deviations of the observed response variable from the estimates provided by the solution.

1.5.1 Derivation of the least squares estimator

The OLSE minimizes the residual errors

$$\begin{aligned}&\sum (Y_i - \hat{Y}_i)^2\\ &= (Y - X\beta)'(Y - X\beta)\\ &= Y'Y - Y'X\beta - (X\beta)'Y + (X\beta)'X\beta\end{aligned}$$

Differentiation with respect to β leads to the following

$$\frac{\partial(Y'Y - Y'X\beta - (X\hat{\beta})'Y + (X\beta)'X\beta)}{\partial\beta} = -2X'Y + 2X'X\beta$$

Equating the differential to zero and solving for β yields the following:

$$2X'X\beta = 2X'Y$$

$$X'X\beta = X'Y$$

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y\\ &= V\Delta^{-1}U'Y \qquad \ldots using \quad 1.6 \quad and \quad 1.11\\ &= \sum_{i=1}^{p} \frac{v_i u_i' Y}{\sqrt{\lambda_i}} \end{aligned} \tag{1.20}$$

1.5.2 Properties of $\hat{\beta}$

1.5.2.1 Expectation

$$
\begin{aligned}
E[\hat{\beta}] &= E[(X'X)^{-1}X'Y] \\
&= (X'X)^{-1}X'E[X\beta+\epsilon] \\
&= (X'X)^{-1}X'X\beta \quad since \quad E[\epsilon]=0 \\
&= \beta
\end{aligned} \tag{1.21}
$$

1.5.2.2 Bias

$$
Bias[\hat{\beta}] = E[\hat{\beta}] - \beta = 0 \qquad from \quad 1.16 \tag{1.22}
$$

1.5.2.3 Variance

$$
\begin{aligned}
Var(\hat{\beta}) &= Var[(X'X)^{-1}X'Y] \\
&= (X'X)^{-1}X' \quad Var[Y] \quad X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1} \\
&= \sigma^2 V\Delta^{-2}V' \qquad \ldots using \quad 1.11
\end{aligned} \tag{1.23}
$$

1.5.2.4 Mean squared error

$$
\begin{aligned}
MSE[\hat{\beta}] &= Var[\hat{\beta}] + (E[\hat{\beta}]-\beta)(E[\hat{\beta}]-\beta)' \qquad from \quad 1.18 \\
&= Var[\hat{\beta}] + 0 \\
&= \sigma^2 V\Delta^{-2}V' \qquad from \quad 1.23
\end{aligned} \tag{1.24}
$$

1.5.2.5 Total mean squared error

$$
\begin{aligned}
TMSE[\hat{\beta}] &= Trace(MSE[\hat{\beta}]) \\
&= Trace(\sigma^2(X'X)^{-1}) \\
&= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \qquad \ldots using \quad 1.14
\end{aligned} \tag{1.25}
$$

For model (1.5), the orthogonal least squares estimator is denoted by the following

$$
\hat{\alpha} = (Z'Z)^{-1}Z'Y = V'\hat{\beta} \tag{1.26}
$$

1.5.3 Properties of $\hat{\alpha}$

1.5.3.1 Expectation

$$\begin{aligned} E[\hat{\alpha}] &= E[(Z'Z)^{-1}Z'Y] \\ &= (Z'Z)^{-1}Z'Z\alpha \\ &= \alpha \end{aligned} \tag{1.27}$$

1.5.3.2 Bias

$$Bias[\hat{\alpha}] = E[\hat{\alpha}] - \alpha = 0 \qquad (using \quad 1.16) \tag{1.28}$$

Hence $\hat{\alpha}$ is an unbiased estimator of α.

1.5.3.3 Variance

$$\begin{aligned} Var(\hat{\alpha}) &= Var[(Z'Z)^{-1}Z'Y] \\ &= (Z'Z)^{-1}Z' \quad Var[Y] \quad Z(Z'Z)^{-1} \\ &= \sigma^2(Z'Z)^{-1} \\ &= \sigma^2\Delta^{-2} \qquad \ldots using \quad 1.10 \end{aligned} \tag{1.29}$$

1.5.3.4 Mean squared error

$$\begin{aligned} MSE[\hat{\alpha}] &= Var[\hat{\alpha}] + (E[\hat{\alpha}] - \alpha)(E[\hat{\alpha}] - \alpha)' \qquad from \quad 1.18 \\ &= Var[\hat{\alpha}] \\ &= \sigma^2\Delta^{-2} \qquad from \quad 1.29 \end{aligned} \tag{1.30}$$

1.5.3.5 Total mean squared error

$$\begin{aligned} TMSE[\hat{\alpha}] &= Trace(MSE[\hat{\alpha}]) \\ &= Trace(\sigma^2(Z'Z)^{-1}) \\ &= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \qquad \ldots using \quad 1.15 \end{aligned} \tag{1.31}$$

The OLSE has good qualities if it satisfies all the assumptions stated in section 1.3. Violation of the assumption of independence of the variables results in the OLSE exhibiting high values of the variance and mean squared error, being unstable and sensitive to minor changes in the data, thus being unacceptably unrealistic. This often means that the data vectors for the predictors are not orthogonal; that is, the matrix X is not full column rank hence there exist near-dependencies and or dependencies among the columns of X. Some of the reasons or sources of these dependencies and near dependencies include the following:

- over definition of the model such that the number of observations is less than that of the variables.
- generating other variables as function of others

These distortions and others usually result in collinearity: a serious problem in regression analysis, explicitly defined in chapter 2. Under collinear conditions, the least squares estimator remains unbiased but $MSE[\hat{\beta}]$ and $Var[\hat{\beta}]$ increase, hence $\hat{\beta}$ becomes unreliable. It is in these conditions when shrinkage estimators become a necessity; the 'fly in the ointment' with the least squares criterion is its requirement of unbiasedness (Marquardt and Snee, 1975).

Chapter 2

The Research Problem

Existence of dependencies and near dependencies among the independent variables (collinearity) has long been studied in statistics but even today, there are still good reasons to study and identify the potential harm of such conditions on regression modelling. In the event that one or more independent variables are defined by linear combinations of other independent variables, the OLSE can become unacceptable; the coefficients may be too large, be of wrong signs and be extremely sensitive and unstable. Further, the variances and standard errors of estimates may be inflated, leading to imprecision of the estimates. All these often create difficulties in inference of the separate influence of the independent variables on the response variable.

The research problem is defined in this chapter. Collinearity is regarded a broad problem from which the main research problem stems. We narrow our focus to the remedy for collinearity and instability of least squares estimates.

2.1 Defining the problem

In this thesis, we investigate shrinkage estimation as an alternative to least squares when the data are collinear. The motivation for this investigation has been induced by the following facts

- Collinear designs of matrices result in instability of the OLSE, thus unreliability and inconsistency of the least squares estimates. Failure to remedy collinearity is guaranteed to result in poor estimation.
- There are lots of shrinkage methods of estimation available but it is not clear which one is ideal. Most of the currently available shrinkage methods are not robust to collinearity; the methods are based on an unstable least squares solution.

Hence, we endeavor to solve the above problems by introducing a new procedure/method on which the shrinkage factors can be based. The new procedure depends on a stable solution hence we expect it to be more reliable and consistent, compared to the existing procedure. We setup a simulation study to assess the efficiency of different shrinkage estimators and select the potentially best alternative to least squares estimation.

2.2 Collinearity

Collinearity has long been and still is one of the major problems in statistical research that arises when there exist near-linear dependencies among the vectors of explanatory variables (Wetherill, 1986). If $\eta \geq 0$ is specified such that there exists a column vector

$$\mathbf{c} = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_p \end{pmatrix},$$

whose elements are not all equal to zero such that

$$\sum_{j=0}^{p} c_j X_j = \delta \qquad with \parallel \delta \parallel \quad \leq \quad \eta \parallel c \parallel, \tag{2.1}$$

then, collinearity exists among the covariates.

A special case of equation (2.1) for which $\delta = 0$ indicates that exact collinearity exists, and if δ is very small, the relationship is approximately true hence near-collinearity exists (Thiart, 1994). Variates are collinear if they lie on the same line or if the angle between them is very small or when the data vectors for predictors are not orthogonal (Hoerl and Kennard, 1980). Collinearity results when at least one dimension of the X-space is poorly defined such that the dispersion almost does not exist among the data points in that dimension. In a nutshell, collinearity describes a set of problems created when some combinations of the columns of matrix X are nearly zero such that the same information is provided in more than one way (redundancy).

Other terms for collinearity are: ill-conditioning (Gunst and Mason, 1977; Belsley et al., 1980; Belsley, 1987; Walker and Page, 2001; Liu, 2003), non-orthogonality (Farrar and Glauber, 1967; Hoerl and Kennard, 1980; Sundberg, 1993), over fitting (Le Cessie and Van Houwelingen, 1992), near collinearity (Mandel, 1982; Stewart, 1987; Thiart, 1994; Firinguetti and Rubio, 2000), conditioning (Belsley and Oldford, 1986), confluence analysis (Frish, 1934), multicollinearity (Hocking et al., 1976; Gunst et al., 1976; Winchern and Churchill, 1978; Askin and Montgomery, 1980; Dorsett et al., 1983; Gunst, 1983; Nomura, 1988; Ohtani, 1986; Oman, 1991; Troskie and Chalton, 1996; Allison, 1999; Wencheko, 2000), clustering (Grohn et al., 2003), singularity (Stewart, 1987), near singularity and near rank deficiency (Sengupta and Bhimasankaram, 1997; Knight and Fu, 2000).

2.2.1 Effects of Collinearity

Collinearity plagues multiple regression and other multivariate techniques. If not corrected for, it may cause serious problems in least squares regression particularly if the primary intention is to find separate influences of independent variables. Collinearity makes it virtually impossible to separate the marginal effects of the independent variables on the response variable.

In collinearity designs of matrices, the X matrix becomes non-orthogonal and is said to be of less than full column rank ($r(X) < p$). As a consequence, some of the singular values of X become very small and tend to have an adverse impact on the regression coefficients, variances, and reliability of estimation in general. An insight into specific effects of collinearity on least squares is provided below.

2.2.1.1 Unstable least squares estimates and poor prediction

In the presence of collinearity, the least squares estimate of β becomes unstable and sensitive to

- the computational method used, and
- errors in regressor variables (Sengupta and Bhimasankaram, 1997).

Little perturbations in either X or Y may result in unstable least squares coefficients (Thiart, 1990). This sensitivity and instability of least squares estimators make their predictions and forecasts generally unreliable.

2.2.1.2 Inflated variances of least squares estimates

Small eigenvalues inflate the variances of the corresponding regression coefficients and lead to wrong predictions and improper conclusions about the estimated regression coefficients (Thiart, 1994; Wetherill, 1986). Inflated variances result in large standard errors of regression coefficients, thus statistically insignificant coefficients, which may sometimes not be truly so. From 1.23, the variance of least squares estimates may be expressed as

$$Var(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p} \frac{v_i v_i'}{\lambda_i}$$

Hence, in the presence of collinearity, the eigenvalues ($\lambda_i's$), corresponding to collinear vectors become extremely small (near zero) and it follows automatically that, $\frac{v_i v_i'}{\lambda_i}$ and $Var(\hat{\beta})$ become inflated since the denominator is a number close to zero.

2.2.1.3 Unexpected coefficient signs

It is often assumed that signs of coefficients are known by intuition. However, when the data are collinear, the coefficients bear unanticipated signs.

2.2.1.4 Large coefficients

Collinearity leads to unacceptably large coefficients of the correlated variables. We note that

$$\hat{\beta} = \sum_{i=1}^{p} \frac{v_i u_i' Y}{\sqrt{\lambda_i}} \qquad \text{(from 1.20)}$$

When collinearity is present in the data, the singular values ($\sqrt{\lambda_i}s$) that correspond to collinear vectors approach zero, hence, $\frac{v_i u_i' Y}{\sqrt{\lambda_i}}$ increases, resulting in large values of coefficients.

2.2.2 Collinearity Measures

Several techniques are available in the statistical literature for detecting collinearity however, there is no particular ideal technique for detection of collinearity therefore more than one techniques should be used. The most important factor is being able to observe and identify collinearity in the data.

Before proceeding to the techniques, the following important observation needs to be made. All the techniques are applied on the correlation form (scaled and centered) of the X matrix. However, it is also important to note that centering removes the intercept from the models hence the models in which the intercept plays a vital role may not be practically sound when the X matrix is centered. Therefore it is sometimes necessary to scale and not center.

- **Centering**

  Centering the X matrix entails subtracting the column means from corresponding elements of columns of X to eliminate collinearites that may be due to the origins of the predictor variables.

- **Scaling**

  The columns of a matrix are said to be scaled when each element of the column is divided by the root of the sums of squares of all the elements in the corresponding column such that the length of each column of X is one. Scaling ensures uniformity in the measurement of the predictor variable.

- **Standardizing**
  Standardizing means centering and scaling such that the matrix $X'X$ is in correlation form.

The following techniques are some of the detective measures of collinearity

2.2.2.1 Correlation Matrix

The correlation matrix, denoted by $X'X$, presents the values of correlations between pairs of independent variables (bivariate correlations). Correlations close to 1 indicate serious collinearity between the pairs of independent variables.

Despite its importance, the correlation matrix cannot be relied upon for full diagnosis of collinearity since it cannot detect existence of more than two dependencies in a matrix.

2.2.2.2 Variance Inflation Factors

The i^{th} variance inflation factor (VIF_i) (Chatterjee and Price, 1977), is defined by

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2.2}$$

where R_i^2 is the magnitude of variation in the i^{th} independent variable X_i, explained by the variation in the rest of the independent variables when the regression model is such that X_i is the response variable and other independent variables are the explanatory variables. The dependencies involving X_i and other independent variables is signified by the closeness of R_i^2 to 1, thus a high VIF_i. Any variance inflation factor greater than 10 indicates collinearity (Wetherill, 1986).

Although variance inflation factors may be reliable, they are unable to detect more than two coexisting dependencies or near dependencies, just like the correlation matrix.

2.2.2.3 Farrar and Glauber Technique

The Farrar and Glauber Technique (Farrar and Glauber, 1967) measures collinearity based on the following assumptions:

- Matrix X is a sample of size (n) from a p-variate Gaussian (Normal) distribution.
- X has orthogonal columns

This technique employs both the determinant of the correlation matrix, $(X'X)$ and the variance inflation factors. The procedure involves transformation of the determinant of $(X'X)$ and the use of variance inflation factors as indicators of variates involved in collinearity.

Nonetheless, the Farrar and Glauber technique is usually not used as a statistical test for collinearity due to the fact that it uses the determinant of $(X'X)$ and the determinant is very sensitive to scaling hence may not be trusted. The technique also depends on the correlation matrix and from what we observe from 2.2.2.1, the correlation matrix cannot be entirely relied upon for diagnosis of collinearity. Further, the technique relies on orthogonality of the X matrix and from the definition of collinearity and the effect, we note that X becomes non-orthogonal in the presence of collinearity.

2.2.2.4 Bunch Maps

Bunch Maps (Belsley et al., 1980) are graphical investigations of the possible relationships among sets of data. They indicate location of dependencies but do not determine the degree to which regression results are degraded by their presence. However, the bunch maps are not recommended for use as a major tool in regression because their extension to dependencies among more than two variates is time consuming and subjective (Belsley et al., 1980).

2.2.2.5 Small eigenvalues

If a matrix has one or more eigenvalues that are almost zero or too small compared to others, then collinearity exists in the data. Small eigenvalues correspond to large elements of eigenvectors therefore any of the two may be a sign of collinearity (Belsley et al., 1980).

2.2.2.6 A small determinant

A matrix is not invertible or near-singular if its determinant is zero. Near-dependencies and dependencies can clearly be detected from the determinant that is extremely small such that the inverse almost does not exist. Since the determinant is generally computed as the product of the eigenvalues of a square matrix and in the presence of collinearity, it tends to zero and the matrix becomes singular. Nonetheless, as pointed out under the Farrar and Glauber technique, the determinant is very sensitive to scaling and cannot be fully relied upon for diagnosis.

2.2.2.7 Condition Number

The condition number (Belsley et al., 1980) of matrix X is the ratio of the largest to the smallest singular value;

$$C = \left(\frac{\lambda_{max}}{\lambda_{min}}\right)^{1/2}$$

It measures the sensitivity of the solution to small changes in X or Y. A condition number greater than 100 indicate extreme collinearity.

2.2.2.8 Condition Index

The condition indices (Belsley et al., 1980) identify the dimensions of the X-space where the dispersion is limited to cause problems in the least squares solution. The h^{th} condition index is denoted by

$$k_h = \left(\frac{\lambda_{max}}{\lambda_h}\right)^{1/2} \qquad h = 1, \ldots, p$$

Condition Indices measure collinearity in the following manner

Condition index	Collinearity
$10 - 30$	weak
$30 - 100$	moderate - strong
100+	extreme

The problem with using condition indices is that it is not easy to identify the columns responsible for collinearity.

2.2.2.9 Multicollinearity index (mci)

The multicollinearity index (Thisted, 1980) is a measure of collinearity that involves the ratios of the squares of eigenvalues of $X'X$. The mci is defined by

$$mci = \sum_{i=1}^{p} \left(\frac{\lambda_p}{\lambda_i}\right)^2$$

where λ_p is the smallest eigenvalue of $X'X$.

Values of mci close to 1 indicate high collinearity and mci greater than 2 indicate little or no collinearity (Thisted, 1980).

2.2.2.10 Variance Decomposition

Unlike the collinearity measures discussed so far, the variance decomposition (Belsley et al., 1980) makes it possible to identify the columns of X involved in collinearity. The variance of j^{th} component of $\hat{\beta}$ may be defined to be

$$Var(\hat{\beta}_j) = \sigma^2 \sum_{i=1}^{p} \frac{v_{ji}v'_{ji}}{\lambda_i} \qquad (using \quad 1.23)$$

Where v_{ji} = the j^{th} element of the i^{th} eigenvector.

The above equation decomposes the variance of the j^{th} coefficient into a sum of components, each associated with one of the eigenvalues. The variance-decomposition proportion is the variance of the j^{th} regression coefficient associated with the i^{th} component of the decomposition of j^{th} regression coefficient.

The variance-decomposition proportions are calculated by

$$\pi_{ij} = \frac{\phi_{ji}}{\phi_j} \qquad i, j = 1, \ldots, p. \tag{2.3}$$

where

$$\phi_{ji} = \frac{v_{ji}v'_{ji}}{\lambda_i}, \qquad \phi_j = \sum_{i=1}^{p} \phi_{ji} \qquad i = 1, \ldots, p.$$

We note that when collinearity is present in the data, the eigenvalues corresponding to covariates involved in dependencies become small. From the equation above, we observe that if the eigenvalues are small, then the variance proportions increase hence, we may identify the columns involved in collinearity. Collinearity becomes a problem when the variance proportions of at least two regression coefficients associated with components that correspond to small eigenvalues exceed 50%.

2.2.3 Coping with collinearity

There are lots of remedial measures for collinearity reported in the literature. These include among others additional data (2.2.3.1), deletion of collinear variables (2.2.3.2), transformation of variables (2.2.3.3), bayesian methods (2.2.3.4), detrending of variables (2.2.3.5), first differencing (2.2.3.6) and shrinkage estimation (2.2.3.7). It must be emphasized that not all the remedial measures are effective; hence, it is imperative to find the most appropriate.

According to Marquardt and Snee (1975), several methods proposed to handle collinearity are usually not met in practice. We maintain that most of the remedial measures of collinearity suggested in the statistical literature, have recognizable disadvantages that should not just be ignored. Failure to recognize these disadvantages is likely to hinder effectiveness of the corresponding remedial measures. In the light of the above considerations, some of the remedial measures of collinearity are discussed below and the corresponding shortcomings identified.

2.2.3.1 Additional data

Obtaining additional data or collecting new data is considered one of the methods for solving collinearity problems. More often, the additional data is taken in the direction of the collinearities such that the X-space is expanded to eliminate the dependencies (Rawlings et al., 1998).

Although the procedure sounds simple, it is usually regarded impractical because

- analysts may not generally be in control of variables to obtain well-behaved data (Jagpal, 1982).
- Data collection may be expensive and/or time-consuming.

2.2.3.2 Deletion of collinear variables

Deletion of the correlated variables reduces collinearity but also reduces the interpretability of the regression equation. Also, the deleted variables may tend to be the most important variables of the model, thus improper or false estimation and sometimes biasness may result.

2.2.3.3 Transformation of variables

Although collinearity may sometimes be removed by appropriate transformations of the explanatory variables (Wetherill, 1986), transformations may completely change the models, thereby leading to estimation of models that differ from the original.

2.2.3.4 Bayesian methods

The use of a priori information or Bayesian methods of estimation is sometimes useful when collinearity is a problem (Gruber, 1980). However, Bayesian methods require a priori information about the distribution of regression parameters which may sometimes not be available.

2.2.3.5 Detrending of variables

Expressing the variables in terms of deviations from their linear trends reduces collinearity (Gruber, 1980). However, the procedure reduces dependency of Y on X and also changes the original specification of the regression equation.

2.2.3.6 First Differences

Expression of the variables in the first differences (the differences between the current and the previous values) overcomes collinearity (Gruber, 1980). However, the procedure produces even greater variances of parameter estimates than OLS (Sujan and Condik, 1979).

2.2.3.7 Shrinkage estimation

Shrinkage estimation defines a class of biased methods of estimation, known to shrink the least squares estimators $\hat{\beta}$ proportionally towards zero. By allowing for a little bias, the methods stabilize the regression and provide estimates with smaller variance (a trade off between high coefficient variances and a little bias) (Le Cessie and Van Houwelingen, 1992). The methods are vital and critically useful in cases whereby collinearity causes the least squares parameter estimates to be too large in absolute values (Gruber, 1998). Unlike discrete procedures such as model selection, shrinkage methods are continuous and therefore do not exhibit high variance (Hastie et al., 2001)

In a nutshell, shrinkage estimation is more important when the existing dependencies among the covariates lead to the following problems:

- unstable coefficients,
- inflated variances, hence large standard errors,
- insignificant coefficients and
- poor prediction and or improper modelling.

Besides being the most practical way to correct for collinearity, shrinkage methods are easy to deal with. Even more important is the fact that there are lots of shrinkage techniques to choose from; there is always at least one shrinkage method appropriate to curb problems attributed to collinearity.

2.3 Summary

In this chapter, the instability of OLS estimates and computation of the shrinkage factors from the OLS solution were identified as the problems addressed by this thesis.

Collinearity was broadly discussed as a phenomenon from which the main research problem of this thesis stems. The definition of collinearity, detective measures, effects and the approaches to collinearity were provided in details. It was noted that presence of collinearity among the independent variables leads to high variability and computational instability of the OLSE; the sampling variances of the estimates become very large hence, the distance between the estimates and the true values becomes extremely large.

Several remedial measures for collinearity were provided and the respective disadvantages highlighted. Shrinkage estimation was considered a class of the most effective approaches to the problems attributed to collinearity since it directly reduces the mean squared error of estimates.

From our point view, usage of parameter estimation techniques that minimize the mean square error of estimates is the most effective way to deal with problems associated with collinearity. Shrinkage estimation is one such technique, of which the estimates are biased but have smaller mean squared errors compared to least squares when collinearity is the problem. We stress that usage of biased methods of estimation in collinear designs of matrices is the most practical way to reach meaningful conclusions.

Chapter 3

Ridge and Shrinkage Estimators

We consider shrinkage estimation a powerful alternative to OLS, for which there is low risk of imprecision of estimates and poor estimation when collinearity is the problem. Shrinkage methods result in biased estimates but most importantly, lead to a significant reduction in the variances and mean squared error values of estimates when the OLS estimates exhibit high variances. We view shrinkage as the most practically convenient and reasonable way to estimate the parameters when the least squares estimates are unstable and imprecise.

For each shrinkage method, there is a specific shrinkage factor for which the variances of the corresponding shrinkage estimates are less than the variances of the least squares estimates. By sacrificing a little bias, each shrinkage estimation method reduces the mean squared error values of estimates, stabilizes the coefficients and produces estimates that are highly precise and almost accurate. This is one attractive feature of shrinkage estimation methods that makes them highly important in research application.

This chapter is organized as follows. The general shrinkage estimator and its properties are defined in the next section. Some of the special cases of shrinkage estimation are discussed in subsequent sections; specifically, we discuss Stein estimation (Stein, 1956), ridge and generalized ridge regression (Hoerl and Kennard, 1970a), Liu and generalized Liu estimation techniques (Liu, 1993) and principal components regression (Kendall, 1957).

For each of the shrinkage methods, we specify the following

* the shrinkage factor,

* the expectation,

* the bias.

* the variance of the estimator,

- * the mean squared error of the estimator and
- * the total mean squared error.

Considerable attention is drawn to ridge regression and the different ways in which the ridge biasing constants are selected.

3.1 The shrinkage estimator

We denote the general shrunken estimator $\hat{\beta}_{sh}$ by the following

$$\begin{aligned}\hat{\beta}_{sh} &= d_{sh}\hat{\beta}\\ &= d_{sh}(X'X)^{-1}X'Y \qquad (using \quad 1.20)\\ &= d_{sh}V\Delta^{-1}U'Y \end{aligned} \tag{3.1}$$

where

d_{sh} = a shrinkage factor within the bounds $0 < d_{sh} < 1$
$\hat{\beta}$ = a vector of least squares coefficients, (defined in § 1.5.1)
V, U and Δ are defined in § 1.2.1.

3.1.1 Properties of $\hat{\beta}_{sh}$

3.1.1.1 Expectation

$$\begin{aligned}E[\hat{\beta}_{sh}] &= E[d_{sh}\hat{\beta}]\\ &= d_{sh}E[\hat{\beta}]\\ &= d_{sh}\beta \qquad (using \quad 1.21)\end{aligned} \tag{3.2}$$

Hence, shrinkage estimates are biased for β.

3.1.1.2 Bias

$$\begin{aligned}Bias[\hat{\beta}_{sh}] &= E[\hat{\beta}_{sh}] - \beta \quad (using \quad 1.16)\\ &= d_{sh}\beta - \beta\\ &= (d_{sh} - 1)\beta\end{aligned} \tag{3.3}$$

3.1.1.3 Variance

$$\begin{aligned}Var[\hat{\beta}_{sh}] &= Var[d_{sh}\hat{\beta}]\\ &= d_{sh}^2\sigma^2(X'X)^{-1} \qquad (using \quad 1.23)\\ &= d_{sh}^2\sigma^2V\Delta^{-2}V'\end{aligned} \tag{3.4}$$

3.1.1.4 Mean Squared Error

$$MSE[\hat{\beta}_{sh}] = Var[\hat{\beta}_{sh}] + (d_{sh} - 1)^2\beta\beta' \qquad (using \quad 1.18)$$

$$= d_{sh}^2\sigma^2(X'X)^{-1} + (d_{sh} - 1)^2\beta\beta'$$

$$= d_{sh}^2\sigma^2 V\Delta^{-2}V' + (d_{sh} - 1)^2\beta\beta' \tag{3.5}$$

3.1.1.5 Total Mean Squared Error

$$TMSE[\hat{\beta}_{sh}] = Trace[MSE(\hat{\beta}_{sh})] \qquad (from \quad 1.19)$$

$$= d_{sh}^2\sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i} + (d_{sh} - 1)^2\sum_{i=1}^{p}\beta_i^2 \tag{3.6}$$

where λ is defined in § 1.2.1.

3.1.2 Desirable qualities of shrinkage estimators

From the properties of $\hat{\beta}_{sh}$, the following important qualities of biased estimators may be observed:

(i) $$Var[\hat{\beta}_{sh}] = d_{sh}^2\sigma^2 V\Delta^{-2}V'$$

$$= d_{sh}^2 Var[\hat{\beta}]$$

Implying that $Var[\hat{\beta}_{sh}] < Var[\hat{\beta}]$ since $0 < d_{sh} < 1$,

(ii) The squared length of $\hat{\beta}_{sh}$ is shorter than that of $\hat{\beta}$. That is

$$\hat{\beta}_{sh}'\hat{\beta}_{sh} = d_{sh}^2\hat{\beta}'\hat{\beta} < \hat{\beta}'\hat{\beta}$$

since $d_{sh}^2 < 1$

(iii) $TMSE[\hat{\beta}_{sh}] < TMSE[\hat{\beta}]$

The Admissibility Condition (Mayer and Willke, 1973) states that a shrinkage estimator is said to be mean square admissible if and only if there exists a shrinkage factor d_{sh} such that $TMSE[d_{sh}\hat{\beta}] < TMSE[\hat{\beta}]$. This condition is satisfied only when

$$d_{sh} > \frac{\sum_{i=1}^{p}\beta_i^2 - TMSE[\hat{\beta}]}{\sum_{i=1}^{p}\beta_i^2 + TMSE[\hat{\beta}]} = \frac{\beta'\beta - Trace\Big(MSE[\hat{\beta}]\Big)}{\beta'\beta + Trace\Big(MSE[\hat{\beta}]\Big)}$$

To find out whether or not $\hat{\beta}_{sh}$ is admissible, we minimize 3.6 subject to d_{sh}, equate the result to zero and finally solve for d_{sh}. If d_{sh} satisfies the above stated condition, then $\hat{\beta}_{sh}$ is mean square admissible, hence $TMSE[\hat{\beta}_{sh}] < TMSE[\hat{\beta}]$.

$$TMSE[\hat{\beta}_{sh}] = d_{sh}^2\sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i} + (d_{sh}-1)^2\sum_{i=1}^{p}\beta_i^2 \qquad (from \quad 3.6)$$

$$= d_{sh}^2 TMSE[\hat{\beta}] + (d_{sh}-1)^2\sum_{i=1}^{p}\beta_i^2 \qquad (using \quad 1.25)$$

We find the partial derivatives of $TMSE[\hat{\beta}_{sh}]$ with respect to d_{sh}:

$$\frac{\partial(TMSE[\hat{\beta}_{sh}])}{\partial d_{sh}} = \frac{\partial(d_{sh}^2\sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i})}{\partial d_{sh}} + \frac{\partial((d_{sh}-1)^2\sum_{i=1}^{p}\beta_i^2)}{\partial d_{sh}}$$

$$= 2d_{sh}\sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i} + 2(d_{sh}-1)\sum_{i=1}^{p}\beta_i^2$$

$$= 2d_{sh}TMSE[\hat{\beta}] + 2(d_{sh}-1)\sum_{i=1}^{p}\beta_i^2$$

Thus

$$\frac{\partial(TMSE[\hat{\beta}_{sh}])}{\partial d_{sh}} = 0 \qquad implies \quad that$$

$$d_{sh}TMSE[\hat{\beta}] + d_{sh}\sum_{i=1}^{p}\beta_i^2 = \sum_{i=1}^{p}\beta_i^2$$

Hence

$$d_{sh} = \frac{\sum_{i=1}^{p}\beta_i^2}{TMSE[\hat{\beta}] + \sum_{i=1}^{p}\beta_i^2} > \frac{\beta'\beta - Trace\Big(MSE[\hat{\beta}]\Big)}{\beta'\beta + Trace\Big(MSE[\hat{\beta}]\Big)}$$

Therefore $\hat{\beta}_{sh}$ is mean square admissible thus $TMSE[\hat{\beta}_{sh}] < TMSE[\hat{\beta}]$. This implies that there is always a shrinkage factor for which the shrinkage estimator is guaranteed to have a smaller total mean squared error compared to OLSE.

Some of the special cases of shrinkage estimation are discussed below. Their properties stem from the properties of $\hat{\beta}_{sh}$ hence we do not repeat them. Rather, we use the final expressions from the general properties to specify the properties of each of the cases considered; we simply substitute appropriate shrinkage factors for d_{sh}.

3.2 Stein Estimation

Stein estimation is a shrinkage method, first proposed by Stein (1956) and later reviewed by James and Stein (1961), Dempster (1973), Efron and Morris (1973), Wind (1973), Zellner and Vanele (1974), Gruber (1979; 1998) to mention a few. The estimation method has long been used in the statistical analysis to substitute least squares estimation when the OLSE is unsatisfactory; hence it is still one of the important estimation procedures that may be put into practice.

3.2.1 The Stein estimator

The Stein estimator is a shrinkage estimator of which the shrinkage factor d_{sh}=c, and is defined by the following:

$$\hat{\beta}_s = c\hat{\beta} \qquad for \quad 0 < c < 1$$

3.2.1.1 Properties of $\hat{\beta}_s$

From the properties of $\hat{\beta}_{sh}$, the following may be specified for $\hat{\beta}_s$, for which d_{sh} is replaced by c.

Expectation:

$$E[\hat{\beta}_s] = c\beta \qquad (from \quad 3.2) \qquad (3.7)$$

Bias:

$$Bias[\hat{\beta}_s] = (c-1)\beta \qquad (from \quad 3.3) \qquad (3.8)$$

Variance:

$$Var[\hat{\beta}_s] = c^2\sigma^2 V\Delta^{-2}V' \qquad (from \quad 3.4) \qquad (3.9)$$

Mean Squared Error:

$$MSE[\hat{\beta}_s] = c^2\sigma^2 V\Delta^{-2}V' + (c-1)^2\beta\beta' \qquad (from \quad 3.5) \qquad (3.10)$$

Total Mean Squared Error:

$$TMSE[\hat{\beta}_s] = c^2\sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} + (c-1)^2 \sum_{i}^{p} \beta_i^2 \qquad (from \quad 3.6) \qquad (3.11)$$

3.2.2 The Stein shrinkage factor

Examples of the most common suggestions for estimation of c include the following

- **James and Stein (1961)**

  Given that $X'X = I_p$ and $p \geq 3$, James and Stein proposed the following expression for c.

$$c = max\left(0, \left[1 - \frac{(p-2)(n-p)\hat{\sigma}^2}{(n-p+2)\hat{\beta}'\hat{\beta})}\right]\right) \tag{3.12}$$

  where

  $\hat{\sigma}^2$ is the least squares variance, obtained after fitting the least squares estimator $\hat{\beta}$, max(0,a) returns a the largest number in a set of values ranging between 0 and a.

- **Sclove (1968)**

  Sclove modified the above expression by substituting the coefficients from the orthogonal or canonical form of least squares for $\hat{\beta}'\hat{\beta}$. That is

$$c = max\left(0, \left[1 - \frac{(p-2)(n-p)\hat{\sigma}^2}{(n-p+2)\sum_{i=1}^{p}\lambda_i\hat{\alpha}_i^2}\right]\right) \tag{3.13}$$

  Further, based on the canonical form of the model, Sclove proposed that a subset of the least squares parameters should be shrunken such that the shrinkage estimator may be expressed as

$$\mathbf{d}_{\mathbf{sh}} = \begin{bmatrix} I_r & 0 \\ 0 & cI_{p-r} \end{bmatrix} \qquad 0 < c < 1,$$

  where

  r eigenvalues are significantly different from zero (rank of X =r) and

  c only shrinks the last p-r components of $\hat{\beta}$ that correspond to the smallest eigenvalues.

- **Van Houwelingen and Le Cessie (1990)**

  Van Houwelingen and Le Cessie proposed cross-validation calibration for estimation of c. The procedure is carried out in following subsequent steps:

  - For all i, compute $\hat{\beta}_{-i}$ as a vector of coefficients estimated from a regression in which the i^{th} observation has been excluded. That is $\hat{Y}_{-i} = X'_{-i}\hat{\beta}_{-i}$.
  - Perform a single variable linear regression of Y_i on $\hat{Y}_{-i}$.
  - Use the resulting coefficient (slope) as an estimate of the shrinkage factor $\hat{c}$.

The procedure was further investigated by Le Cessie and Van Houwelingen, 1992; Vach et al., 2001; Van Houwelingen, 2001; Sauerbrei, 1999. all of who acknowledge that the procedure overcomes large variances of least squares estimates caused by dependencies and or near-dependencies of the independent variables.

* **Breiman's Garrote**

Breiman (1995) defined a procedure for which the corresponding estimators was called Breiman's Garrote. For a given threshold $t \geq 0$, the Garotte shrinkage factor c is obtained from

$$\mathbf{c} = \begin{bmatrix} c_1 & 0 & 0 & 0 \\ 0 & c_2 & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & 0 & c_p \end{bmatrix}$$

under a constraint

$$\sum_{j=1}^{p} c_j \leq t \qquad for \quad c_j \geq 0$$

If t is predetermined to be p, that is if $t = p$, then $c_j = 1$ hence c is a diagonal matrix of ones; $c = I_p$. Thus the Garotte estimator becomes equivalent to $\hat{\beta}$; if for convenience, we denote the Garotte estimator by $\hat{\beta}^G$, then $\hat{\beta}^G = \hat{\beta}$ when t=p. On the other hand, if t is predetermined to be small, then some of the c_j tend to zero hence the corresponding coefficients also approach zero. The optimal value of t is selected by crossvalidation (Vach et al., 2001). That is, t is selected such that the following function is minimized

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} x_{ij} \hat{\beta}^G_{j(-i)}(t) \right)^2$$

where $\hat{\beta}^G_{j(-i)}(t)$ is the j^{th} Garotte estimate, computed from a specific value of t, with the i^{th} observation omitted.

3.3 Ridge Regression

Ridge regression is a shrinkage procedure, originating from a response surface technique known as ridge Analysis and first introduced by Hoerl (1959). The technique was used to graphically portray the behaviour of high-dimensional quadratic response surfaces and to locate optimal regions and the first publication on its application on regression problems was on the chemical plant data (Hoerl, 1962). This was further extended by Hoerl and Kennard (1968) to include

Bayesian interpretation of ridge and the comparison of ridge and estimation of β when constrained to a bounded convex set.

Although ridge analysis had been proved to be important in locating the optimal predicted variables in spaces of predictor variables, computational instability of least squares estimates remained a problem when the data were collinear, therefore, Hoerl and Kennard (1970a, 1970b) proposed and later published a new estimation method (ridge regression), to address problems that could be attributed to collinearity. The family of estimates given by the ridge biasing parameter $[k \geq 0]$ in the newly introduced ridge regression seemed mathematically similar to portrayal of quadratic response functions (ridge analysis) hence the analysis built around the new technique was been labelled 'ridge regression' (Hoerl and Kennard, 2000).

Subsequent to its publication, ridge regression was further investigated and given so much attention that it masked ridge analysis, leading to sparse literature on the ridge analysis technique (Hoerl, 1985).

3.3.1 Definition

Since 1970, the following expressions have been interchangeably used to define ridge regression.

- Ridge regression is a biased estimation technique and a formal procedure that has been developed to compensate for effects of collinearity (Swindel, 1981; Hawkins and Yin, 2002; Akdeniz et al., 2003; Sundberg, 1993; Walker and Birch, 1988; Gunst, 1980).

- It is an important estimation technique in the theory of point estimation which provides estimators with smaller mean square error than Least Squares when collinearity is present in the data (Halawa and El Bassiouni, 2000; Ngo et al., 2003; Gunst and Mason, 1977; Elston and Proe, 1995; Hoerl and Kennard, 1970a).

- Ridge regression is sometimes regarded a restricted or constrained least squares estimation method (Gibbons and McDonald, 1984; Grob, 2003) that may be used to portray the sensitivity of the estimates to the set of data in use (Hoerl and Kennard, 1970a).

- It is a classical statistical algorithm that imposes a penalty or a restriction on the size of coefficients to obtain stable results (Le Cessie and Van Houwelingen, 1992; Hong et al., 2004; Hastie et al., 2001).

- It is a procedure intended to overcome 'ill-conditioned' situations, where near dependencies between columns of X cause:

    - near-singularity of the correlation matrix $(X'X)$ and

  - instability in the parameter estimates (Swindel, 1981).

- Ridge regression is an alternative to least squares estimation, used as a tool to alleviate collinearity or non-orthogonality (Wan, 2002; Troskie and Chalton, 1996; Hoerl, 1985; Kidwell and Brown, 1982; Conniffe and Stone, 1973; Thiart, 1990).

- Ridge regression may also be considered an estimation procedure based on the equation that defines a class of estimators indexed by a scalar parameter (McDonald and Galarneau, 1975).

Ideally, ridge regression is an estimation procedure based on adding small positive quantities (bias, biasing parameters or characterizing scalars (Dwivedi et al., 1980)) to the diagonal of the correlation matrix of independent variables, hence it produces biased estimators. Although ridge estimators are biased, they are less affected by small changes in the data and are much more stable than least squares estimators when prediction vectors are not orthogonal. Ridge estimates may be used to obtain point estimates with minimum MSE in cases where the estimates are sensitive to particular sets of data being used (Hoerl and Kennard, 2000).

3.3.2 The ridge estimator

The ridge estimator $\hat{\beta}_R$ is a shrinkage estimator for which the shrinkage matrix is defined by:

$$[V\Delta^2 V' + kI]^{-1} V\Delta^2 V' \qquad k > 0$$

where Δ and V are defined in section 1.2.1, and k is the ridge constant, also known as the biasing constant or the shrinkage parameter.

3.3.2.1 Derivation of the ridge estimator

$\hat{\beta}_R$ minimizes

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

subject to

$$\sum_{j=1}^{p} \beta_j^2 \leq t \tag{3.14}$$

where t is an arbitrary constant and x_{ij} is the element in the i^{th} row and the j^{th} column of X.

Equivalently, we may say the ridge estimator shrinks the OLSE by imposing a penalty on their size as follows:

$$\hat{\beta}_R = argmin[\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + k\sum_{j=1}^{p}\beta_j^2] \tag{3.15}$$

where $k\sum_{j=1}^{p}\beta_j^2$ is the penalty and k is a shrinkage parameter that has a direct relationship with t in (3.14).

Let $f(\beta) = [(Y - X\beta)'(Y - X\beta) + k\beta^2]$
$= Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta + k\beta^2$

To derive the ridge estimator, we differentiate $f(\beta)$ once with respect to β and equate the derivative to zero and solve for the unknown coefficient:

$$\frac{\partial f(\beta)}{\partial \beta} = -2X'Y + 2X'X\beta + 2k\beta = 0$$

$$2X'Y = 2X'X\beta + 2k\beta = 0$$

$$X'Y = (X'X + kI)\beta$$

$$\hat{\beta}_R = (X'X + kI)^{-1}X'Y \qquad k > 0 \tag{3.16}$$

3.3.2.2 Relation to OLSE

The ridge solution is a linear transform of least squares solution, where the transform depends on the biasing parameter k.

Let

$$W = (X'X + kI)^{-1} \tag{3.17}$$

It follows from simple substitution of (3.17) into (3.16) that $\hat{\beta}_R = WX'Y$.

Post-multiplication of W and pre-multiplication of $X'Y$ by the Identity matrix $I = X'X(X'X)^{-1}$ leads to

$$\hat{\beta}_R = WX'X\underbrace{(X'X)^{-1}X'Y}_{\hat{\beta}}$$

Implying that

$$\hat{\beta}_R = WX'X\hat{\beta}$$
$$= \Big((V\Delta^2V' + kI)^{-1}V\Delta^2V'\Big)\hat{\beta} \qquad k > 0$$

Let

$$G = WX'X = (V\Delta^2V' + kI)^{-1}V\Delta^2V' = I - k(V\Delta^2V' + kI)^{-1} \tag{3.18}$$

then

$$\hat{\beta}_R = G\hat{\beta} \tag{3.19}$$

Hence $\hat{\beta}_R$ is a shrinkage estimator of which the shrinkage matrix is G.

Considering the orthogonal form of the linear model (1.5), we may express the ridge estimator as follows:

$$\begin{aligned}\hat{\beta}_R &= (Z'Z + kI)^{-1}Z'Z\hat{\beta} \\ &= (\Delta^2 + kI)^{-1}\Delta^2\hat{\beta} \qquad \textit{using} \quad 1.9 \end{aligned} \tag{3.20}$$

Hence for the orthogonal form of the linear model,

$$G = \begin{bmatrix} \frac{\lambda_1}{\lambda_1+k} & 0 & 0 & 0 \\ 0 & \frac{\lambda_2}{\lambda_2+k} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & \frac{\lambda_p}{\lambda_p+k} \end{bmatrix}$$

Each i^{th} element of $\hat{\beta}$ is shrunk by $\dfrac{\lambda_i}{\lambda_i + k}$, implying that

$$\hat{\beta}_{R_i} = \frac{\lambda_i}{\lambda_i + k}\hat{\beta}_i$$

where $\hat{\beta}_{R_i}$ is the i^{th} element of $\hat{\beta}_R$ and the i^{th} shrinkage factor is

$$\frac{\lambda_i}{\lambda_i + k}, \qquad 0 < \frac{\lambda_i}{\lambda_i + k} < 1$$

3.3.2.3 Properties of $\hat{\beta}_R$

From the general properties of the shrinkage estimator specified in 3.2, 3.3, 3.4, 3.5 and 3.6, the following properties may be specified for the ridge estimator $\hat{\beta}_R$.

Expectation

$$\begin{aligned} E[\hat{\beta}_R] &= G\beta \\ &= (V\Delta^2V' + kI)^{-1}V\Delta^2V'\beta \\ &= \Big[I - k(V\Delta^2V' + kI)^{-1}\Big]\beta \qquad (\textit{using} \quad 3.18) \end{aligned} \tag{3.21}$$

Bias

$$\begin{aligned} Bias[\hat{\beta}_R] &= (G - I)\beta \\ &= \Big[(V\Delta^2V' + kI)^{-1}V\Delta^2V' - I\Big]\beta \\ &= -k(V\Delta^2V' + kI)^{-1}\beta \qquad (\textit{using} \quad 3.21) \end{aligned} \tag{3.22}$$

Variance

$$\begin{aligned} Var(\hat{\beta}_R) &= \sigma^2 G(X'X)^{-1}G' \\ &= \sigma^2 GV\Delta^{-2}V'G' \\ &= \sigma^2\Big(I - k(V\Delta^2V' + kI)^{-1}\Big)V\Delta^{-2}V'\Big(I - k(V\Delta^2V' + kI)^{-1}\Big) \\ &= \sigma^2(V\Delta^2V' + kI)^{-1}V\Delta^2V'(V\Delta^2V' + kI)^{-1} \end{aligned} \tag{3.23}$$

Mean Squared Error

$$\begin{aligned} MSE(\hat{\beta}_R) &= \sigma^2 G(X'X)^{-1}G' + (G - I)\beta\beta'(G - I)' \\ &= \sigma^2(V\Delta^2V'+kI)^{-1}V\Delta^2V'(V\Delta^2V'+kI)^{-1}+k^2(V\Delta^2V'+kI)^{-1}\beta\beta'(V\Delta^2V'+kI)^{-1} \end{aligned} \tag{3.24}$$

Total Mean Squared Error

$$TMSE(\hat{\beta}_R) = \sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i + k)^2} + k^2\sum_{i=1}^{p}\frac{\beta_i^2}{(\lambda_i + k)^2} \tag{3.25}$$

3.3.2.4 Other Properties

- There always exists a positive constant k such that the $MSE[\hat{\beta}_R]$ is minimized (Hoerl and Kennard, 1970a; Marquardt, 1970; Gruber, 1980). The Ridge Existence Theorem (Vinod and Ullah, 1981) states that in the presence of collinearity, there is always an arbitrary constant from which the MSE for ridge estimates may be computed to be less than that of least squares estimates. That is: there exists a constant k $[0 < k < (2\sigma^2/\beta'\beta)]$ such that

  $$MSE(\hat{\beta}_R) < MSE(\hat{\beta}) \tag{3.26}$$

  Hence in situations whereby the existence theorem holds true, ridge estimators are considered much more reliable than least squares estimators since minimal $MSE[\tilde{\beta}]$ of an estimator implies that the corresponding estimator is the closest to the true parameters.

- The ridge estimator is not invariant to scaling and other linear transformations.

- The bias of $\hat{\beta}_R$ is a function of the orientation of the unknown parameter vector β to the eigenvectors; the bias is minimized when $\beta = v_1$ and maximized when $\beta = v_p$; where v_1 and v_p represent the 1^{st} and the p^{th} right singular vectors of X (Newhouse and Oman, 1971; Gruber, 1980).

- The ridge estimator is equivalent to the augmented OLSE: where the augmentation is as follows:

$$\mathbf{Y}_{(n+p)\times 1} = \begin{bmatrix} Y \\ 0_p \end{bmatrix} \quad , \quad \mathbf{X}_{(n+p)\times p} = \begin{bmatrix} X \\ \sqrt{k}I_p \end{bmatrix}$$

where

X = the $(n \times p)$ standardized matrix of independent variables
Y = a $(n \times 1)$ vector of response observations
0_p = the $(p \times 1)$ null vector
I_p = a $(p \times p)$ identity matrix

$\sqrt{k}$ = the bias, added to each of the standardized explanatory variables, with no observable effect on the dependent variable (Gruber, 1980; Marquardt, 1970).

3.3.2.5 Selection of k, the biasing parameter

The ridge theory postulates that biasing parameter k should be chosen to minimize mean squared error of ridge estimators. However, there is no specific biasing parameter assured to yield good ridge estimates for all unknown coefficient vectors. The optimal k value is a function of the unknown parameters β and σ^2 and in practice, k should be estimated from the data or be determined subjectively (Gruber, 1980).

The literature suggests several ways in which the optimal shrinkage parameter may be chosen and these include the following:

Ridge Traces

A ridge trace is a graphical presentation or a plot of individual ridge coefficients ($\hat{\beta}_R$) versus the corresponding ridge constants (k), used as a guide for selecting the optimal ridge constant in a given problem (McDonald, 1980). The procedure was originally suggested by Hoerl and Kennard (1970a) to investigate a variety of k values and their impact upon changes in ridge coefficient estimates.

The ridge trace may be defined as a path through the likelihood space that provides an insight into the structure of the factor space and the sensitivity of the results to particular sets of data (Hoerl and Kennard, 2000). It may also be referred to as a two-dimensional plot of the ridge solutions against the corresponding k parameters in the interval [0,1] that serves to portray the complex interrelationships that exist between collinear prediction variables and their effects on the estimation of β (Hoerl and Kennard, 1970a).

The criteria used to examine a ridge trace include stability, magnitudes and sign changes of the estimated coefficients and the inflation of residual sum of squares (McDonald, 1980). All these are primarily subjective on the range of k plotted for the ridge trace. From the graph, the optimal biasing parameter k is selected at a point where the traces stabilize.

Disadvantages

- The exercise of running multiple ridge regression models with different k values is time consuming and may be tedious.
- There is always an uncertainty in determining the optimal k from the ridge traces. The optimal value is not obvious from the graph; a rough estimate is usually made, depending on the stability of the traces.

Other procedures

Besides the ridge traces, there are lots of mathematical equations provided in the statistical literature for estimation of the optimal k; some of the examples are outlined below. Note that $\hat{\beta}$ and $\hat{\alpha}$ represent the OLS coefficients for models (1.1) (standard classical model) and (1.5) (orthogonal linear model) respectively.

- **Hoerl and Kennard (1970a)**

$$k_{hk} = \frac{\hat{\sigma}^2}{\hat{\beta}^2_{max}} \qquad (3.27)$$

where

$\hat{\beta}^2_{max}$ = the square of the maximum or largest least squares coefficient.
$\hat{\sigma}^2$ = the least squares residual variance

Disadvantage

- k_{hk} depends on $\hat{\sigma}^2$ and $\hat{\beta}^2_{max}$ hence it is likely to be affected by collinearity. If the data are highly collinear, least squares coefficients tend to be extremely larger than the true values, leading to false or inaccurate estimation. Further, $\hat{\beta}_{max}$ becomes the most misleading and inappropriate coefficient to rely on.

- **Mallows (1973)**

Mallows proposed that k should be chosen to minimize the following function

$$C_L = \frac{SSR_k}{\hat{\sigma}^2} + 2trace(H_k) - (n-2) \tag{3.28}$$

where

$SSR_k = (Y - X\hat{\beta}_R)'(Y - X\hat{\beta}_R) =$ Sum of squared residuals using $\hat{\beta}_R$

$H_k = X(X'X + kI)^{-1}X'$

Disadvantages

- Identifying k that minimizes C_L may not be practically simple. The process could take too long and be tiresome.
- The act of minimizing C_L is vulnerable to erratic computations.

- **Hoerl, Kennard and Baldwin (1975)**

$$k_{hkb} = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \tag{3.29}$$

where

$\hat{\beta}'\hat{\beta}$ = the sum of squares of least squares regression coefficients excluding the constant.

p = the number of variables in the model, excluding the constant.

$\hat{\sigma}^2$ = the least squares residual variance.

Disadvantage

- Although k_{hkb} is simple to compute, it depends on $\hat{\sigma}^2$ and $\hat{\beta}^2_{max}$ both of which are unstable and unreliable in collinear designs of matrices.

- **McDonald and Galarneau (1975)**

McDonald and Galarneau suggested selection of k such that the squared length of its corresponding vector of coefficients equals the squared length of the true parameter. The suggested unbiased estimator of the squared length of the true parameter is defined as

$$Q = \hat{\beta}'\hat{\beta} - \hat{\sigma}^2 \sum_{i=1}^{p} \lambda_i^{-1} \tag{3.30}$$

Selection of k is based on the following rules:

- **Rule 1**

  Choose k=0 such that $\hat{\beta}_R = \hat{\beta}$. $\hat{\beta}$ is required for computation of Q.

- **Rule 2**

  Choose k=0 such that $\hat{\beta}_R = \hat{\beta}$ for $Q < 0$ otherwise choose k to satisfy $\hat{\beta}'_R\hat{\beta}_R = Q$

- **Rule 3**

  Choose $k = \infty$ so that $\hat{\beta}_R = 0$ for $Q < 0$ otherwise choose k to satisfy $\hat{\beta}'_R\hat{\beta}_R = Q$

Rules 2 and 3 are alternative default values of k when Q is negative. From our point of view. rules 2 and 3 are the same; the implementation rules for each are not clear.

Disadvantages

Although the McDonald and Galarneau procedure is one of the well known methods by which the ridge constant may be selected, the following disadvantages may be highlighted:

* In highly collinear designs, $\hat{\beta}'\hat{\beta}$ becomes unstable. implying that some parameters may deviate considerably from the mean. hence elements in $(\hat{\beta}_R)$ may be also have high variances (Gruber, 1980).

* $\hat{\beta}'_R\hat{\beta}_R = Q$ does not consider a lower bound of the sum of squared coefficients $(\hat{\beta}'\hat{\beta})$ and in some cases it may lead to negative values (Gruber, 1980).

* In highly collinear data, Q may be unstable as a result of the instability of the least squares solution.

- **Iteration** (Hoerl and Kennard, 1976)

  Shortly after introducing k_{hkb}, Hoerl and Kennard developed an iteration on k_{hkb} on the basis that the squared length of $\hat{\beta}$ $(\hat{\beta}'\hat{\beta})$ is large when X is collinear, hence k_{hkb} may potentially be too small. The iterative procedure is summarized below.

$$\begin{array}{lll} \hat{\beta}: & \ldots k_0 & = \dfrac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \\ \hat{\beta}_R(k_0): & \ldots k_1 & = \dfrac{p\hat{\sigma}^2}{[\hat{\beta}_R(k_0)]'[\hat{\beta}_R(k_0)]} \\ \hat{\beta}_R(k_1): & \ldots k_2 & = \dfrac{p\hat{\sigma}^2}{[\hat{\beta}_R(k_1)]'[\hat{\beta}_R(k_1)]} \\ \hat{\beta}_R(k_t): & \ldots k_{t+1} & = \dfrac{p\hat{\sigma}^2}{[\hat{\beta}_R(k_t)]'[\hat{\beta}_R(k_t)]} \end{array}$$

where

k_i = the estimate of k on the i^{th} iteration

$\hat{\beta}_R(k_i)$ = a vector of ridge coefficients for the i^{th} iteration, computed from the k_{i-1}.

t = the t^{th} iteration

$\hat{\beta}_R(k_i) : \dots k_{i+1}$ $\Rightarrow \hat{\beta}_R(k_i)$ is used to estimate k_{i+1}

Initially, $\hat{\beta}$ is used to estimate $k_0 = \dfrac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$. Then, k_0 is used in estimating the ridge coefficients in $\hat{\beta}_R(k_0)$, which in turn, are input in computation of k_1, so on and so forth. The sequence is terminated when

$$\frac{k_{i+1} - k_i}{k_i} \leq \delta = 20T^{1.3},$$

where $T = \dfrac{Trace(X'X)^{-1}}{p}$ (Hoerl and Kennard, 1980)

Disadvantage

- Iteration is a long process that is open to errors.

- **Lawless and Wang (1976)**

$$k_{lw} = p\hat{\sigma}^2 / \sum \hat{\alpha}_i^2 \lambda_i = \frac{\hat{\sigma}^2 p}{Trace(\hat{\beta}'X'X\hat{\beta})} \tag{3.31}$$

where the unknowns are described in the previous sections
Disadvantage

- k_{lw} depends on the least squares solution and may easily be affected by high or extreme collinearity.

- **Vinod (1976)**
The proposed biasing parameter is selected to minimize index of stability of the relative magnitudes of parameters (ISRM), defined as

$$ISRM(k) = \sum_{i=1}^{p} \left[\Big(p\frac{\lambda_i}{(\lambda_i + k)^2} \Big) / \Big(\sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} \Big) - 1 \right]^2 \tag{3.32}$$

Disadvantage

- Selection of k that minimizes ISRM(k) may be practically a tiresome exercise.

- **Hocking, Speed and Lynn (1976)**

$$k_{hsl} = \hat{\sigma}^2 \frac{\sum_{i=1}^{p}(\lambda_i \hat{\beta}_i)^2}{(\sum_{i=1}^{p} \lambda_i \hat{\beta}_i^2)^2} \tag{3.33}$$

Disadvantage

- k_{hsl} depends on the least squares solution and is likely to be impacted on by extreme collinearity.

- **Brown (1993)**

Brown made two suggestions for k, originating from the k_{lw} and k_{hkb}, hence we label them k_{lwm} and k_{hkbm} respectively.

$$k_{lwm} = \frac{(r-2)\hat{\sigma}^2 \sum \lambda_i}{r\hat{\beta}' X' X \hat{\beta}} \tag{3.34}$$

where r=rank(X)

$$k_{hkbm} = \frac{(r-2)\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \tag{3.35}$$

In order for 3.34 and 3.35 to be positive the condition $r(X) > 2$ has to hold.

Disadvantages

- Both k_{hkbm} and k_{lwm} have high chances of being affected by collinearity since $\hat{\sigma}^2$ and $\hat{\beta}$ are adversely impacted on by collinearity.

- **Kibria (2003)**

Kibria made the following three suggestions for k; based on the arithmetic mean (am), the geometric mean (gm) and the median (med) of $(\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2})$ respectively.

$$k_{am} = \frac{1}{p}\sum_{i=1}^{p}\left(\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}\right) \tag{3.36}$$

$$k_{gm} = \frac{\hat{\sigma}^2}{(\prod_{i=1}^{p} \hat{\alpha}_i^2)^{\frac{1}{p}}} \tag{3.37}$$

$$k_{med} = median\left(\frac{\hat{\sigma}^2}{\hat{\alpha}_1^2}, \ldots, \frac{\hat{\sigma}^2}{\hat{\alpha}_p^2}\right) \tag{3.38}$$

Disadvantage

- Extreme cases of collinearity are likely to have an adverse effect on k_{am} and k_{gm} since the two depend on the least squares solution.

3.3.3 Ridge regression procedure

Ridge regression generally works with centred and scaled matrices (§ 2.2.2) of independent variables so that the sum of squares and product matrices are in the correlation form (§ 2.2.2). The analysis procedure is usually carried out logically in the following subsequent steps:

- Center and scale the matrix of independent variables to standardize the measurement units and to remove possible collinearities that may involve the intercept. However, it is important to note that centring removes the constant from the regression model hence any model in which the constant plays a vital role may lack practical sense when the constant has been removed. Therefore, it is sometimes essential to scale and not center.
- For selection of k from the suggested formulae,
  - Compute Ordinary least squares solutions in terms of the centred and scaled matrices of independent variables.
  - Substitute the least squares solutions in at least one of the biasing parameter suggestions provided in the previous section or any other method, not discussed in this study.
  - Compute the ridge solution from the selected biasing parameter
- For selection of k from the ridge traces,
  - Compute ridge solutions for different k parameters; $k_i \geq 0$.
  - Plot the ridge estimates against the different k parameters (ridge traces) and select the optimal biasing parameter
- In this study we transform the solutions corresponding to the optimal k back to the original form (the unstandardized form) before computing any measures of efficiencies of estimators.

3.3.4 Application and accessibility of ridge regression programs

For years, ridge regression has had a considerable amount of application varying from all fields of research: marketing scoring models (Malthouse, 1999), molding conditions for thermosets (Talwar and Ashlock, 1970), price and production (Bettman, 1973), mortality and air pollution (McDonald and Schwing, 1973), agricultural research (Jeffery and McKinney, 1975) to mention a few.

In all applications, ridge regression has been proved to be important when dealing with estimation problems that arise out of collinearity. It produces excellent estimates that are relatively simple to calculate (Feig, 1978). The algorithm and computations for ridge regression are straight forward and can be made with simple modifications to a standard linear regression program. The procedure involves inversion of the matrix $X'X + kI$ instead of a near singular matrix $X'X$, where k is selected to remove the singularity, thus stabilizing the estimators of β (Thiart, 1990).

A lot of computer programs have been developed to undertake ridge regression analysis. For example:

- Hoerl (1959) coded a full ridge regression program in FORTRAN IV to compute the coefficients for 32 different biasing parameters.
- Bradley and McGann (1977) wrote RIDGEREG to improve the precision of regression estimates for nonorthogonal data. It calculates the standardized covariance matrix, its determinant and the regression coefficients for the parameter (k) varying from 0 to 0.5 in steps of 0.1.
- Gunst (1979) proposed the guide to an efficient programming of biased regression algorithms, taking advantage of the mathematical similarities among them.
- ε*RIDGE* was described by MIT (1975) as a program that implements the main ridge regression algorithm.
- Bolding and Houston (1974) wrote a Fortran program to compute ridge regression coefficients.
- RRIDGE was developed by Jain et al. (1977) as a Fortran IV program that handles up to 30 factors, 200 observations and 20 values of k. Its output includes the correlation matrix, eigenvalues and parameters for different values of the biasing factor, k.
- (Carmer and Hsieh (1979); Sinha and Hardy (1979)) described SAS macros for ridge regression computations.
- Bush (1980) coded a comprehensive FORTRAN IV program to compute ridge regression. The program computes parameters for various k parameters, including those of:
  - (Dempster et al., 1977)
  - (Hoerl et al., 1975)
  - (Kasarda and Shih, 1977)
  - (Lawless and Wang, 1976)
  - (McDonald and Galarneau, 1975)

Today, an option for ridge regression is available in most statistical software packages. For instance, Statistica provides an option for computation of ridge solutions; only k has to be specified. Also, Eviews and R are other statistical software programs with built-in functions that allow easy computations and programming.

3.4 Generalized Ridge Regression

The generalized ridge estimation is a shrinkage method, first proposed by Hoerl and Kennard (1970a) to improve on ridge regression. The improvement is explained by the following:

Previously,

$$\hat{\beta}_R = \left\{(\Delta^2 + kI)^{-1}\Delta^2\right\}\hat{\beta}$$

From the above function, we observe that only one value of k is used to shrink all the components of $\hat{\beta}$. Implying that the components associated with large eigenvalues may possibly be shrunk more than necessary, and those that are associated with small eigenvalues may be shrunk less than required. Hence, allowing the choice of different constants may signify an improvement.

On this basis, Hoerl and Kennard (1970a) introduced generalized ridge regression, for which the additive constants vary across the different components of $\hat{\beta}$. Therefore, the generalized ridge estimation method is defined as a general form of ridge regression for which the biasing parameter k is defined by a diagonal matrix of different constants $k_i's$, with components $k_i \geq 0$.

3.4.1 The estimator

Consider the orthogonal linear model (1.5) from which the least squares estimator of α is defined by

$$\hat{\alpha} = (Z'Z)^{-1}Z'Y = V'\hat{\beta} \qquad (from \quad 1.26)$$

The generalized ridge estimator is defined by:

$$\hat{\alpha}_{GR} = (Z'Z + K)^{-1}Z'Y \tag{3.39}$$

where

$$\mathbf{K} = \begin{bmatrix} k_1 & 0 & 0 & 0 \\ 0 & k_2 & 0 & \vdots \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & 0 & k_p \end{bmatrix} \qquad 0 < k_i, \qquad i = 1, \dots, p$$

Notice that unlike in ridge regression where k is the same for all diagonal entries, generalized ridge utilizes different $k_i's$.

3.4.1.1 Relation to Least Squares

In terms of the least squares estimator $\hat{\alpha}$, $\hat{\alpha}_{GR}$ is defined as :

$$\begin{aligned}\hat{\alpha}_{GR} &= (\Delta^2 + K)^{-1} Z'Y \\ &= \underbrace{(\Delta^2 + K)^{-1}\Delta^2}\quad \hat{\alpha} \\ &= \qquad \delta \qquad \hat{\alpha} \\ &= \overbrace{(I - K(\Delta^2 + K)^{-1})}\hat{\alpha}\end{aligned}$$

Hence, $\hat{\alpha}_{GR}$ is a shrinkage estimator, of which the shrinkage matrix is

$$\delta = \begin{bmatrix} \frac{\lambda_1}{\lambda_1+k_1} & 0 & 0 & 0 \\ 0 & \frac{\lambda_2}{\lambda_2+k_2} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & \frac{\lambda_p}{\lambda_p+k_p} \end{bmatrix}$$

Therefore, the i^{th} generalized ridge estimator may be expressed as:

$$\hat{\alpha}_{GR_i} = \frac{\lambda_i}{\lambda_i + k_i}\hat{\alpha}_i$$

and the i^{th} generalized ridge shrinkage factor may be defined by

$$\delta_i = \frac{\lambda_i}{\lambda_i + k_i}$$

3.4.1.2 Properties of generalized ridge estimators

Expectation

$$\begin{aligned}E[\hat{\alpha}_{GR}] &= (\Delta^2 + K)^{-1}(\Delta^2)\alpha = (I - K(\Delta^2 + K)^{-1})\alpha \\ &= \delta\alpha \end{aligned} \tag{3.40}$$

Bias

$$\begin{aligned}Bias[\hat{\alpha}_{GR}] &= (\delta - I)\alpha \\ &= -K(\Delta^2 + K)^{-1}\alpha\end{aligned} \tag{3.41}$$

Variance

$$Var(\hat{\alpha}_{GR}) = \sigma^2\delta(Z'Z)^{-1}\delta' = \sigma^2\delta\Delta^{-2}\delta' \qquad (using \quad 1.29)$$

$$= \sigma^2(\Delta^2 + K)^{-1}\Delta^2(\Delta^2 + K)^{-1} \tag{3.42}$$

Mean Squared Error

$$MSE(\hat{\alpha}_{GR}) = \sigma^2(\Delta^2 + K)^{-1}\Delta^2(\Delta^2 + K)^{-1} + (\delta - I)\alpha\alpha'(\delta - I)$$

$$= \sigma^2(\Delta^2 + K)^{-1}\Delta^2(\Delta^2 + K)^{-1} + K(\Delta^2 + K)^{-1}\alpha\alpha' K(\Delta^2 + K)^{-1} \tag{3.43}$$

Total Mean Squared Error

$$TMSE(\hat{\alpha}_{GR}) = \sum_{i=1}^{p}\left(\frac{\sigma^2\lambda_i}{(\lambda_i + k_i)^2} + \frac{(\alpha_i k_i)^2}{(\lambda_i + k_i)^2}\right)$$

$$= \sum_{i=1}^{p}\left(\frac{\sigma^2\lambda_i + \alpha_i^2 k_i^2}{(\lambda_i + k_i)^2}\right) \tag{3.44}$$

3.4.2 Suggestions for the optimal set of $k_i's$

The optimal set of components $k_i's$ leads to the minimum $MSE[\hat{\alpha}_{GR}]$. We provide the following suggested formulae for ideal $k_i's$.

* **Hoerl and Kennard (1970a;1970b;1980)**

$$k_{hk_i} = \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2} \tag{3.45}$$

* **Troskie and Chalton (1996)**

$$k_{tc_i} = \frac{(\lambda_i)}{((\lambda_i\hat{\alpha}_i^2/\hat{\sigma}^2) + 1)} \qquad = \frac{\lambda_i}{F_i + 1} \tag{3.46}$$

Where $\quad F_i = \lambda_i\hat{\alpha}_i^2/\hat{\sigma}^2$

Comment

The fact that k_{hk_i} and k_{tc_i} depend on least squares is a disadvantage since the least squares solution is unreliable and unstable in collinear designs of matrices.

3.5 Liu Estimation

Liu estimation was proposed by Liu (1993) as one of the shrinkage methods of estimation that may be used when least squares estimates are unsatisfactory. The method was developed on the basis that the two mostly common biased methods of estimation, Stein and ridge regression have the following drawbacks:

- **Stein:**
  Stein shrinks all components of $\hat{\beta}$ with the same factor hence the Stein estimator does not behave well in practice (Liu, 1993).

- **Ridge:**
  Computations of ridge shrinkage parameter using suggestions like those of Mcdonald and Galarneau (1975) and C_L criterion (Mallows, 1973) complicate estimation of k (Liu, 1993).

This shrinkage method was further investigated by Akdeniz and Kaciranlar, 1995; Gruber, 1998; Liu, 2003; Akdeniz, 2001; Arslan and Billor, 2000; Kaciranlar and Sakallioglu, 2001, and Kaciranlar et al., 1999 to mention a few. This series of investigations resulted in a lot of amendments and corrections and an example has been provided at the end of this section.

3.5.1 The Liu estimator

The Liu estimator is defined by the least squares solution to the following linear system:

$$Y^{**} = X^{**}\beta + \epsilon^{**} \tag{3.47}$$

where

$$\mathbf{Y}^{**}_{(n+p)\times 1} = \begin{bmatrix} Y \\ d\hat{\beta} \end{bmatrix}, \quad \mathbf{X}^{**}_{(n+p)\times p} = \begin{bmatrix} X \\ I \end{bmatrix}, \quad \epsilon^{**}_{(n+p)\times 1} = \begin{bmatrix} \epsilon \\ \epsilon^{*} \end{bmatrix}$$

where d is an arbitrary constant, within the range $0 < d < 1$.

Thus the Liu estimator is

$$\begin{aligned}\hat{\beta}_L &= (X'X + I)^{-1}(X'Y + d\hat{\beta}) \\ &= (X'X + I)^{-1}(X'X + dI)\hat{\beta} \\ &= (V\Delta^2V' + I)^{-1}(V\Delta^2V' + dI)\hat{\beta} \end{aligned} \tag{3.48}$$

Let $L = (V\Delta^2V' + I)^{-1}(V\Delta^2V' + dI)$
Then, $\hat{\beta}_L = L\hat{\beta}$; hence the Liu shrinkage matrix is defined by L.

3.5.1.1 Properties of the Liu estimator

Expectation

$$\begin{aligned} E[\hat{\beta}_L] &= L\beta \\ &= (V\Delta^2V' + I)^{-1}(V\Delta^2V' + dI)\beta \end{aligned} \tag{3.49}$$

Bias

$$\begin{aligned}Bias[\hat{\beta}_L] &= (L-I)\beta \\ &= \Big((V\Delta^2V'+I)^{-1}(V\Delta^2V'+dI)-I\Big)\beta \qquad (3.50)\end{aligned}$$

Variance

$$\begin{aligned}Var[\hat{\beta}_L] &= \sigma^2 L(X'X)^{-1}L' \\ &= \sigma^2(V\Delta^2V'+I)^{-1}(V\Delta^2V'+dI)V\Delta^{-2}V'\Big[(V\Delta^2V'+I)^{-1}(V\Delta^2V'+dI)\Big]' \qquad (3.51)\end{aligned}$$

Mean Squared Error

$$MSE[\hat{\beta}_L] = \sigma^2 L(X'X)^{-1}L' + (L-I)\beta\beta'(L-I)' \qquad (3.52)$$

Total Mean Squared Error

$$TMSE[\hat{\beta}_L] = \sigma^2\sum_{i=1}^{p}\frac{(\lambda_i+d)^2}{\lambda_i(\lambda_i+1)^2} + (d-1)^2\sum_{i=1}^{p}\frac{\beta_i^2}{(\lambda_i+1)^2} \qquad (3.53)$$

3.5.1.2 The choice of d

Liu 1993

Liu suggests the following criteria for selection of the parameter d.

* $$\hat{d}_{mm} = 1-\hat{\sigma}^2\frac{\sum_{i=1}^{p}1/\lambda_i(\lambda_i+1)}{\sum_{i=1}^{p}\hat{\alpha}_i^2/(\lambda_i+1)^2} \qquad (3.54)$$

$\hat{d}_{mm}$ is defined as a point where $MSE[\hat{\beta}_L]$ obtains the minimum hence the label mm. Liu 1993 uses the term 'the minimum MSE estimate' to refer to $\hat{d}_{mm}$.

Disadvantage

- $\hat{d}_{mm}$ depends on $\hat{\sigma}^2$ and $\hat{\alpha}_i^2$, both of which are adversely affected by collinearity.

* $$\hat{d}_{cl} = 1-\hat{\sigma}^2\frac{\sum_{i=1}^{p}1/(\lambda_i+1)}{\sum_{i=1}^{p}\lambda_i\hat{\alpha}_i^2/(\lambda_i+1)^2} \qquad (3.55)$$

$\hat{d}_{cl}$ is computed as the minimum of

$$C_L = \frac{SSR_d}{\hat{\sigma}^2} + 2tr(H_d) - n + 2$$

where SSR_d = Sum of squared residuals from estimation of $\hat{\beta}_L$ and

$H_d = X(X'X+I)^{-1}(X'X+dI)X'$

Disadvantage

- $\hat{d}_{cl}$ is more likely to be impacted on negatively by collinearity since it depends on the least squares solution.

$$* \qquad \hat{d}_k = 1 - \left\{ \frac{\sum_{i=1}^{p} \frac{\hat{\alpha}_i^2}{\lambda_i + 1} - \left\{ \left(\sum_{i=1}^{p} \frac{\hat{\alpha}_i^2}{\lambda_i + 1} \right)^2 - \hat{\sigma}^2 \sum_{i=1}^{p} \frac{\hat{\alpha}_i^2}{(\lambda_i + 1)^2} \sum_{i=1}^{p} \frac{1}{\lambda_i} \right\}^{\frac{1}{2}}}{\sum_{i=1}^{p} \frac{\hat{\alpha}_i^2}{(\lambda_i + 1)^2}} \right\} \qquad (3.56)$$

Disadvantages

- $\hat{d}_k$ is practically complicated and likely to be computed with errors.
- $\hat{d}_k$ depends on $\hat{\sigma}^2$ and $\hat{\alpha}_i^2$ both of which are impacted on by collinearity.

* **Iteration of** d_{mm}, d_{cl} *or* d_k

Liu also proposes iteration of either d_{mm}, d_{cl} *or* d_k. For example,
Iteration of d_{mm}

$$\begin{aligned}
\hat{\alpha}: \qquad d_0 &= 1 - \hat{\sigma}^2 \left\{ \sum_{i=1}^{p} \frac{1}{\lambda_i(\lambda_i + 1)} \Big/ \sum_{i=1}^{p} \frac{\hat{\alpha}_i^2}{(\lambda_i + 1)^2} \right\} \\
\hat{\beta}_L(d_0): \qquad d_1 &= 1 - \hat{\sigma}^2 \left\{ \sum_{i=1}^{p} \frac{1}{\lambda_i(\lambda_i + 1)} \Big/ \sum_{i=1}^{p} \frac{\hat{\beta}_{L(i)}^2(d_0)}{(\lambda_i + 1)^2} \right\} \\
\hat{\beta}_L(d_1): \qquad d_2 &= 1 - \hat{\sigma}^2 \left\{ \sum_{i=1}^{p} \frac{1}{\lambda_i(\lambda_i + 1)} \Big/ \sum_{i=1}^{p} \frac{\hat{\beta}_{L(i)}^2(d_1)}{(\lambda_i + 1)^2} \right\}
\end{aligned}$$

where $\hat{\beta}_L(d_i)$ = a vector of Liu coefficients estimated from the i^{th} d.
Initially, the least squares vector of coefficient $\hat{\alpha}$ is used to estimate d_0. Then, d_0 is used to estimate the components in $\hat{\beta}_L(d_0)$, which in turn, are input in computation of d_1, so on and so forth.

The sequence is only terminated when

$$\frac{d_i - d_{i+1}}{d_i} < 20T^{-1.3}$$

where $T = \dfrac{Trace(X'X)^{-1}}{p}$

Disadvantages

- Iteration is time consuming
- The mathematical functions that are being iterated are highly likely to be muddled in the process.

For simulation, we select $\hat{d}_{cl}$ and $\hat{d}_{mm}$ since both functions are easy to compute and time saving.

3.5.2 The modified Liu estimator $\left(\text{Liu-type estimator } (\hat{\beta}_{kd})\right)$

We note without further investigation that Liu (2003) extended the Liu estimation theory by introducing the estimator called the Liu-type estimator $\hat{\beta}_{kd}$. Liu 2003 argues that unlike $\hat{\beta}_L$, $\hat{\beta}_{kd}$ includes two constants, k and d so that k may not just be restricted to small values while d may be adjusted to reduce the bias that may be introduced by large values of k.

The improved estimate $\hat{\beta}_{kd}$ is derived as the least squares solution to the following:

$$Y^* = X^*\beta + \epsilon^* \tag{3.57}$$

where

$$\mathbf{Y}^*_{(n+p)\times 1} = \begin{bmatrix} Y \\ \left(\frac{-d}{\sqrt{k}}\right)\tilde{\beta} \end{bmatrix}, \qquad \mathbf{X}^*_{(n+p)\times p} = \begin{bmatrix} X \\ \sqrt{k}I \end{bmatrix}, \qquad \epsilon^*_{(n+p)\times 1} = \begin{bmatrix} \epsilon \\ \check{\epsilon} \end{bmatrix}$$

Thus the Liu-type estimator is

$$\hat{\beta}_{kd} = (X'X + kI)^{-1}(X'Y - d\tilde{\beta}) \tag{3.58}$$

where

k and d = parameters; $k > 0$ and $-\infty < d < \infty$ respectively and

$\tilde{\beta}$ = Any estimator of β.

Note that for $\hat{\beta}_L, 0 < d < 1$ whereas for $\hat{\beta}_{kd}, -\infty < d < \infty$.

If we substitute $\hat{\beta}$ for $\tilde{\beta}$, then

$$\begin{aligned} \hat{\beta}_{kd} &= (X'X + kI)^{-1}(X'X - dI)\hat{\beta} \\ &= (V\Delta^2V' + kI)^{-1}(V\Delta^2V' - dI)\hat{\beta} \end{aligned}$$

From the above equation, $(V\Delta^2V' + kI)^{-1}(V\Delta^2V' - dI)$ is the shrinkage matrix.

However, the fact that d is not restricted to small values is a huge drawback. If d is set to be a positive value close to ∞, the diagonal entries of $(V\Delta^2V' + kI)^{-1}(V\Delta^2V' - dI)$ approach $-\infty$ hence, $(V\Delta^2V' + kI)^{-1}(V\Delta^2V' - dI)$ does not qualify to be a shrinkage matrix. That is, the diagonal elements of $(V\Delta^2V' + kI)^{-1}(V\Delta^2V' - dI)$ do not fall within the range $0 < d_{sh} < 1$. Similarly, if d is a large negative value, the diagonal elements of $(V\Delta^2V' + kI)^{-1}(V\Delta^2V' - dI)$ become too large to be considered shrinkage factors.

Further, if we substitute $\hat{\beta}_R$ for $\tilde{\beta}$, then

$$\hat{\beta}_{kd} = (X'X + kI)^{-1}(X'Y - d\hat{\beta}_R)$$

$$= (X'X + kI)^{-1}X'Y - d(X'X + kI)^{-1}\hat{\beta}_R$$

$$= \hat{\beta}_R - d(X'X + kI)^{-1}\hat{\beta}_R$$

$$= (I - d(X'X + kI)^{-1})\hat{\beta}_R$$

By observation, negative values of d inflate $\hat{\beta}_R$ while positive values between 0 and 1 lead to shrinkage of $\hat{\beta}_R$.

From our view point, inflating and or shrinking $\hat{\beta}_R$ does not make sense. Ridge coefficients are among the mostly favoured and potentially accurate estimates when collinearity is the problem, hence, inflating or shrinking $\hat{\beta}_R$ only destroys the existing good qualities of the ridge coefficients.

- Inflating $\hat{\beta}_R$ is not a good procedure in that it forces the ridge coefficients to be too large relative to the true values.
- Also, from experience, $\hat{\beta}_R$ already shrinks $\hat{\beta}$ enough to correct for the problems attributed to collinearity. Hence, shrinking $\hat{\beta}_R$ implies that $\hat{\beta}$ is being shrunk more than necessary.

So far, we have only substituted $\hat{\beta}_R$ and $\hat{\beta}$ into 3.58 and the outcome of substitution is not convincing that $\hat{\beta}_{kd}$ is a good estimator. However, we note without further details that the following estimators have been practically investigated for substitution into 3.58 for $\tilde{\beta}$ and the results have shown that $\hat{\beta}_{kd}$ can be regarded a good estimator:

- Ridge estimator ($\hat{\beta}_R$) (Liu, 2003),
- Principal component estimator ($\hat{\beta}_{pc}$) (Kaciranlar and Sakallioglu, 2001) and
- the M-estimator (Arslan and Billor, 2000).

Further investigation into the matter is beyond the scope of this thesis therefore we simply note the information for interest's sake.

3.6 Generalized Liu Estimation

The generalized Liu method of estimation is another shrinkage method, suggested by Liu (1993) as a general form of Liu estimation. The method follows the exact same procedure as Liu except that like in generalized ridge, generalized Liu method substitutes d by a diagonal matrix of $d_i's$, where each d_i estimates a single coefficient.

3.6.1 The estimator

The generalized Liu estimator is defined by the following

$$\begin{aligned}\hat{\beta}_{GL} &= (X'X + I)^{-1}(X'Y + D\hat{\beta}) \\ &= (X'X + I)^{-1}(X'X + D)\hat{\beta} \\ &= (V\Delta^2V' + I)^{-1}(V\Delta^2V' + D)\hat{\beta}\end{aligned}$$

where

$$D = \begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & \vdots \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & 0 & d_p \end{bmatrix} \qquad 0 < d_i < 1, \qquad i = 1, ..., p$$

Let $S = (V\Delta^2V' + I)^{-1}(V\Delta^2V' + D)$

Then, $\hat{\beta}_{GL} = S\hat{\beta}$ therefore S is the shrinkage matrix for generalized Liu estimation.

This implies that each i^{th} element of $\hat{\beta}_{GL}$ is shrunk by

$$\frac{\lambda_i v_i v_i' + d_i}{\lambda_i v_i v_i' + 1}$$

3.6.1.1 Properties of $\hat{\beta}_{GL}$

Expectation

$$\begin{aligned}E[\hat{\beta}_{GL}] &= S\beta \\ &= (V\Delta^2V' + I)^{-1}(V\Delta^2V' + D)\beta \qquad (3.59)\end{aligned}$$

Bias

$$\begin{aligned}Bias[\hat{\beta}_{GL}] &= [S - I]\beta \\ &= \Big((V\Delta^2V' + I)^{-1}(V\Delta^2V' + D) - I\Big)\beta \qquad (3.60)\end{aligned}$$

Variance

$$\begin{aligned}Var[\hat{\beta}_{GL}] &= \sigma^2 SV\Delta^{-2}V'S' \\ &= \sigma^2(V\Delta^2V' + I)^{-1}(V\Delta^2V' + D)(V\Delta^{-2}V')\Big((V\Delta^2V' + I)^{-1}(V\Delta^2V' + D)\Big)' \qquad (3.61)\end{aligned}$$

Mean Squared Error

$$MSE[\hat{\beta}_{GL}] = \sigma^2 SV\Delta^{-2}V'S' + [S-I]\beta\beta'[S-I]' \tag{3.62}$$

Total Mean Squared Error

$$TMSE[\hat{\beta}_{GL}] = \sigma^2 \sum_{i=1}^{p} \frac{(\lambda_i + d_i)^2}{\lambda_i(\lambda_i+1)^2} + \sum_{i=1}^{p} \frac{(d_i - 1)^2\beta_i^2}{(\lambda_i+1)^2} \tag{3.63}$$

3.6.1.2 Selection of d_i

Liu (1993) suggests the following function for estimating the optimal set of components of d_i.

$$d_i = 1 - \hat{\sigma}^2 \frac{(\lambda_i + 1)}{\lambda_i \hat{\alpha}_i^2} \tag{3.64}$$

3.7 Principal Components Regression

Principal components regression (Kendall, 1957) is one of the most widely used shrinkage methods of estimation. The method entails deletion of the dimensions of the X-space that cause dependencies among the independent variables. The principal components of $X = U\Delta V'$ are defined as linear functions of the independent variables specified by the column vectors of V.

Consider the orthogonal linear model (1.5)
$Y = Z\alpha + \epsilon$
where

- the least squares estimator of α is $\hat{\alpha} = (Z'Z)^{-1}Z'Y$ and
- each column of Z represents one of the principal components.

Suppose we partition Z into $[Z_a : Z_b]$ such that

- Z_a is a matrix of dimension $(n \times (p-m))$, containing p-m principal components corresponding to p-m largest eigenvalues
- Z_b is a matrix of dimension $(n \times m)$, containing m principal components corresponding to m smallest eigenvalues.

We can the rewrite equation 1.5 as follows

$$\begin{aligned}\mathbf{Y} &= \begin{bmatrix} Z_a & Z_b \end{bmatrix} \begin{bmatrix} \alpha_a \\ \alpha_b \end{bmatrix} + \epsilon \\ &= Z_a\alpha_a + Z_b\alpha_b + \epsilon\end{aligned}$$

where
α_a = a vector of dimension $((p-m)\times 1)$, corresponding to Z_a
α_b = a vector of dimension $(m\times 1)$, corresponding to Z_b.

From the relationship $Z = XV$ (*section* 1.2.1), V can also be partitioned into

$$\mathbf{V} = \begin{bmatrix} V_a & V_b \end{bmatrix}$$

such that

$$\mathbf{Z} = XV = \underbrace{X}_{(n\times p)} \begin{bmatrix} \underbrace{V_a}_{(p\times(p-m))} & \underbrace{V_b}_{(p\times m)} \end{bmatrix}$$

Principal components regression entails deletion of the m principal components associated with the dimensions of X that cause collinearity. Suppose we set the last m eigenvalues to zero, it then follows that $Z_b\alpha_b = 0$ thus $\alpha_b = 0$. Hence for the p-m remaining components, the least squares estimate of α_a (principal component estimator of α_a) becomes:

$$\hat{\alpha}_a = (Z_a'Z_a)^{-1}Z_a'Y$$

3.7.1 The estimator

The principal components estimator $(\hat{\beta}_{pc})$ is defined by the following

$$\begin{aligned}\hat{\beta}_{pc} &= V_a\hat{\alpha}_a\\ &= V_aV_a'\hat{\beta} \qquad (3.65)\end{aligned}$$

hence $\hat{\beta}_{pc}$ is a shrinkage estimator of which the shrinkage matrix is V_aV_a'.

Comment
Unlike other shrinkage matrices/factors discussed in the preceding sections, the shrinkage matrix for $\hat{\beta}_{pc}$ does not depend on the unknown values. This is an important quality of $\hat{\beta}_{pc}$ that even makes principal components regression more reliable and easier to deal with.

3.7.1.1 Properties of $\hat{\beta}_{pc}$

Expectation

$$E[\hat{\beta}_{pc}] = V_aV_a'\beta \qquad (3.66)$$

Bias

$$\begin{aligned} Bias[\hat{\beta}_{pc}] &= (V_aV_a' - I)\beta \\ &= -V_bV_b'\beta \end{aligned} \qquad (3.67)$$

Variance

$$Var[\hat{\beta}_{pc}] = \sigma^2 V_aV_a'(X'X)^{-1}V_aV_a' \qquad (using \quad 3.4)$$

But

$$\begin{aligned} (X'X)^{-1} &= V\Delta^{-2}V' \\ &= \left[\begin{array}{cc} V_a & V_b \end{array}\right] \left[\begin{array}{cc} \Delta_a & 0 \\ 0 & \Delta_b \end{array}\right]^{-2} \left[\begin{array}{c} V_a' \\ V_b' \end{array}\right] \\ &= V_a\Delta_a^{-2}V_a' + V_b\Delta_b^{-2}V_b' \end{aligned}$$

Therefore

$$\begin{aligned} Var[\hat{\beta}_{pc}] &= \sigma^2 V_aV_a'[V_a\Delta_a^{-2}V_a' + V_b\Delta_b^{-2}V_b']V_aV_a' \\ &= \sigma^2(V_a\Delta_a^{-2}V_a') \qquad (since \quad V_a \quad and \quad V_b \quad are \quad orthogonal.) \end{aligned} \qquad (3.68)$$

Mean Squared Error

$$MSE[\hat{\beta}_{pc}] = \sigma^2(V_a\Delta_a^{-2}V_a') + V_bV_b'\beta\beta'V_bV_b' \qquad (3.69)$$

Total mean squared error

$$\begin{aligned} TMSE[\hat{\beta}_{pc}] &= \sigma^2 \sum_{i=1}^{p-m} \frac{1}{\lambda_i} + \beta'V_bV_b'V_bV_b'\beta \\ &= \sigma^2 \sum_{i=1}^{p-m} \frac{1}{\lambda_i} + \beta'V_bV_b'\beta \qquad since \quad V_b'V_b = I \end{aligned} \qquad (3.70)$$

The principal components regression procedure includes the following:

- Estimation of the least squares estimate of α to assess significance of different variables.

- Deletion of the principal components corresponding to the smallest eigenvalue(s)

- Least squares estimation of the remaining components.

3.7.2 Criteria for eliminating the principal components

The following decisive factors are vital in selection/deletion of the principal components.

- *Small eigenvalues*

  The principal components associated with eigenvalues that are near zero should be deleted. If the eigenvalue is close to zero, then the corresponding coefficient has a large variance (Rawlings et al., 1998).

- *Significance of individual components*

  The individual components should be tested for significance. Less significant components should be eliminated (Brown, 1993). The classical F test may be used to evaluate the hypothesis that $Z_b\alpha_b = 0$ (Thiart, 1990; Hill et al., 1977). Kendall (1957) recommends usage of t-tests to test the significance of the components.

- *Correlation between the response variable and the components.*

  Components that are significantly correlated with the response variable should not be eliminated. Graphs may be used to determine the kind of correlation between the components and the response variable; a nearly perfect or a perfect linear relationship implies the importance of the corresponding principal component in the model.

- *High variance*

  The principal components with high variance should be retained (Jolliffe, 1982). Components with small variance are unlikely to be important in regression (Mosteller and Tukey, 1977; Gunst and Mason, 1980). However, Jeffers (1967, p.230) argues that the components with small variances may possibly be highly correlated with the response variable hence turn out to be important in the model.

- *Small prediction error*

  The principal components for which the regression model has the minimum prediction error should be retained (Brown, 1993).

- *Small mean squared error*

  The principal components for which the regression model has the minimum MSE error should be retained (Hill et al., 1977).

3.8 Summary

In this chapter, we defined shrinkage estimation as a family of biased estimation techniques of which the error risk is lower than that of least squares estimation when collinearity is a problem.

We brought together Stein estimation, ridge regression, generalized ridge, Liu, generalized Liu and principal components regression into a common framework of shrinkage estimation. We characterized each of these methods by a unique shrinkage factor or matrix. Further, the properties of each estimator were specified in line with the general properties of the shrinkage estimator.

To wrap up the chapter, we summarize the properties underlying the discussed shrinkage estimators in tables 3.1, 3.2 and 3.3. We provide tables to summarize the shrinkage factors, TMSE's, bias, expectations and the variances of the shrinkage methods considered in this chapter.

	Shrinkage Factor / Matrix	**TMSE**
$\hat{\beta}_s$	$0 \leq c \leq 1$	$c^2\sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} + (c-1)^2 \sum_i^p \beta_i^2$
$\hat{\beta}_R$	$(V\Delta^2V' + kI)^{-1}V\Delta^2V'$, $k > 0$	$\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2} + k^2 \sum_{i=1}^{p} \frac{\beta_i^2}{(\lambda_i+k)^2}$
$\hat{\beta}_{GR}$	$(\Delta^2 + K)^{-1}\Delta^2$	$\sum_{i=1}^{p} \left(\frac{\sigma^2\lambda_i}{(\lambda_i+k_i)^2} + \frac{(\alpha_i k_i)^2}{(\lambda_i+k_i)^2} \right)$
$\hat{\beta}_L$	$(V\Delta^2V' + I)^{-1}(V\Delta^2V' + dI)$, $0 < d < 1$	$\sigma^2 \sum_{i=1}^{p} \frac{(\lambda_i+d)^2}{\lambda_i(\lambda_i+1)^2} + (d-1)^2 \sum_{i=1}^{p} \frac{\beta_i^2}{(\lambda_i+1)^2}$
$\hat{\beta}_{GL}$	$(V\Delta^2V' + I)^{-1}(V\Delta^2V' + D)$	$\sigma^2 \sum_{i=1}^{p} \dfrac{(\lambda_i + d_i)^2}{\lambda_i(\lambda_i + 1)^2} + \displaystyle\sum_{i=1}^{p} \dfrac{(d_i - 1)^2\beta_i^2}{(\lambda_i + 1)^2}$
$\hat{\beta}_{pc}$	V_aV_a'	$\sigma^2 \sum_{i=1}^{p-m} \frac{1}{\lambda_i} + \beta' V_b V_b' \beta$

Table 3.1: Shrinkage factors and TMSE's

	Expectation	Bias
$\hat{\beta}_{sh}$	$d_{sh}\beta$	$(d_{sh}-1)\beta$
$\hat{\beta}_s$	$c\beta$	$(c-1)\beta$
$\hat{\beta}_R$	$(V\Delta^2V'+kI)^{-1}V\Delta^2V'\beta$	$\left[(V\Delta^2V'+kI)^{-1}V\Delta^2V'-I\right]\beta$
$\hat{\beta}_{GR}$	$(\Delta^2+K)^{-1}(\Delta^2)\alpha$	$\left((\Delta^2+K)^{-1}(\Delta^2)-I\right)\alpha$
β_L	$(V\Delta^2V'+I)^{-1}(V\Delta^2V'+dI)\beta$	$\left((V\Delta^2V'+I)^{-1}(V\Delta^2V'+dI)-I\right)\beta$
β_{GL}	$(V\Delta^2V'+I)^{-1}(V\Delta^2V'+D)\beta$	$\left(\left[(V\Delta^2V'+I)^{-1}(V\Delta^2V'+D)\right]-I\right)\beta$
β_{pc}	$V_aV_a'\beta$	$-V_bV_b'\beta$

Table 3.2: Expected values and Bias expressions for the biased estimators

	Variance
$\hat{\beta}_{sh}$	$d_{sh}^2\sigma^2V\Delta^{-2}V'$
$\hat{\beta}_s$	$c^2\sigma^2V\Delta^{-2}V'$
$\hat{\beta}_R$	$\sigma^2(V\Delta^2V'+kI)^{-1}V\Delta^2V'(V\Delta^2V'+kI)^{-1}$
$\hat{\beta}_{GR}$	$\sigma^2(\Delta^2+K)^{-1}\Delta^2(\Delta^2+K)^{-1}$
β_L	$\sigma^2(V\Delta^2V'+I)^{-1}(V\Delta^2V'+dI)V\Delta^{-2}V'\left[(V\Delta^2V'+I)^{-1}(V\Delta^2V'+dI)\right]'$
β_{GL}	$\sigma^2(V\Delta^2V'+I)^{-1}(V\Delta^2V'+D)(V\Delta^{-2}V')\left((V\Delta^2V'+I)^{-1}(V\Delta^2V'+D)\right)'$
β_{pc}	$\sigma^2(V_a\Delta_a^{-2}V_a')$

Table 3.3: Variances of biased estimators

Chapter 4

A new approach to ridge shrinkage estimation

Ridge and generalized ridge estimation methods are among the mostly favoured and practically convenient shrinkage methods of estimation, vital when the OLS estimates are unreliable. For this reason, it is inevitably crucial to review the methods and identify potential improvement tactics as the need arises.

From previous discussions, it should be observed that there is a variety of suggestions for ridge shrinkage factors all of which are said to play a vital role in ridge regression. For each unique ridge shrinkage factor/matrix, there exists a corresponding ridge shrinkage method, the reliability of which is wholly determined by the shrinkage factor/matrix. In other words, each shrinkage factor/matrix determines the goodness and effectiveness of the corresponding ridge shrinkage method and each method is differentiated from the rest by the shrinkage factor/matrix. Therefore, it is critically important to look into the best ways in which the shrinkage factors may be estimated.

There is a heated controversy in application of ridge shrinkage methods; different suggestions are emerging in the literature to improve on the existing concepts but there is still no particular ridge shrinkage method, proven to be generally superior. Noteworthy is the fact that all the ridge shrinkage factors are currently dependent on the least squares regression coefficients and/or variance. This, from our point of view, is a huge drawback that directly impacts on performance of ridge estimators since least squares is unreliable when collinearity is present in the data.

In this chapter, we propose a new convenient method for estimating the ridge and generalized ridge shrinkage factors/matrices. First, we review the current standard criteria through which the shrinkage factors are being estimated and outline the potential hazard of these criteria on the existing methods. To conclude the chapter, we suggest a new approach for estimation of ridge and generalized ridge shrinkage factors.

4.1 The current procedure

From the previous chapter, we note that the ridge and generalized ridge shrinkage factors depend on the unknown parameters which have to be chosen or estimated from the data. Usually, the unknowns are estimated using the OLS solutions. Specifically, the coefficients and the variance estimates from ordinary least squares estimation are used to estimate the ridge and generalized ridge shrinkage matrices.

4.1.1 Ridge

The ridge shrinkage matrix was previously specified to be $G = WX'X = (V\Delta^2V'+kI)^{-1}V\Delta^2V'$ (from 3.18), where k is an unknown parameter. From a range of suggestions for estimation of k provided in chapter 3, we observe that $\hat{\sigma}^2$ and ($\hat{\alpha}$ or $\hat{\beta}$) are essential in finding the unknown k. That is

Label	Expression	Equation number
k_{hk}	$\dfrac{\hat{\sigma}^2}{\hat{\beta}^2_{max}}$	3.27
C_L	$\dfrac{SSR_k}{\hat{\sigma}^2} + 2trace(H_k) - (n-2)$	3.28
k_{hkb}	$\dfrac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$	3.29
Q	$\hat{\beta}'\hat{\beta} - \hat{\sigma}^2\sum_{i=1}^{p}\lambda_i^{-1}$	3.30
k_{lw}	$p\hat{\sigma}^2/\sum\hat{\alpha}_i^2\lambda_i$	3.31
k_{hsl}	$\hat{\sigma}^2\dfrac{\sum_{i=1}^{p}(\lambda_i\hat{\beta}_i)^2}{(\sum_{i=1}^{p}\lambda_i\hat{\beta}_i^2)^2}$	3.33
k_{lwm}	$\dfrac{(r-2)\hat{\sigma}^2\sum\lambda_i}{r\hat{\beta}'X'X\hat{\beta}}$	3.34
k_{hkbm}	$\dfrac{(r-2)\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$	3.35
k_{am}	$\dfrac{1}{p}\sum_{i=1}^{p}\left(\dfrac{\hat{\sigma}^2}{\hat{\alpha}_i^2}\right)$	3.36
k_{gm}	$\dfrac{\hat{\sigma}^2}{(\prod_{i=1}^{p}\hat{\alpha}_i^2)^{\frac{1}{p}}}$	3.37
k_{med}	$median\left(\dfrac{\hat{\sigma}^2}{\hat{\alpha}_1^2},\ldots,\dfrac{\hat{\sigma}^2}{\hat{\alpha}_p^2}\right)$	3.38

Table 4.1: Traditional ridge constants

4.1.2 Generalized ridge

Previously, we specified the following shrinkage matrix for generalized ridge regression

$$\delta = (\Delta^2 + K)^{-1}\Delta^2$$

where K is a diagonal matrix of unknown $k_i's$ that have to be estimated. Again, from the suggested functions provided in the previous chapter, $k_i's$ depend on the OLS solution.

Label	Expression	Equation number
k_{hk_i}	$\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}$	3.45
k_{tc_i}	$\frac{(\lambda_i)}{((\lambda_i\hat{\alpha}_i^2/\hat{\sigma}^2)+1)}$	3.46

Table 4.2: Traditional generalized ridge constants

4.1.3 The hazard of the current procedure

Estimation of ridge and generalized ridge shrinkage factors from the OLS solution has been in application since 1970 when the corresponding estimation methods were first introduced. The results have always been and are still the better, compared to those of OLS when collinearity is a problem. However, without ruling out this fact, it is critical to take cognisance of the potential harm that the OLS solution has on ridge and generalized ridge (in terms of calculating the biasing constants k/k_i).

Without unnecessarily repeating the details provided in chapter 2, we emphasize that least squares estimation is highly influenced by collinearity; the coefficients tend to be too large, extremely sensitive, unstable and even bear wrong signs. Even more important is the fact that the variances and standard errors of estimates are inflated hence the estimates deviate significantly from the true values (imprecision of the estimates) when the data are collinear. It should be understood that imprecision of estimates imply that the estimates are vague and not likely to provide the correct information. Hence, it follows logically that ridge and generalized ridge solutions of which the shrinkage factors are dependent on the OLS solution are vulnerable to collinearity.

From our point of view, the instability of the OLS solution is highly likely to result in erratic ridge and generalized ridge shrinkage estimates. This is one aspect that has not received attention but which has the potential to perturb the ridge estimation methods. In view of this, we suggest a new procedure for estimating the shrinkage factors independent from the OLS solution.

4.2 The new procedure

We propose that the solution to principal components, setting the smallest root to zero should rather be substituted for the OLS solution in estimation of the shrinkage factors/matrices.

We use the following criteria to select the alternative for OLS.

- Robustness to collinearity:
  We select a procedure that is more stable than OLS and less likely to be impacted on by collinearity. Like any other shrinkage estimation procedure, principal components regression deleting at least one smallest root is a remedy for collinearity and is less likely to be adversely effected by collinearity. By eliminating the dimensions of the X-space that are causing the problem, principal components regression removes collinearity instantly (chapter 3).

- Non-dependence on OLS solution:
  We choose a procedure that does not depend on the OLS solution. Unlike most of the shrinkage estimators of which shrinkage factors require the OLS solution, the principal components shrinkage matrix does not rely on the OLS solution.

Since principal components regression meets both criteria, we consider it the most eligible procedure to substitute the OLS solution in estimation of the ridge and generalized shrinkage factors. However the following points should be taken into consideration.

- We suggest and emphasize deletion of the smallest root because the stability of principal components estimates is observed when the extremely small singular values are removed from the regression model. We are not proposing deletion of more than one roots because most often, the principal components regression solution stabilizes after deletion of the smallest root. Nonetheless, we cannot generalize and rule out the possibility that principal component regression deleting more than one roots could also be substituted for OLS in estimation of the shrinkage factors. Where necessary, more than one smallest roots may be set to zero; the important issue is to avoid estimation of shrinkage factors from the OLS solution when the data are collinear.

- If zero singular values are eliminated, the principal components regression and OLS give similar results, hence the proposed substitution would not make sense. The principal components estimator becomes a shrinkage estimator only when at least one smallest singular value is removed.

The proposed procedure imposes changes in the functions for the unknown parameters used to estimate the ridge and generalized ridge shrinkage factors. We illustrate the changes below and assess performance of the new ridge and generalized ridge estimators (those that are based on the new procedure) in the simulation study (chapter 5). Note that the subscript 'pcdel1' implies the use of the solution for principal components regression, setting the smallest singular value to zero.

4.2.1 Ridge

Table 4.3 presents the functions for the new proposed parameters for the ridge shrinkage factor.

Label	**Expression**
k_{hk-new}	$\frac{\hat{\sigma}^2_{pcdel1}}{\hat{\beta}_{pcdel1}{}^2_{max}}$
C'_{L-new}	$\frac{SSR_k}{\hat{\sigma}^2_{pcdel1}} + 2trace(H_k) - (n-2)$
$k_{hkb-new}$	$\frac{p\hat{\sigma}^2_{pcdel1}}{\hat{\beta}'_{pcdel1}\hat{\beta}_{pcdel1}}$
Q_{new}	$\hat{\beta}'_{pcdel1}\hat{\beta}_{pcdel1} - \hat{\sigma}^2_{pcdel1}\sum_{i=1}^{p}\lambda_i^{-1}$
k_{lw-new}	$p\hat{\sigma}^2_{pcdel1}/\sum_{i=1}^{p-m}\hat{\alpha}^2_{pcdel1_i}\lambda_i$
$k_{hsl-new}$	$\hat{\sigma}^2_{pcdel1}\frac{\sum_{i=1}^{p}(\lambda_i\hat{\beta}_{pcdel1_i})^2}{(\sum_{i=1}^{p}\lambda_i\hat{\beta}^2_{pcdel1_i})^2}$
$k_{lwm-new}$	$\frac{(r-2)\hat{\sigma}^2_{pcdel1}\sum_{i=1}^{p-m}\lambda_i}{r\hat{\beta}'_{pcdel1}X'X\hat{\beta}_{pcdel1}}$ m=the number of the eliminated eigenvalues
$k_{hkbm-new}$	$\frac{(r-2)\hat{\sigma}^2_{pcdel1}}{\hat{\beta}'_{pcdel1}\hat{\beta}_{pcdel1}}$
k_{am-new}	$\frac{1}{(p-m)}\sum_{i=1}^{p-m}\left(\frac{\hat{\sigma}^2_{pcdel1}}{\hat{\alpha}^2_{pcdel1_i}}\right)$
k_{gm-new}	$\frac{\hat{\sigma}^2_{pcdel1}}{(\prod_{i=1}^{p-m}\hat{\alpha}^2_{pcdel1_i})^{\frac{1}{p-m}}}$
$k_{med-new}$	$median\left(\frac{\hat{\sigma}^2_{pcdel1}}{\hat{\alpha}^2_{pcdel1_1}},\ldots,\frac{\hat{\sigma}^2_{pcdel1}}{\hat{\alpha}^2_{pcdel1_{p-m}}}\right)$

Table 4.3: New ridge constants

4.2.2 Generalized ridge

Table 4.4 present the new proposed generalized ridge parameters:

$$\delta = I - K(\Delta^2 + K)^{-1} = (\Delta^2 + K)^{-1}\Delta^2$$

Label	Expression
k_{hk-new_i}	$\frac{\hat{\sigma}^2_{pcdel1}}{\hat{\alpha}^2_{pcdel1_i}}$
k_{tc-new_i}	$\frac{(\lambda_i)}{((\lambda_i\hat{\alpha}^2_{pcdel1_i}/\hat{\sigma}^2_{pcdel1}) + 1)}$

Table 4.4: New generalized ridge constants

A crucial point to take note of is that for

$$k_{hk-new_i} = \frac{\hat{\sigma}^2_{pcdel1}}{\hat{\alpha}^2_{pcdel1_i}} \tag{4.1}$$

the condition $\hat{\alpha}_{pcdel1_p} = 0$ leads to $k_{hk-new_p} = \infty$.

Therefore we propose substitution of $min(k_{hk-new_1}, \cdots k_{hk-new_{p-1}})$ for $k_{hk-new_p} = \infty$ (the p^{th} element of the shrinkage matrix) to avoid computational complications. Hence the proposed matrix shrinkage is the following

$$\mathbf{K} = \begin{bmatrix} k_{hk-new_1} & 0 & 0 & 0 \\ 0 & k_{hk-new_2} & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & \vdots & 0 & min(k_{hk-new_1}, \cdots k_{hk-new_{p-1}}) \end{bmatrix}$$

We investigate performance of the estimators associated with the above stated biasing factors in a simulation study; details are provided in chapter 5. We compare the new estimators with their corresponding known estimators '(old)' and the rest of other shrinkage estimators considered in chapter 3.

4.3 Summary

A new approach to estimation of the unknowns for ridge and generalized ridge shrinkage matrices was presented in this chapter. The new approach entails substitution of the OLS solution by the solution from principal components regression, setting the smallest root to zero. The new method is expected to improve on the existing one since it depends on a stable solution.

Chapter 5

The Simulation study

Shrinkage estimation has received a wide application in statistical research. Hence, a lot of papers are available in a variety of journals where shrinkage methods are proven to be effective in dealing with the problems attributed to collinearity. An extensive application of shrinkage methods over a range of problems is considered a necessity to judge the value of the emerging improvements over the old shrinkage methods.

In this chapter, we review some of the past simulation studies and present a simulation study in which we compare performances of 24 biased estimators relative to the OLSE. The estimators consist of 2 principal components estimators (deleting one and two roots), the Stein estimator of James and Stein (1961), the generalized Liu estimator and 2 Liu estimators suggested by Liu (1993), 14 ridge estimators; 7 of which are based on the new proposed method and 4 generalized ridge estimators; of which 2 are based on the new method proposed in this study. The chapter is structured to provide:

(1) *Past simulation studies (§ 5.1)*:
We review and summarize some of the past simulation studies on comparison of shrinkage methods.

(2) The distinction between this study and past simulation studies is drawn in (§ 5.2).

(3) *The design of the simulation study (§ 5.3)*:
We present our simulation study in the following manner

- The structure of the X matrix (§ 5.3.1),
- The collinearity level of X (§ 5.3.1.1),
- The structure of Y (§ 5.3.2),
- Distributions of error terms used to generate Y (§ 5.3.2.1)

(3) The basis for comparison of estimates (§ 5.4)

(4) The simulation program (§ 5.5)

5.1 Past simulation studies

We reviewed 16 simulation studies on shrinkage estimators. For convenience, these studies are summarized in Appendix D.

In all the studies

- OLS was used as a yardstick
- Measure of effectiveness ranges from MSE, TMSE, residual prediction error, relative efficiencies and Pitman measures (column 2 of table D1).

In these comparisons, all the authors had at least one of the 'ridge family' estimators in the list of estimators compared. Thiart (1994) had the most comprehensive list while FU (1998) considered one of the recently introduced special cases of ridge, lasso.

The authors reported that ridge performs the best. However, Thiart (1994) reported that ridge estimators outperform OLSE, but that it was not necessarily the best; the author could not identify a unique 'best' shrinkage estimator.

5.2 Why is this study unique?

In the previous simulation studies very little has been done on performance of shrinkage estimators across different error distributions. In almost all the reported studies, the error terms were always assumed normal. We carry out a simulation study in line with Thiart (1994) and Thiart et al. (1993) and identify the following features that make our study different from previous studies.

- We consider a range of error distributions and different variance levels; four distributions and three different variance levels.
- We propose a new method for estimation of shrinkage factors/matrix.
- We focus our attention on performance of 24 shrinkage estimators, including 6 of the 13 investigated by Thiart et al. (1993) and Thiart (1994).
- We use a different method of data generation. For a linear model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \epsilon_i \qquad i = 1 \ldots 30$$

both Thiart et al. (1993) and Thiart (1994) employed the Mcdonald and Garlarneau (1975) data generation method in which

$$x_{ij} = (1-\alpha_1^2)^{\frac{1}{2}} z_{ij} + \alpha_1 z_{i6} \qquad i = 1, \ldots 30 \quad and \quad j = 1, 2, 3$$

and

$$x_{ij} = (1-\alpha_2^2)^{\frac{1}{2}} z_{ij} + \alpha_2 z_{i6} \qquad i = 1, \ldots, 30 \quad and \quad j = 4, 5$$

where

$z_{ij}'s$ =N(0,1) independent variables
x_{ij} =the i^{th} element of the j^{th} column of X
$\alpha_i' s$ =parameters that determine the level of dependencies among the independent variables.

However, we specify integers for X_1 and X_3 and compute the other three columns as combinations of the two (§ 5.3). Further, we specify one arbitrary vector of $\beta_i' s$ whereas Thiart (1994) selected two β_i' eigenvectors of $X'X$ corresponding to the smallest and the largest eigenvalues.

5.3 The design of the simulation study

5.3.1 Generating the X matrix

The X matrix is an extension of a small illustrative example in Rawlings et al. (1998, p.372). We extend Rawling's (20 x 4) matrix to a (100 x 6) matrix of independent variables.
The X matrix is generated as follows:

For n=1,....,100

- The first column consists of ones, thus
  $X_1 = 1$
- Column 2 (X_2) is a sequence of numbers from 20:29 with an increment of 1, repeated to make 100 observations.
- Column 3 is column 2 with 25 subtracted from it and observations 1 and 11 changed to -4 to avoid direct collinearity.
  $X_3 = X_2 - 25$.

- The fourth column is a periodic sequence running (5,4,3,2,1,2,3,4,5,6) repeated to make 100 observations.
- Column 5 is 5 plus the difference between X_2 and X_4. To avoid direct collinearity, we change observations 54 and 96 to 5 and 2 respectively.
  $X_5 = X_2 - X_4 + 5$
- The last column is column 4 with 10 subtracted and observations 38 and 100 both changed to 3.
  $X_6 = X_4 - 10$

Thus:

$$\mathbf{X} = \begin{bmatrix} 1 & 20 & -4 & 5 & 20 & -5 \\ 1 & 21 & -4 & 4 & 22 & -6 \\ 1 & 22 & -3 & 3 & 24 & -7 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 28 & 3 & 5 & 28 & -5 \\ 1 & 29 & 4 & 6 & 28 & 3 \end{bmatrix}$$

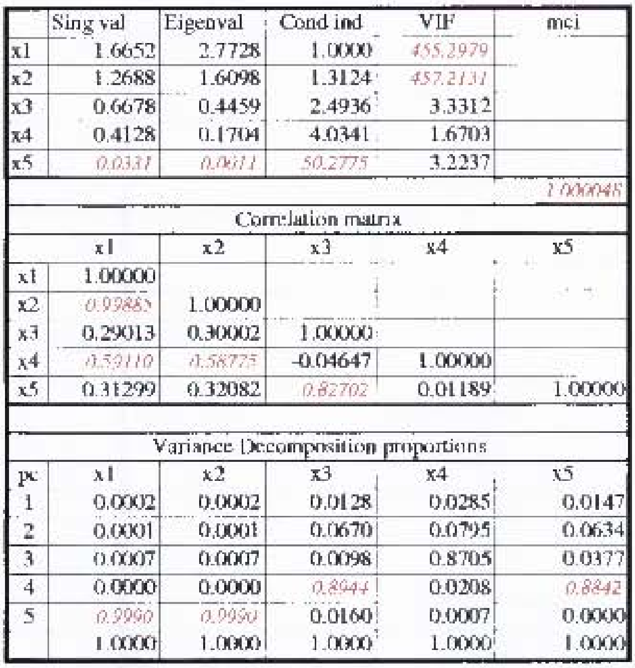
One of the primary objectives of this study is to compare different shrinkage estimators when there exists extremely high collinearity among the independent variables. Hence, it is critically important to ensure existence of extreme collinearity in X before proceeding any further.

5.3.1.1 Collinearity measures of X

By observation, we could conclude that the generated X matrix is collinear since some columns are generated from others. However, this would not be good enough to expose the magnitude of collinearity present in X. Also, we would not even know whether or not the existing level of collinearity satisfies the requirements for our objectives. Thus, we summarize the collinearity diagnostic results for the standardized X matrix table 5.1.

We indicate extreme collinearity in red and italics. From the diagnostics, we conclude that X is extremely collinear. Two of the five VIF values are larger than 10, both the last eigenvalue and the last singular value are very small compared to others, the condition number (last condition index) is around 50 and the mci is very close to 1. In addition, the correlation matrix, indicates four strong bivariate correlations between the following pairs of variables: (X_1 and X_2), (X_1 and X_4), (X_2 and X_4), and (X_3 and X_5).

Looking at the variance proportions, it is clear that two components with small eigenvalues contribute more than 50% to two regression coefficients; that is, the 5^{th} (last) pc contributes



	Sing val	Eigenval	Cond ind	VIF	mci
x1	1.6652	2.7728	1.0000	455.2979	
x2	1.2688	1.6098	1.3124	457.2131	
x3	0.6678	0.4459	2.4936	3.3312	
x4	0.4128	0.1704	4.0341	1.6703	
x5	0.0331	0.0011	50.2775	3.2237	
					1.000048
Correlation matrix					
	x1	x2	x3	x4	x5
x1	1.00000				
x2	0.99862	1.00000			
x3	0.29013	0.30002	1.00000		
x4	0.59110	0.58775	-0.04647	1.00000	
x5	0.31299	0.32082	0.82702	0.01189	1.00000
Variance Decomposition proportions					
pc	x1	x2	x3	x4	x5
1	0.0002	0.0002	0.0128	0.0285	0.0147
2	0.0001	0.0001	0.0670	0.0795	0.0634
3	0.0007	0.0007	0.0098	0.8705	0.0377
4	0.0000	0.0000	0.8944	0.0208	0.8842
5	0.9990	0.9990	0.0160	0.0007	0.0000
	1.0000	1.0000	1.0000	1.0000	1.0000

Table 5.1: Collinearity diagnostics for standardized X

more than 99% to X_1 and X_2 while the second last (4^{th}) contributes more than 88% to X_3 and X_5. We note with certainty that the independent variables are highly correlated. Hence we use the collinear **X** matrix to generate the dependent variable **Y**.

5.3.2 Generating the dependent variable, Y

The response variable is generated from the following model:

$$Y = X\beta_T + \epsilon \tag{5.1}$$

where

β_T is a (6×1) vector of true coefficients chosen such that

$$\beta_T' = \begin{bmatrix} 10 & 0.4 & 0.5 & 0.25 & 0.3 & 4.5 \end{bmatrix}$$

and **X** is defined by

$$\mathbf{X} = \begin{bmatrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \end{bmatrix}$$

with X_i being the i^{th} 100×1 column of the matrix defined earlier in this chapter.

ϵ is a 100×1 vector of random error terms with mean zero, fixed across different variance levels ($\sigma^2 = 2, 25$ and 100) and following either Normal, Laplace, Exponential or Student's t distributions. Our experiment consists of

$$\underset{estimators}{24} \quad \times \quad \underset{\substack{variance\\ levels}}{3} \quad \times \quad \underset{distributions}{4}$$

For each experiment, 500 monte carlo simulations (repetitions) are made.

5.3.2.1 Generating the error terms

We are keenly interested in finding a robust biased estimator that is not tied into normality hence our selection of error distributions includes long tailed distributions (non-normal). We want to observe whether or not heavy tailed distributions of error terms influence performance of estimators. The biased estimators are expected to perform outstandingly better than the OLSE when the error distribution is long tailed (Student's t) since the latter is too sensitive to extreme values.

All the programming is done in R; a command based statistical software package developed for statistical analysis, freely accessible at http://cran.r-project.org. The pseudo-random variables are generated from R built-in functions. The built-in functions generate the random variables in the following manner: the R pseudo-random generator produces a 32-bit integer whose top 31 bits are divided by 2^{31} to produce a real number in the range (0,1) (details are provided by Ripley (1987) and Venables and Ripley (1994)). Once the integers are generated, R uses them to produce values from the different distributions.

Full program details are given in appendix A.

Normal error terms

The probability density function of the Normal distribution is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \qquad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0 \tag{5.2}$$

Where μ and σ represent the mean and the standard deviation, used to specify location of the data and the spread of the distribution respectively. In this study, $\mu = 0$ and σ is varied across the three values as shown below

We generate each column of 100 random normal error terms from the following function

$$rnorm(100, 0, \sigma_i), \qquad i = 1, 2, 3$$

where

$\sigma_i = 0.1414, 5, 10$; corresponding to the variance levels $\sigma_i^2 = 2, 25, 100$ respectively

$\mu = E(\epsilon) = 0$
n $= 100$

Student's t error terms

The probability density function of the Student's t distribution is defined by

$$f_X(x) = \frac{\{\Gamma[(v+1)/2]\}}{(\pi v)^{1/2}\Gamma(v/2)[1+(x^2/v)]^{(v+1)/2}} \qquad -\infty < x < \infty, \quad v > 0 \tag{5.3}$$

where v is the degrees of freedom.

We generate each Student's t error variate as a ratio of a standard normal variate to the root of a chi-squared variate divided by the corresponding degrees of freedom as shown below. The numerator and the denominator are independent.

$$\frac{rnorm(100, 0, 1)}{\sqrt{rchisq(100, v_i, ncp = 0)/v_i}} \tag{5.4}$$

where v_i represents the degrees of freedom (non-negative) and ncp is the non-centrality parameter (non-negative), set to zero.

The variance expression for a Student's t distributed variable is $\sigma^2 = \dfrac{v}{v-2}$ for all even numbers greater than two ($v > 2$) and zero otherwise. Hence, computation of v (the degrees of freedom) from the formula $\sigma^2 = \dfrac{v}{v-2}$ requires the condition that $\sigma^2 > 1$.

The Student's t degrees of freedom used for simulation in this study are computed to be the following:

i	**1**	**2**	**3**
variance (σ_i^2)	2	25	100
$v_i = \dfrac{2\sigma_i^2}{\sigma_i^2 - 1}$	4	2.083	2.020

Table 5.2: The choice of v

Hence, we extract the degrees of freedom (v_i) and generate each of columns of the Student's t error terms using equation 5.4.

Laplace error terms

The Laplace density function is given by

$$fx(x) = \frac{1}{2c} \quad e^{-\frac{|x-a|}{c}}, \qquad -\infty < x < \infty, \quad -\infty < a < \infty, \quad c > 0$$

Hence the general distribution function may be defined by

$$F(x) = \begin{cases} \frac{1}{2} \quad e^{\frac{(x-a)}{c}} & x < a \\ 1 - \frac{1}{2} \quad e^{-\frac{(x-a)}{c}} & x > a \end{cases}$$

with a = mean and c defined such that the variance of X= $2c^2$.

We find the inverse function of the distribution function and express the Laplace variable in terms of a uniformly distributed variable in the following manner:

Let $u = F(x)$ and $x = F(u)^{-1}$:

For $x < a$,

$$\begin{aligned} u &= \frac{1}{2}e^{\frac{(x-a)}{c}} \\ 2u &= e^{\frac{(x-a)}{c}} \\ ln(2u) &= \frac{(x-a)}{c} \\ c\{ln(2u)\} &= x - a \\ x &= \text{a+c } \{\text{ln(2u)}\} \end{aligned}$$

For $x > a$,

$$\begin{aligned} u &= 1 - \tfrac{1}{2}e^{-\frac{(x-a)}{c}} \\ 2\{1-u\} &= e^{-\frac{(x-a)}{c}} \\ ln\{2(1-u)\} &= -\frac{(x-a)}{c} \\ c\{ln[2(1-u)]\} &= a - x \\ x &= \text{a - c}\{\text{ln[2(1-u)]}\} \end{aligned}$$

where u is a random number between 0 and 1 ($u \sim U(0, 1)$)

However, since the concern in this study is mainly on zero mean distributions, we set a=0 and compute the Laplace error values from the following expressions:

$$x = c \quad ln(2u) \qquad for \quad 0 \leq u \leq 0.5 \quad and \quad x < 0 \tag{5.5}$$

and

$$x = -c \quad ln(2[1-u]) \qquad for \quad 0.5 \leq u \leq 1 \quad and \quad x > 0 \tag{5.6}$$

We compute c to correspond to the desired variance levels as shown in table 5.3.

i	**1**	**2**	**3**
variance (σ_i^2)	2	25	100
$c_i = \sqrt{\frac{\sigma_i^2}{2}}$	1	3.536	7.07

Table 5.3: c estimates for Laplace distribution

Hence, we compute each Laplace error term from

$errlap[i] = c_i * log(2 * runif(1,0,1))$ for $x < 0$
and
$errlap[i] = -c_i * log(2 * (1 - runif(1,0,1)))$ for $x > 0$

where $runif(1,0,1)$ = a random number between 0 and 1.

Exponential error terms

An exponentially distributed random variable X has the density function

$$f(x) = \lambda e^{-\lambda x} \qquad x > 0, \quad \lambda > 0.$$

with mean $\frac{1}{\lambda}$ and variance $\frac{1}{\lambda^2}$.

We generate the columns of exponential error terms from the function

$$rexp(100, \lambda_i), \qquad i = 1, 2, 3$$

where $\lambda_i's$ correspond to the three desired variance levels.

Since $\sigma_i^2 = \frac{1}{\lambda_i^2}$, we equate σ_i^2 to each of the desired levels of variance and solve for the unknown parameter λ_i as shown in table 5.4.

i	**1**	**2**	**3**
variance (σ_i^2)	2	25	100
$\lambda_i = \sqrt{\frac{1}{\sigma_i^2}}$	0.707	0.2	0.1

Table 5.4: λ estimation for Exponential distribution

One way to generate X would be to find the inverse function of the Exponential distribution function

$$F(x) = 1 - e^{-\lambda x} \qquad x > 0 \tag{5.7}$$

and solve for x. The resulting expression would be

$$x = \frac{-ln(1-u)}{\lambda}, \tag{5.8}$$

where u is a random number between zero and one; ($u \sim U(0, 1)$)

5.4 The basis for comparison of estimators

The differences between the estimated coefficients and the true coefficients form the basis for assessment and comparison of the listed biased estimation methods. We note from the previous chapters that when two or more independent variables are collinear, the OLSE exhibits large variance and mean squared error, hence, the bias becomes a requirement for reduction of the MSE and the variance. In this study, we are mainly concerned with comparison and identification of the biased estimators of which the estimates deviate the least from the known true parameters, relative to the rest of the estimators considered for simulation.
The comparison of estimators is based on the following

- minimum squared Euclidean distance between the estimates and the true values (TMSE) and

- maximum efficiency of each estimator relative to the OLSE.

For each estimator, we define the relative efficiency to be a ratio of the total mean squared error of the OLSE to the total mean squared error of the estimator; denoted by

$$RE[\tilde{\beta}] = \frac{TMSE(\hat{\beta})}{TMSE(\tilde{\beta})}$$

where $\tilde{\beta}$ is the estimator of which the efficiency is being computed, relative to $\hat{\beta}$.

The relative efficiencies allow direct comparison of the biased estimators to the OLSE hence we use least squares as a yardstick. We characterize an efficient estimator by a large value of RE. From the definition of $RE[\tilde{\beta}]$ it should be easy to observe that if $TMSE(\tilde{\beta})$ is smaller than $TMSE(\hat{\beta})$, then $RE[\tilde{\beta}]$ is expected to be large and vice versa. For simplicity and convenience, we interpret the magnitude of relative efficiencies in the following manner:

> An estimator $\tilde{\beta}$ is more efficient than the OLSE if RE$[\tilde{\beta}]$ is greater than 1; greater than 1 in this context means anything beyond 1.01. Note the following:

* $1 \leq RE[\tilde{\beta}] \leq 1.01$ is considered equivalent to 1 and the corresponding estimator $\tilde{\beta}$ is said to be similar to the OLSE.

* Estimators whose RE values are less than 1 ($RE < 1$) are considered less efficient.

Based on the 500 repetitions, a biased estimator of which the relative efficiency is the highest is said to be the most efficient, compared to other shrinkage estimators.
The comparison is performed through the following subsequent steps.

- For each method of estimation, we estimate the regression coefficients from the simulated data and obtain 500 sets of betas. Each set of coefficients is a $(p+1) \times 1$ column vector.

- Subsequent to obtaining the coefficients, we compute the squared sum of the difference between the obtained coefficients and the prior known coefficients (β_T). For each method of estimation, a summary of the 500 replications is given as

$$\sum_{j=1}^{6} \sum_{i=1}^{500} (\tilde{\beta}_{ji} - \beta_{T_j})^2$$

where

$\tilde{\beta}_{ji}$ = the j^{th} estimate in the i^{th} replication, corresponding to any of the estimators under consideration

β_{T_j} = the j^{th} elements of β_T.

- Hence the relative efficiencies are computed from

$$\sum_{j=1}^{6} \sum_{i=1}^{500} (\hat{\beta}_{ji} - \beta_{T_j})^2 / \sum_{j=1}^{6} \sum_{i=1}^{500} (\tilde{\beta}_{ji} - \beta_{T_j})^2$$

where $\hat{\beta}_{ji}$ = the j^{th} ordinary least squares estimate in the i^{th} replication.

The method of estimation that results in the minimum TMSE and the maximum relative efficiency (RE) relative to other methods is the most preferred and is considered 'best' in the context of this thesis.

The 24 estimators used in this study are summarized in appendix C.

5.5 The simulation program

A summary of the R simulation program is provided in this section. We do not specify the syntax in this summary but rather concentrate on the flow of the program or the sequence of simulation steps. Full program details are provided in appendix A.

For each of the four selected distributions (Normal, Student's t, Laplace and Exponential) and each of the three variance levels ($\sigma^2 = 2, 25$ and 100), we run 500 simulations. We compute the X matrix and error terms in R and store them as Excel files. For each of the 500 simulations, the error terms and the X matrix are used to compute the dependent variable.

The dependent variable and the predictor variables are standardized; X is centred and scaled to be in correlation form. With the standardized variables, we do least squares estimation and principal components regression deleting no roots, one root and two roots.

We extract the coefficients and the variance estimates from PCdel1 and the OLS and compute the required unknown parameters for ridge, generalized ridge, Stein, Liu and generalized Liu estimation methods. The respective estimators are computed and transformed back to the unstandardized form. Hence the mean squared error, total mean squared error and the relative efficiencies for each of the estimators are computed and written to Excel files.

We define the relationship between the standardized and the unstandardized regression coefficients by the following expression

$$\beta^* = \frac{\beta \times (s_x)}{(s_y)}, \qquad \textit{implying that} \qquad \beta = \frac{\beta^* \times (s_y)}{(s_x)}$$

where

β^* = a vector of standardized coefficients

β = vector of unstandardized coefficients

s_x = the root of the sum of squared columns of the centered X.

s_y = the root of the sum of squared centered observations in Y.

5.6 Summary

In this chapter, the simulation study was presented and some of the previous studies were reviewed, compared and differentiated from this study. Further, the basis for comparison of estimators and the simulation program summary were provided.

Chapter 6

Discussion of Simulation results

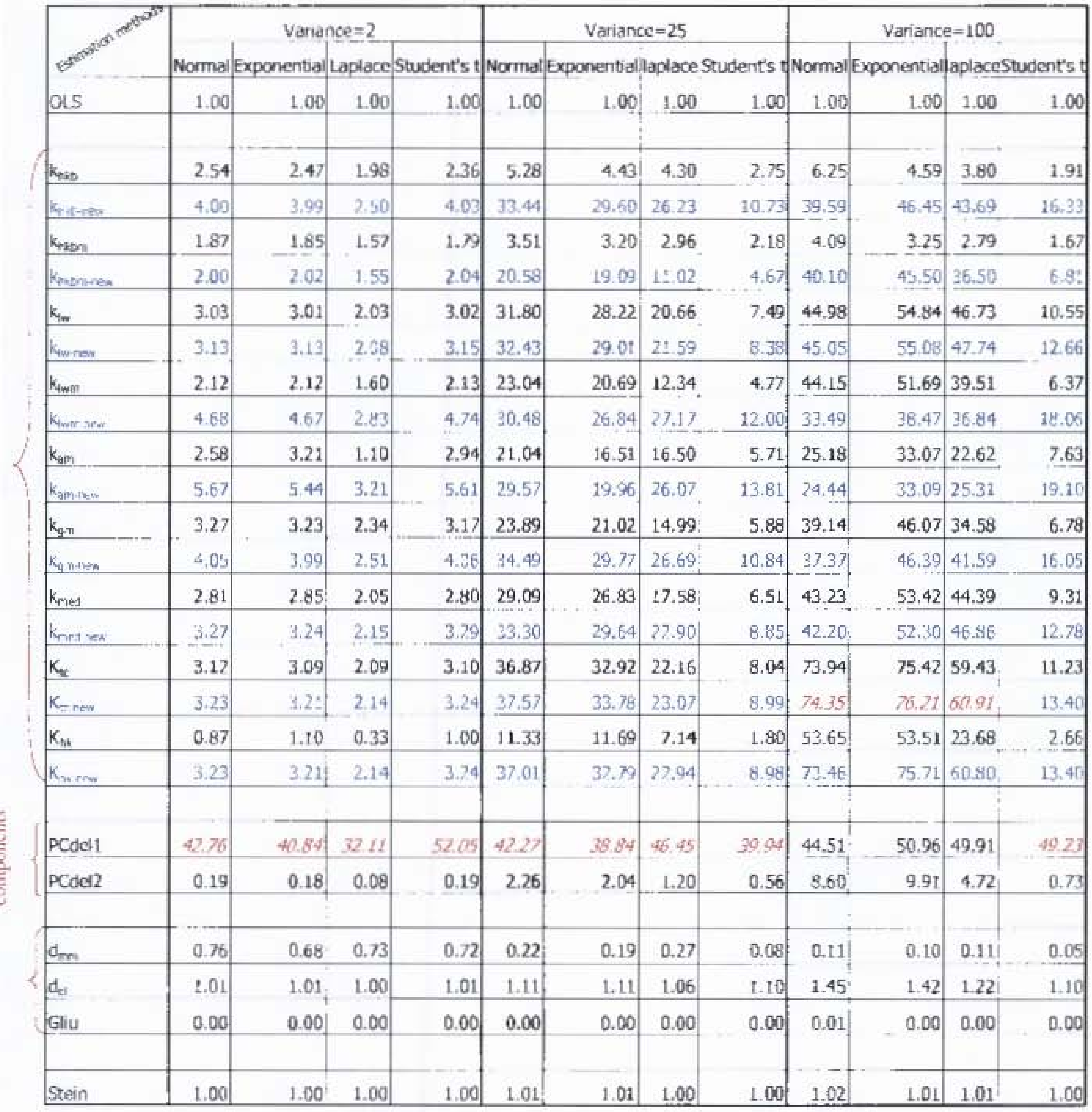
Results of the simulation study are presented and discussed in this chapter. The relative efficiencies of 24 shrinkage estimators at three levels of variances of error terms and four different distributions are discussed. We split up the discussion into the following sub-sections:

- *Overall performance of estimators:* we discuss the general performance of shrinkage estimators relative to the OLSE. Further, the issue of whether or not the new estimators improve on the old ones is investigated.
- *Relative performance of estimators by variance levels × distributions:* the efficiencies of individual biased estimators across the variance levels and distributions are discussed.
- *General performance across different families of estimates:* the discussion focuses on performance and comparison of families of estimators.

6.1 The Results

We present the relative efficiencies (REs) of the biased estimators in table 6.1. Except for principal components regression methods, generalized Liu and Stein estimation, each estimator is represented by the corresponding parameter. The rows correspond to the estimation methods and the columns correspond to the distributions and the variance levels.

We do not tabulate the REs in any particular order since for each variance, the values fluctuate across the four distributions hence sorting becomes difficult. For each distribution at a particular level of variance, the largest RE value is written in italics and coloured in red. Further, new estimation methods proposed in this study are highlighted in blue to differentiate them from the rest.


components

Estimation methods	Variance=2				Variance=25				Variance=100			
	Normal	Exponential	Laplace	Student's t	Normal	Exponential	laplace	Student's t	Normal	Exponential	laplace	Student's t
OLS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
k_{HKB}	2.54	2.47	1.98	2.36	5.28	4.43	4.30	2.75	6.25	4.59	3.80	1.91
$k_{HKB-new}$	4.00	3.99	2.50	4.03	33.44	29.60	26.23	10.73	39.59	46.45	43.69	16.33
k_{HKBm}	1.87	1.85	1.57	1.79	3.51	3.20	2.96	2.18	4.09	3.25	2.79	1.67
$k_{HKBm-new}$	2.00	2.02	1.55	2.04	20.58	19.09	11.02	4.67	40.10	45.50	36.50	6.81
k_{lw}	3.03	3.01	2.03	3.02	31.80	28.22	20.66	7.49	44.98	54.84	46.73	10.55
k_{lw-new}	3.13	3.13	2.08	3.15	32.43	29.01	21.59	8.38	45.05	55.08	47.74	12.66
k_{lwm}	2.12	2.12	1.60	2.13	23.04	20.69	12.34	4.77	44.15	51.69	39.51	6.37
$k_{lwm-new}$	4.68	4.67	2.83	4.74	30.48	26.84	27.17	12.00	33.49	38.47	36.84	18.06
k_{am}	2.58	3.21	1.10	2.94	21.04	16.51	16.50	5.71	25.18	33.07	22.62	7.63
k_{am-new}	5.67	5.44	3.21	5.61	29.57	19.96	26.07	13.81	24.44	33.09	25.31	19.10
k_{gm}	3.27	3.23	2.34	3.17	23.89	21.02	14.99	5.88	39.14	46.07	34.58	6.78
k_{gm-new}	4.05	3.99	2.51	4.06	34.49	29.77	26.69	10.84	37.37	46.39	41.59	16.05
k_{med}	2.81	2.85	2.05	2.80	29.09	26.83	17.58	6.51	43.23	53.42	44.39	9.31
$k_{med-new}$	3.27	3.24	2.15	3.79	33.30	29.64	27.90	8.85	42.20	52.30	46.86	12.78
K_{sc}	3.17	3.09	2.09	3.10	36.87	32.92	22.16	8.04	73.94	75.42	59.43	11.23
K_{sc-new}	3.23	3.21	2.14	3.24	37.57	33.78	23.07	8.99	*74.35*	*76.21*	*60.91*	13.40
K_{hk}	0.87	1.10	0.33	1.00	11.33	11.69	7.14	1.80	53.65	53.51	23.68	2.66
K_{hk-new}	3.23	3.21	2.14	3.74	37.01	32.79	27.94	8.98	73.46	75.71	60.80	13.40
PCdel1	*42.76*	*40.84*	*32.11*	*52.05*	*42.27*	*38.84*	*46.45*	*39.94*	44.51	50.96	49.91	*49.23*
PCdel2	0.19	0.18	0.08	0.19	2.26	2.04	1.20	0.56	8.60	9.91	4.72	0.73
d_{mm}	0.76	0.68	0.73	0.72	0.22	0.19	0.27	0.08	0.11	0.10	0.11	0.05
d_{cl}	1.01	1.01	1.00	1.01	1.11	1.11	1.06	1.10	1.45	1.42	1.22	1.10
Gliu	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Stein	1.00	1.00	1.00	1.00	1.01	1.01	1.00	1.00	1.02	1.01	1.01	1.00

Table 6.1: Relative efficiencies of 24 biased estimator and the OLSE. The most superior estimators are is indicated by red, families of estimators indicated by braces and labels and new estimators indicated by blue.

6.2 Discussion of results

6.2.1 Overall performance of estimators

- 19 out of 24 biased estimators are more efficient than the OLSE at all three variance levels and four distributions. The efficient estimators include all ridge and generalized ridge estimators and the PCdel1 estimator. The Stein estimator and two of the three Liu estimators do not show any efficiency at all, regardless of the distribution nor the variance level. Further, PCdel2 and one of the Liu estimators show efficiency only when the level of variance is high and the distributions are not long tailed.
- The new estimators show an improvement over the corresponding old estimators. The following cases are the only exceptions for which no improvement is observed.

Estimator	Distribution	Variance level
$k_{hkb-new}$	-Laplace	$\sigma^2 = 2$
$k_{lwm-new}$	-Normal -Exponential -Laplace	$\sigma^2 = 100$
k_{am}	-Normal	$\sigma^2 = 100$

Table 6.2: Exceptional cases for improvement

6.2.2 Relative performance by variance levels × distributions

We observe an overall positive relationship between the variance of error terms and performance of shrinkage estimators relative to the OLSE. The relative efficiencies of estimators increase drastically when the variance increases, implying that the biased estimators are more advantageous over the OLSE when the variance is large.

6.2.2.1 Relative performance when $\sigma^2 = 2$

We outline the 'best' eight performing estimators in table 6.3. There is a similarity between the distributions: PCdel1, k_{am-new}, $k_{lwm-new}$ and k_{gm-new} are consistently the four most superior in all the distributions. This implies that the distributions do not influence performance of the estimators when the variance is small.

	1	2	3	4	5	6	7	8
N	PCdel1	k_{am-new}	$k_{lwm-new}$	k_{gm-new}	$k_{hkb-new}$ k_{gm}	$k_{med-new}$ K_{tc-new}	K_{hk-new}	k_{lw-new}
Exp	PCdel1	k_{am-new}	$k_{lwm-new}$	k_{gm-new} $k_{hkb-new}$	$k_{med-new}$	k_{gm} K_{tc-new}	$K_{hk--new}$ k_{lw-new} k_{am}	k_{lw-new}
Lap	PCdel1	k_{am-new}	$k_{lwm-new}$	k_{gm-new}	$k_{hkb-new}$	k_{gm}	$k_{med-new}$ K_{hk-new}	K_{tc-new}
t	PCdel1	k_{am-new}	$k_{lwm-new}$	k_{gm-new}	$k_{hkb-new}$	$k_{med-new}$ K_{hk-new}	K_{tc-new}	k_{lw-new}

Table 6.3: Eight 'best' performing estimators, ranked from 1^{st} to 8^{th}: $\sigma^2 = 2$

6.2.2.2 Relative performance when $\sigma^2 = 25$

There is a significant increase in the REs compared to those of $\sigma^2 = 2$, implying that the variance might influence performance of estimators. We summarize eight 'best' performing estimators over the distributions when $\sigma^2 = 25$ in table 6.4.

	1	2	3	4	5	6	7	8
N	PCdel1	K_{tc-new}	K_{tc}	K_{hk-new}	k_{gm-new}	$k_{med-new}$	$k_{hkb-new}$	k_{lw-new}
Exp	PCdel1	K_{tc-new}	K_{tc}	K_{hk-new}	k_{gm-new}	$k_{med-new}$	$k_{hkb-new}$	k_{lw-new}
Lap	PCdel1	$k_{lwm-new}$	k_{gm-new}	$k_{hkb-new}$	k_{am-new}	K_{tc-new}	K_{hk-new}	$k_{med-new}$
t	PCdel1	k_{am-new}	$k_{lwm-new}$	k_{gm-new}	$k_{hkb-new}$	K_{tc-new}	K_{hk-new}	$k_{med-new}$

Table 6.4: Eight 'best' performing estimators, ranked from 1^{st} to 8^{th}: $\sigma^2 = 25$

- The PCdel1 estimator is consistently outperforming the rest of the estimators for all distributions; a similar result was observed when the variance was small. This might be the implication that PCdel1 is not influenced by neither the variance level nor the distribution.

- Normal and Exponential distributions are similar, perhaps because the two are of the same family.

- The new estimators are superior to all other estimators except the PCdel1 estimator; implying that the new proposed method is effective.

6.2.2.3 Relative performance when $\sigma^2 = 100$

Generally, the RE values are much higher than those observed when $\sigma^2 = 2$ and $\sigma^2 = 25$. In Table 6.5, the 'best' eight performing estimators are summarized over the distributions, when

$\sigma^2 = 100$.

	1	2	3	4	5	6	7	8
N	K_{tc-new}	K_{tc}	K_{hk-new}	k_{lw-new}	K_{hk}	k_{med}	PCdel1	k_{gm-new}
Exp	K_{tc-new}	K_{tc}	K_{hk-new}	k_{lw-new}	K_{hk}	k_{med}	PCdel1	k_{gm-new}
Lap	PCdel1	$k_{lwm-new}$	k_{gm-new}	$k_{hkb-new}$	k_{am-new}	K_{tc-new}	K_{hk-new}	$k_{med-new}$
t	PCdel1	k_{am-new}	$k_{lwm-new}$	$k_{hkb-new}$	k_{gm-new}	K_{tc-new}	K_{hk-new}	$k_{med-new}$

Table 6.5: Eight 'best' performing estimators, ranked from 1^{st} to 8^{th}: $\sigma^2 = 100$

- We observe an incredible performance by the estimators corresponding to K_{tc-new}, K_{tc} and K_{hk-new} when the error terms follow Normal and Exponential distributions.

- There is no significant change in the relative efficiencies for the Student's t distribution; the order of performance of estimators is the same for all variance levels. Hence it may be said that the variance does not play an important role in the Student's t distribution.

- There is more fluctuation in performance estimators. PCdel1 and k_{gm} are ranked 7 and 8 in Normal and Exponential distribution however, for Laplace and Student's t, PCdel1 is dominant and k_{gm} is among the first five 'best' performing estimators.

6.2.3 General performance across different families of estimators

- All ridge estimators except K_{hk} (at $\sigma^2 = 2$) perform better than the OLSE at all variance levels and distributions.

- We cannot generalize on the principal components family, however, we note that the PCdel1 estimator is outstandingly a good, stable estimator. PCdel2 does not perform well hence we do not consider it advantageous in this study.

- In this study, Stein is not superior to OLS at all orientations; four distributions and three variance levels.

- The Liu family is also not doing too well; none of Liu estimators significantly outperforms the OLSE.

6.3 Summary

In this chapter, we presented the simulation results. It has been found that not all the shrinkage estimators considered in this study are more efficient than the OLSE. The PCdel1 estimator is superior to all other estimators. Also, there is a positive relationship between performance of estimators and the variance of error terms.

Estimators associated with the new proposed method are generally performing better than (superior to) the OLSE and other biased estimators. The relative efficiencies increase with the variance. It has been observed that there is more variation of performance of estimators across the four distributions as the variance level increases, implying that the small variance does not influence the esimators.

Chapter 7

Summary and Recommendations

In this chapter, we evaluate the objectives of this study, draw conclusions and make recommendations.

7.1 Evaluation of objectives

We re-examine the objectives of this study with a view to determine whether or not they have been achieved.

The primary objectives of this study were the following

- *To propose a new method for estimating the shrinkage factors.*
  The motivation for this objective was provoked by the fact that the traditional methods are vulnerable to collinearity since the methods depend on OLS, a procedure that has been shown to be highly bugged by existence of collinearity. Our prior expectation was that the new method would show a significant improvement over the traditional methods since the new method is independent of OLS. The results show that the new proposed method is indeed an improvement of the traditional methods hence the objective has been achieved.

- *To classify 24 biased estimators under one category of shrinkage estimation with a view to determine the most effective and robust estimator.*
  The desire to achieve this objective was stimulated by the fact that there is currently no outstandingly best performing biased estimator. We hoped to identify one or more predominant biased estimator(s) from the new estimators. From the results, we observe that principal components and the ridge family are the 'best' but we still cannot generalize on the outstandingly superior estimator.

7.2 Conclusions

We draw the following general conclusions:

- *Biased estimators outperform the OLSE when collinearity is present in the data.*
  The implication of this conclusion is that biased estimators are more reliable and closer to the true values than the OLSE when X is not orthogonal. The same conclusion was reached by Hoerl et al. (1975); Marquardt and Snee (1975); Guilkey and Murphy (1975); Lawless and Wang (1976); Hoerl and Kennard (1976); Hocking (1976); Gunst and Mason (1977); Winchen and Churchill (1978); Thiart et al. (1993); Thiart (1994); Breiman (1995); Aldrin (1997); Fu (1998); Kaciranlar and Sakallioglu (2001); Wencheko (2001); and Liu (2003).

- *The new proposed procedure leads to a significant improvement in performance of the shrinkage estimators.*
  From the results, the estimators associated with the new proposed method outperform the known estimators, of which the biasing factors are based on OLS solution. This implies that the new method is effective and should be implemented to improve on shrinkage estimation.

- *The distribution of error terms plays a minimal or no role in performance of biased estimators, especially when the variance is small.*
  From the results, we observe that for the least variance, the RE values are nearly similar for all four distributions. Hence, the distribution of error terms does not influence performance of biased estimators. This conforms with the conclusion drawn by Thiart (1994).

- *The relative efficiencies increase with the variance.*
  We note from the findings that the relative efficiencies of biased estimators increase considerably as the levels of variance increase. This means that the biased estimators are affected by the level of variance of error terms; large variances are likely to indicate good performance of shrinkage estimators. A similar conclusion was reached by Thiart (1990).

- *PCdel1 is optimal for the data set used in this study.*
  The simulation results indicate that the PCdel1 estimator is outstandingly superior to all the estimators considered in this study. However, some of the studies do not report any particular better biased estimator relative to others. This implies that the efficiency

of a biased estimator depends on the level of collinearity and the shrinkage factor/matrix.

Although principal component regression is optimal in this study, we cannot generalize the results. Furthermore, principal component analysis transforms the data into new artificial data-specific variables making the results difficult to interpret in general. In contrast, ridge regression deals with variables in their original form, hence it is much easier for the experimenter working with the data as all the variables are included in their original form in the model.

- *The ridge family of estimators is consistently better than other families.*
  Apart from PCdel1, we observe efficient performance by all ridge and generalized ridge estimators. This implies that ridge regression is imperative and should be considered in many applications.

- *Liu estimators are not ideal for the data set used in this study.*
  The Liu family of estimators performs disastrously at all variance levels. This could imply that

  - Liu estimators are incapable of handling collinearity problems or
  - the effectiveness of biased estimation techniques varies across the kinds of data analyzed; meaning that not all the methods are appropriate for usage all the time.

Owing to the conclusions stated above, we provide the recommendations below.

7.3 Recommendations

* Use biased estimation methods in collinear designs of matrices. Biased estimators outperform the least squares estimator in the presence of collinearity.

* Where possible, refrain from using the OLS solution to estimate the biasing factors. Rather use the principal components solution instead of OLS. This has a critical importance in improvement of some of the biased estimation methods, more especially when principal components estimates exhibit much stability.

* In analysis of collinear data, explore different biased estimation methods to identify one or more that best suit the circumstances and the problem at hand. Include ridge regression in the search for the optimal biased method(s).

* Investigate further into potential improvement of biased/shrinkage estimation and appropriate methods for estimating the biasing factors.

7.4 Further research

- We recommend further investigations into our new proposed method, using different data sets and other orientations.

- It could be interesting to see performance of new estimators under different collinearity levels. Perhaps one weak, medium and extreme orientations of collinearity could portray a better picture.

- The relationship between the variance and relative efficiency of biased estimators could be pursued further.

- Further improvement of the new proposed method could be investigated.

References

Akdeniz, F. (2001). *The examination and analysis of residuals for some biased estimators in linear regression.* Communications in Statistics - Theory and Methods, 30, 1171-1183.

Akdeniz, F. and Kaciranlar, S. (1995). *On the almost unbiased generalized Liu estimator and unbiased estimation of the bias and mse.* Communications in Statistics - Theory and Methods, 24, 789-1797.

Aldrin, M. (1997). *Length modified ridge regression* Computational Statistics and Data Analysis, 25, 377-398.

Allison, P.D. (1999). *Multiple Regression.* Pine Forge Press, California.

Arslan, O. and Billor, N. (2000). *Robust Liu estimator for regression based on an M-estimator.* Journal of Applied Statistics, 27, 39-47.

Askin, G.R. and Montgomery, D.C. (1980). *Augmented Robust Estimators.* Technometrics, 22, 333-341.

Belsley, D.A. (1987). *Collinearity and Least Squares Regression: Comment: Well-Conditioned Collinearity Indices* Statistical Science, 2, 86-91.

Belsley, D. A. and Oldford, R. W. (1986). *The general problem of ill-conditioning and its role in statistical analysis.* Computational Statistics and data analysis, 4, 103-120.

Belsley, A.D., Kuh, E. and Welsch, E.R. (1980). *Regression Diagnostics: Identifying influential data and sources of collinearity.* John Wiley and Sons, New York.

Bettman, J.R. (1973). *Perceived price and product perceptual variables.* Journal of Marketing Research, 10, 100-102.

Bolding, J.T. and Houston, S.R. (1974). *Fortran computer program for computation of ridge regression coefficients.* Educational and Psychological Measurements, 34, 151-152.

Bradley, C.E. and McGann, A.F. (1977). *RIDGEREG: A program to improve the precision of regression estimates for nonorthogonal data.* Journal of Marketing Research, 14, 412-431.

Breiman. L. (1995). *Better subset regression using the nonnegative Garotte.* Technometrics, 37. 373-384.

Brown. P.J. (1993). *Measurement, Regression and Calibration.* Clarendon Press, Oxford, 55-71.

Bush, A.J. (1980). *Ridge: A program to perform ridge regression analysis.* Behaviour Research Methods and Instrumentation, 12, 73-74.

Carmer, S.G. and Hsieh, W.T. (1979). *Exploring biased regression with SAS.* Proceedings of the Fourth Annual SAS Users Group International Conference, SAS Institute Inc., Cary, North Carolina, 27511, 223-228.

Chatterjee, S. and Price, B. (1977). *Regression Analysis by Example.* John Wiley and Sons: New York.

Conniffe, D and Stone, J. (1973). *A Critical View of Ridge Regression.* The Statistician, 22, 181-187.

Dempster, A.P. (1973). *Alternatives to Least Squares in Multiple Regression.* In: Kabe, D.G. and Gupta, R.P.(eds) Multivariate Statistical Inference. Amsterdam, North Holland, 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, 72, 77-106.

Dorsett, D., Gunst, R.F. and Gartland, C.E. (1983). *Multicollinear Effects of Weighted Least Squares Regression.* Elsevier Science Publishers (North Holland).

Dwivedi, T.D., Srivastava, V.K. and Hall, R.L. (1980). *Finite Sample Properties of Ridge Estimators.* Technometrics, 22, 205-212.

Efron, B. and Morris, C. (1973). *Stein's Estimation Rule and its Competitors.* Journal of the American Statistical Association, 65, 117-130.

Elston, D.A. and Proe, M.F. (1995). *Smoothing Regression Coefficients in an Overspecified Regression Model with Interrelated Explanatory Variables.* Applied Statistics, 44, 395-406.

Farrar. D. E. and Glauber, R. R. (1967). *Multicollinearity in regression analysis: the problem revisited.* Review of Economics and Statistics, 49, 92-107.

Feig, D.G. (1978). *Ridge Regression: when biased estimation is better.* Social Science Quarterly, 58, 708-716.

Firinguetti, L. and Rubio, H. (2000). *A note on the moments of stochastic shrinkage parameters in ridge regression.* Communications in Statistics - Simulation and Computation, 29, 995-970.

Frank, I.E. and Friedman, J.H. (1993). *A statistical view of some chemometrics regression tools.* Technometrics, 35, 109-135.

Frisch, R. (1934). *Statistical Confluence Analysis by Means of Complete Regression Systems.* Oslo; Universitetets Okonomiske Institutt, Oslo, Norway.

Fu, W. (1998). *Penalized Regression; The Bridge Versus the Lasso.* Journal of Computational and Graphical Statistics, 7, 397-416.

Gibbons, D.I. and McDonald, C.G. (1984). *A rational Interpretation of the Ridge Trace.* Technometrics, 26, 339-36.

Grob, J. (2003). *Restricted Ridge Estimation.* Statistics and Probability Letters 65, 57-64.

Grohn, Y. T., Schukken, Y. H., McDermott, B. and McDermott, J. J. (2003). *Analysis of correlated discrete observations: background, examples and solutions.* Preventive Veterinary Medicine, 59, 223-240.

Gruber, J. (1979). *Empirical Bayes, James-Stein and Ridge Regression Type Estimators for Linear Models.* Unpublished PhD Thesis, University of Rochester.

Gruber, J. (1980). *Multicollinearity and Biased Estimation.* Proceedings of a Conference at the University of Hagen. Vandenhoeck & Ruprecht in Gottingen.

Gruber, J. (1998). *Improving Efficiency by shrinkage; The James-Stein and Ridge Regression Estimators.* Marcel Dekker, New York.

Guilkey, D.K. and Murphy, J.L. (1975). *Directed ridge regression techniques in cases of multicollinearity.* Journal of American Statistical Association, 70, 769-775.

Gunst, R.F. (1979). *An approach to the programming of biased regression algorithms.* Communications in Statistics - Simulation and Computation, 8, 151-159.

Gunst, R.F. (1980). *A critique of some Ridge Regression Methods: Comment.* Journal of American Statistical Association, 75, 98-100.

Gunst, R. F. (1983). *Regression analysis with multicollinear predictor variables: Definition, Detection and Effects.* Journal of Communication in Statistics - Theory and Methods, 12, 2217-2260.

Gunst, R.F. and Mason, L.R. (1977). *Biased Estimation in regression: an Evaluation Using Mean Squared Error.* Journal of the American Statistical Association, 72, 616-628

Gunst, R.F., Webster, J.T. and Mason, R.L. (1976). *A Comparison of Least Squares and Latent Root Regression Estimators.* Technometrics, 18, 75-83.

Halawa, A.M. and El Bassiouni, M.Y. (2000). *Tests of Regression Coefficients under Ridge Regression Models.* Journal of Statistical Computation, 65, 341-356.

Hastie, T., Tibshirani, R. and Freidman, J. (2001). *The Elements of Statistical Learning; Data Mining, Inference and Prediction.* Springer-Verlag, New York.

Hawkins, D.M. (1975). *Relations Between Ridge Regression and Eigenanalysis of the Augmented Correlation Matrix.* Technometrics, 17, 477-479.

Hawkins, D.M. and Yin, X. (2002). *A faster algorithm for ridge regression of reduced rank data.* Computational Statistics and Data Analysis, 40, 253-262.

Hill, C.R., Fomby, B.T. and Johnson, R.S. (1977). *Component selection norms for principal components regression.* Communications in Statistics - Theory and Methods, 4, 309-334.

Hocking, R.R. (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, 69-78.

Hoerl, A.E. (1962). *Application of ridge analysis to regression problems.* Chemical Engineering Progress, 58, 54-59.

Hoerl, W. (1985). *Ridge Analysis 25 years later.* The American Statistician, 39, 186-192.

Hoerl, A.E. and Kennard, R.W. (1968). *Ridge regression: Biased estimation for nonorthogonal problems.* E.I du Pont de Nemours and Co., Engineering Department Report, Accession No. 13183.

Hoerl, A.E. and Kennard, R.W. (1970a). *Ridge regression: Biased estimation of nonorthogonal problems.* Technometrics, 12, 55-67.

Hoerl, A.E. and Kennard, R.W. (1970b). *Ridge regression: Applications to nonorthogonal problems.* Technometrics, 12, 69-82.

Hoerl, A.E. and Kennard, R.W. (1976). *Ridge regression: iterative estimation of the biasing parameter.* Communications in Statistics, Series A, 5, 77-88.

Hoerl, E. and Kennard, R.W. (1980). *Ridge Regression.* American Journal of Mathematical and Management Sciences, 1, 5-83.

Hoerl, E. and Kennard, R.W. (2000). *Ridge Regression: Biased Estimation for nonorthogonal Problems.* Technometrics, 42, 80.

Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975). *Ridge regression: Some simulations.* Communications in Statistics, 4, 105-123.

Hong, D., Hwang, C. and Ahn, C. (2004). *Ridge estimation for regression models with crisp inputs and Gaussian fuzzy output.* Fuzzy Sets and Systems 142, 307-319.

Jagpal, S.H. (1982). *Multicollinearity in Structural Equation Models With Unobservable Variables.* Journal of Marketing Research, XIX, 431-9.

Jain, A.K., Mahajan, V. and Bergier, M. (1977). *RRIDGE: A program for estimating parameters in the presence of multicollinearity.* Journal of Marketing Research, 14, 561.

James, W. and Stein, C. (1961). *Estimation with quadratic loss.* Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability, 1, 361-380. University of Califonia Press.

Jeffery, H. and McKinney, G. (1975). *Application of Ridge regression on agricultural data.* Journal of the Australian Institute of Agricultural Science, 41, 34-39.

Jeffers, J.N.R. (1967). *Two case studies in the application of principal component analysis. I. Artificial data.* Applied Statistics, 16, 225-236.

Jollifee, I.T. (1982). *A note on the use of principal components in regression.* Applied Statistics, 31, 300-303.

Kaciranlar, S. and Sakallioglu, S. (2001). *Combining the Liu estimator and the Principal component regression estimator.* Communications in Statistics - Theory and Methods, 30, 2699-2705.

Kaciranlar, S. and Sakallioglu, S. Akdeniz, F., Styan, G.P, and Werner, J.H. (1999). *A new biased estimator in linear regression and a detailed analysis of the widely-analysed dataset on Portland Cement.* The Indian Journal of Statistics, Series B, 61, 443-449.

Kasarda, J.D. and Shih, W.P. (1977). *Optimal bias in ridge regression approaches to multicollinearity.* Sociological Methods and Research, 5, 461-470.

Kendall, M.G. (1957). *A Course in Multivariate Analysis.* London: Charles W. Griffin.

Kibria, B.M. (2003). *Performance of Some New Ridge Regression Estimators.* Communication in Statistics - Simulation and Computation, 32, 419-435.

Kidwell, J.S. and Brown, L. H. (1982). *Ridge Regression as a technique for analysing models with Multicollinearity.* Journal of Marriage and the Family, 2, 287-300.

Knight, K. and Fu, W. (2000). *Asymptotics for Lasso-Type Estimators.* The Annals of Statistics, 28, 1356-1378.

Lawless, J.F. and Wang, P. (1976). *A simulation study of ridge and other regression estimators.* Communications in Statistics - Theory and Methods, 4, 307-323.

Le Cessie, S. and Van Houwelingen, J. C. (1992). *Ridge Estimators in Logistic Regression.* Applied Statistics, 41, 191-201.

Liu, K. (1993). *A new class of Biased estimate in Linear regression.* Communications in Statistics - Theory and methods, 22, 393-402.

Liu, K. (2003). *Using Liu-Type Estimator to Combat Collinearity.* Communications in Statistics - Theory and Methods, 32, 1009-1020.

Mallows, C.L. (1973). *Some comments on C_p .* Technometrics, 15, 661-675.

Malthouse, E.C. (1999). *Ridge Regression and Direct Marketing Scoring Models.* Journal of Interactive Marketing, 13, 10-23.

Mandel, J. (1982). *Use of the Singular Value Decomposition in Regression Analysis.* The American Statistician, 36, 15-24

Marquardt, D.W. (1970). *Generalized Inverses, ridge regression, biased linear estimation and nonlinear estimation.* Technometrics, 12, 591-612.

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

Mayer, L.W. and Willkie, T.A. (1973). *On biased estimation in linear models.* Technometrics, 15, 497-508.

McDonald, G.C. (1980). *Some Algebraic Properties of Ridge Coefficients.* Journal of Royal Statistical Society B, 42, 31-34.

McDonald, G.C. and Galarneau, D.I. (1975). *A Monte Carlo evaluation of some ridge-type estimators.* Journal of the American Statistical Association, 70, 407-416.

McDonald, G.C. and Schwing, R.C. (1973). *Instabilities of regression estimates relating air pollution to mortality.* Technometrics, 15, 463-481.

MIT (1975). *TEP: Robust and Ridge Regression.* D0070N, 123 pp. MIT Information Processing Services, Cambridge, MA 02139.

Newhouse, J.P. and Oman, S.D. (1971). *An Evaluation of ridge estimators.* Rand report. No R-716-PR. Rand Corp., Santa Monica, Califonia.

Ngo, S.H., Kemeny, S. and Deak, A. (2003). *Performance of the ridge regression method as applied to complex linear and nonlinear models.* Chemometrics and Intelligent Laboratory Systems 67, 69-78.

Nomura, M. (1988). *On the Almost Unbiased Ridge regression estimator.* Communications in Statistics - Simulation and Computation, 17, 729-743.

Ohtani, K. (1986). *On Small Sample properties of the almost unbiased Generalized ridge estimator.* Communications in Statistics - Theory and methods, 15, 1571-1578.

Oman, S.D. (1991). *Random Calibration with many measurements: An application of Stein Estimation.* Technometrics, 33, 187-195.

Pitman, E. (1937). *The closest estimates of statistical parameters.* Proceedings of the cambridge Philosophical Society, 33, 212-222.

Rawlings, J.O., Pantula, S.G. and Dickey, D.A. (1998). *Applied Regression Analysis; A Research Tool.* Springer-Verlag, New York.

Ripley, B.D. (1987). *Stochastic Simulation.* New York: John Wiley and Sons.

Sauerbrei, W. (1999). *The Use of Resampling Methods to Simplify Regression Models in Medical Statistics.* Applied Statistics, 48, 313-329.

Sclove, S.L. (1968). *Improved estimators for coefficients in linear regression.* Journal of the American Statistical Association, 63, 596-606.

Sengupta, D. and Bhimasankaram, P. (1997). *On the roles of Observations in Collinearity in the linear Model.* Journal of the American Statistical Association, 92, 1024-1032.

Sinha, A.N. and Hardy, K.A. (1979). *A SAS macro for ridge analysis of multivariate general linear models.* Proceedings of the Fourth Annual SAS Users Group International Conference, SAS Institute Inc., Cary, North Carolina 27511, 229-233.

Stein, C. (1956). *Inadmissibility of the usual estimator for the mean of a Multivariate Normal Distribution.* Proceedings of the third Berkeley Symposium on Mathematics, Statistics and Probability. Berkeley: University of California Press, 197-206.

Stein, C. (1960). *Multiple regression.* In Contributions to Probability and Statistics, 424 -443.

Stewart, G.W. (1987). *Collinearity and Least Squares Regression.* Statistical Science, 2, 68-84.

Sujan, I. and Condik, S. (1979). *Methods for Efficient Estimation of Parameters in Regression Models Containing Significant Multicollinearity.* The Sixth Conference on Problems of Building and Estimation of large Econometric models, Polanica Zdroj (Organized by the Institute of Econometrics and Statistics, University of Lordz).

Sundberg, R. (1993). *Continuum Regression and Ridge Regression.* Journal of the Royal Statistical Society. Series B (Methodological), 55, 653-659.

Swindel, B.F. (1981). *Geometry of Ridge Regression Illustrated.* The American Statistician, 35, 12-15.

Talwar, B.L. and Ashlock, L.T. (1970). *Selecting molding conditions for thermosets.* Society of Plastics Engineers Journal, 26, 42-46.

Thiart, C. (1990). *Collinearity and Consequences for estimation: A Study and Simulation.* MSc thesis, University of Cape Town.

Thiart, C. (1994). *Aspects of Estimation in the Linear Model with Special Reference to Collinearity.* PhD thesis, University of Cape Town.

Thiart. C., Dunne, T.T., Troskie, C.G. and Chalton, D.O. (1993). *A Simulation Study of Biased estimators against the Ordinary Least Squares estimator.* Communications in Statistics - Simulation and Computation, 22, 569-589.

Thisted, A.R. (1980). *A Critique of some Ridge Regression Methods: Comment.* Journal of the American Statistical Association, 75, 81-86.

Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society, Series B, 58, 267-288.

Troskie, C.G. and Chalton, D.O. (1996). *A Bayesian Estimate for the Constants in Ridge Regression.* South African Statistical Journal, 30, 119-137.

Vach, K., Sauerbrei, W. and Schumacher, M. (2001). *Variable selection and shrinkage: Comparison of some approaches*, Statistica Neerlandica, 55, 53-75

Van Houwelingen, J. C. (2001). *Shrinkage and penalized likelihood as methods to improve predictive accuracy*, Statistica Neerlandica, 55, 17-34

Van Houwelingen, J.C. and Le Cessie, S. (1990). *Predictive Value of Statistical Models.* Statistics in Medicine, 9, 1303-1325.

Venables, W.N. and Ripley, B.D. (1994). *Modern Applied Statistics with S-Plus.* Springer-Verlag, New York.

Vinod, H.D. (1976). *Application of New Ridge Methods to a Study of Bell System Scale Economies.* Journal of the American Statistical Association, 71, 835-841.

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

Walker, E. and Birch, B.J. (1988). *Influence Measures in Ridge Regression.* Technometrics, 30, 221-227.

Walker, G.S. and Page, J.C. (2001). *Generalized ridge regression and a generalization of the C_p statistic.* Journal of Applied Statistics, 28, 911-922.

Wan, A.T.K. (2002). *On generalized ridge regression estimators under collinearity and balanced loss.* Applied Mathematics and Computation, 129, 455-467.

Wencheko, E. (2000). *Estimation of the signal to noise in the linear regression model.* Statistical Papers, 41, 327-343.

Wencheko, E. (2001). *Comparison of regression estimators using Pitman measures of nearness* Statistical Papers, 42, 375-386.

Wetherill, G. B. (1986). *Regression Analysis with Applications.* Chapman and Hall Ltd, New York

Winchern, D.W. and Churchill, G.A. (1978). *A Comparison of Ridge Estimators.* Technometrics, 20, 301-311.

Wind, S.L. (1973). *An Empirical Bayes Approach to Multiple Linear Regression.* Annals of Statistics, 1, 93-103.

Zellner, A. and Vanele, W. (1974). *Bayes Stein Estimators for k means, Regression and Simultaneous Equation Models.* In: Feinberg, S.E. and Zellner, A. (eds) Studies in Bayesian Econometrics and Statistics. , Amsterdam, North Holland, 628-653.

Appendix A

The Simulation Program

Details of the program are provided in this appendix.

This appendix is sub-divided into three main sections. We first provide the program details for generating and saving the X matrix, proceed to the error matrices and finally present the main program which computes 24 shrinkage estimators and the corresponding relative efficiencies (REs) discussed in the simulation chapter. We fix the matrix of independent variables and the error term matrices. For each distribution, 500 columns of error terms are initially generated in R, written in Excel Comma Delimited files and then later read back into R from Excel.

Details of the main program include the following

- Reading the fixed error terms and a fixed matrix of independent variables from Excel Comma Delimited files,
- Declaration of variables used in the main loop,
- Computation of least squares estimates and the 24 biased estimators, MSEs, TMSEs and REs.

Included in the program, is a highlight of the program developed for estimation of the ridge constants with the PCdel1 solution.

Programming in R is such that all comments and guidelines follow a symbol # whereas the program commands are the explicit statements, written without the symbol # , hence we differentiate between commands and comments in the same manner to avoid confusion.

A.1 Program details for generating the X matrix

We generated the X matrix from the following program

```
# Generating x1 as a 100 × 1 column of ones.
x1=matrix(1,100,1)
# Generating x2 as a vector of values 20:29, repeated 10 times to make 100 observations
x2=as.matrix(rep(c(20:29),10))

# Generating x3 and replacing the 1st and 11th observations by -4 to avoid direct collinearity.
x3=x2-25
x3=replace(x3,1,-4)
x3=replace(x3,11,-4)

# Generating x4 as a vector of values (5,4,3,2,1,2,3,4,5,6), repeated 10 times to make 100 observations
x4=c(5,4,3,2,1,2,3,4,5,6)
x4=as.matrix(rep(x4,10))

# Generating x5 and replacing the 54th and 96th observations by 5 and 2 respectively.
x5=x2-x4+5
x5=replace(x5,96,2)
x5=replace(x5,54,5)

# Generating x6 and replacing the 38th and 100th observations by 3.
x6=x4-10
x6=replace(x6,38,3)
x6=replace(x6,100,3)

# Combining the 6 columns into a 6x6 matrix and writing the matrix to an Excel comma delimited file ’xmatrix’.
x=cbind(x1,x2,x3,x4,x5,x6)
write(x,file=”C:/xmatrix.csv”,ncolumn=6,append=F)
```

A.2 Program details for error terms

This program takes about two minutes to generate 500 columns of error terms from Laplace, Contaminated normal, Exponential and Normal distributions, with each distribution varied across the variance levels $\sigma^2 = 2$,25 and 100 respectively. For each distribution and variance level, the program fixes or saves 500 columns in a unique file in drives C and E.

```
n=100
nrep=500
p=5

errlap2=matrix(0,n,nrep)
errlap25=matrix(0,n,nrep)
errlap100=matrix(0,n,nrep)
errt2=matrix(0,n,nrep)
errt25=matrix(0,n,nrep)
errt100=matrix(0,n,nrep)
errexp2=matrix(0,n,nrep)
errexp25=matrix(0,n,nrep)
errexp100=matrix(0,n,nrep)
errn2=matrix(0,n,nrep)
errn25=matrix(0,n,nrep)
errn100=matrix(0,n,nrep)

for(i in 1:nrep)
 {
   for(i in 1:n)
    {
      # error Laplace for $\sigma^2 = 2$
       if(0< =runif(1,0,1)< =0.5)
        { errlap2[i,j]=1*log(2*runif(1,0,1))} else
        { errlap2[i,j]=-1*log(2*(1-runif(1,0,1)))}

          # error Laplace for $\sigma^2 = 25$
           if(0<=runif(1,0,1)<=0.5)
           { errlap25[i,j]=3.536*log(2*runif(1,0,1))} else
           { errlap25[i,j]=-3.536*log(2*(1-runif(1,0.1)))}
```

```
            # error Laplace for $\sigma^2 = 100$
            if(0< =runif(1,0,1)< =0.5)
                    { errlap100[i,j]=7.07*log(2*runif(1,0,1))} else errlap100[i,j]= -7.07*log(2*(1-runif(1,0,1)))}

           # error t for $\sigma^2 = 2$
            errt2[i,j]=rnorm(1,0,1)/sqrt(rchisq(1,4,ncp=0)/4)

           # error for Student's t for $\sigma^2 = 25$
            errt25[i,j]=rnorm(1,0,1)/sqrt(rchisq(1,2.083,ncp=0)/2.083)

            # error for Student's t for $\sigma^2 = 100$
            errt100[i,j]=rnorm(1,0,1)/sqrt(rchisq(1,2.020,ncp=0)/2.020)
           # error Exponential for $\sigma^2 = 2$
           errexp2[i,j]= rexp(1,rate=0.707)

           # error Exponential for $\sigma^2 = 25$
           errexp25[i,j]= rexp(1,rate=0.2)

           # error Exponential for $\sigma^2 = 100$
           errexp100[i,j]= rexp(1,rate=0.1)

           # error Normal for $\sigma^2 = 2$
          errn2[i,j]=rnorm(1,0,0.1414)

           # error matrix for $\sigma^2 = 25$
          errn25[i,j]=rnorm(1,0,5)

           # error matrix for $\sigma^2 = 100$
          errn100[i,j]=rnorm(1,0,10)
        }
}


# Writing the error files in Excel
write(errlap2,file="C:/errlap2.csv",ncolumn=nrep,append=F)
write(errlap25,file=”C:/errlap25.csv”,ncolumn=nrep,append=F)
write(errlap100,file=”C:/errlap100.csv”,ncolumn=nrep,append=F)

write(errt2,file=”C:/errt2.csv”,ncolumn=nrep,append=F)
write(errt25,file=”C:/errt25.csv”,ncolumn=nrep,append=F)
write(errt100,file=”C:/errt100.csv”,ncolumn=nrep,append=F)

write(errexp2,file=”C:/errexp2.csv”,ncolumn=nrep,append=F)
write(errexp25,file=”C:/errexp25.csv”,ncolumn=nrep,append=F)
write(errexp100,file=”C:/errexp100.csv”,ncolumn=nrep,append=F)

write(errn2,file=”C:/errn2.csv”,ncolumn=nrep,append=F)
write(errn25,file=”C:/errn25.csv”,ncolumn=nrep,append=F)
write(errn100,file=”C:/errn100.csv”,ncolumn=nrep,append=F)
```

A.3 The Main Program

We read the error terms and the matrix of independent variables from Excel Comma Delimited files into R, standardize the X matrix and then compute the coefficient estimates from Least Squares regression and the 24 shrinkage estimation techniques discussed in the simulation study. We then transform the coefficients back to the unstandardized form and compute the corresponding Mean Squared Error values and the Relative Efficiencies.

```
# Read the X matrix from an excel csv-file ”xmatrix”
xm=as.matrix(read.csv(”C:/xmatrix.csv”, sep=”,”,header=T))
# creating the x matrix that does not have a constant

# Specify the true coefficients
beta =as.matrix(c(10,0.4,0.5,0.25,0.3,4.5))

# Specify the number of observations, replications and the number of variables in the X matrix.
n=100
nrep=500
p=5
```

```
# Centre the columns of xm
center=apply(xm,2,mean)
xcntr=sweep(xm,2,center,”-”)

# Scale the X matrix
stde=0
se=0
for(i in 1:ncol(xm))
{
stde[i]=sqrt(crossprod(xcntr[,i]))
}
stde
# NB the first element in stde = zero (corresponding to the intercept) therefore we set
se=stde[,2:p+1] # to exclude the first value
for(i in 1:p)
{
se[i]=stde[i+1]
}
se

# Standardize X (with the first column of the centred matrix eliminated)
Z=sweep(as.matrix(xcntr[,-1]),2,se,”/”)

# Compute the Singular Value Decomposition of Z (the standardized X matrix)
sv=svd(Z)
v=sv$ v
u=sv$ u
d=(sv$ d)
eignv=d2̂

# Product matrices: (X’X) inverse and (Z’Z)
xtxinv=solve(t(xm)%*%xm)
ztz=t(Z)%*%Z
```

We read in the error terms, declare the variables and proceed to the main loop. Note that we consider four distributions of error terms and three levels of variance, hence the programs are developed such that each distribution at each variance level has its own independent program to avoid mistakes.

However. the programs are very similar; computations of estimates follow the exact same steps, the difference lies in labels used within the programs and the file names when the information is either being written or read from Excel.

Declaration statements and the main loop for the Normal distribution for $\sigma^2 = 2$

```
# Read the normal (0,0.1414) error terms from a csv file 'errn2
errn2=read.csv(”C:/errn2.csv”,sep=” ”,header=F)

# Declaration statements
Yn2             =       vector (mode=”list”,length=nrep)
Y               =       vector(mode=”list”,length=nrep)
ycntr           =       vector(mode=”list”,length=nrep)
xty             =       vector(mode=”list”,length=nrep)
zty             =       vector(mode=”list”,length=nrep)
alpha           =       vector(mode=”list”,length=nrep)
alpha1          =       vector(mode=”list”,length=nrep)
alpha11         =       vector(mode=”list”,length=nrep)
olsn2           =       vector(mode=”list”,length=nrep)
ols             =       vector(mode=”list”,length=nrep)
olscoef         =       vector(mode=”list”,length=nrep)
diffols         =       vector(mode=”list”,length=nrep)
diffhkb         =       vector(mode=”list”,length=nrep)
diffhkbnew      =       vector(mode=”list”,length=nrep)
diffhkbm        =       vector(mode=”list”,length=nrep)
diffhkbmnew     =       vector(mode=”list”,length=nrep)
diffkam         =       vector(mode=”list”,length=nrep)
diffkamnew      =       vector(mode=”list”,length=nrep)
diffkgm         =       vector(mode=”list”,length=nrep)
diffkgmnew      =       vector(mode=”list”,length=nrep)
diffkmed        =       vector(mode=”list”,length=nrep)
diffkmednew     =       vector(mode=”list”,length=nrep)
difflw          =       vector(mode=”list”,length=nrep)
difflwnew       =       vector(mode=”list”,length=nrep)
difflwm         =       vector(mode=”list”,length=nrep)
```

```
difflwmnew          =     vector(mode=”list”,length=nrep)
diffgrhkb           =     vector(mode=”list”,length=nrep)
diffgrhkbnew        =     vector(mode=”list”,length=nrep)
diffpcdel0          =     vector(mode=”list”,length=nrep)
diffpcdel1          =     vector(mode=”list”,length=nrep)
diffpcdel2          =     vector(mode=”list”,length=nrep)
diffgrtroskie       =     vector(mode=”list”,length=nrep)
diffgrtroskienew    =     vector(mode=”list”,length=nrep)
diffstein           =     vector(mode=”list”,length=nrep)
diffliumm           =     vector(mode=”list”,length=nrep)
diffliucl           =     vector(mode=”list”,length=nrep)
diffgliu            =     vector(mode=”list”,length=nrep)
kamxtxinv           =     vector(mode=”list”,length=nrep)
kamnewxtxinv        =     vector(mode=”list”,length=nrep)
kgmxtxinv           =     vector(mode=”list”,length=nrep)
kgmnewxtxinv        =     vector(mode=”list”,length=nrep)
kmedxtxinv          =     vector(mode=”list”,length=nrep)
kmednewxtxinv       =     vector(mode=”list”,length=nrep)
ridgehkb            =     vector(mode=”list”,length=nrep)
ridgehkbnew         =     vector(mode=”list”,length=nrep)
ridgehkbm           =     vector(mode=”list”,length=nrep)
ridgehkbmnew        =     vector(mode=”list”,length=nrep)
ridgelw             =     vector(mode=”list”,length=nrep)
ridgelwnew          =     vector(mode=”list”,length=nrep)
ridgelwstd          =     vector(mode=”list”,length=nrep)
ridgelwstdnew       =     vector(mode=”list”,length=nrep)
ridgelwm            =     vector(mode=”list”,length=nrep)
ridgelwmnew         =     vector(mode=”list”,length=nrep)
pcdel0              =     vector(mode=”list”,length=nrep)
pccoefdel0          =     vector(mode=”list”,length=nrep)
pcdel1              =     vector(mode=”list”,length=nrep)
pccoefdel1          =     vector(mode=”list”,length=nrep)
pcdel2              =     vector(mode=”list”,length=nrep)
liumm               =     vector(mode=”list”,length=nrep)
liucl               =     vector(mode=”list”,length=nrep)
gliu                =     vector(mode=”list”,length=nrep)
bta                 =     vector(mode=”list”,length=nrep)
```

```
btaliudmm          =     vector(mode=”list”,length=nrep)
btaliudcl          =     vector(mode=”list”,length=nrep)
btagliu            =     vector(mode=”list”,length=nrep)
kambta             =     vector(mode=”list”,length=nrep)
kamnewbta          =     vector(mode=”list”,length=nrep)
kgmbta             =     vector(mode=”list”,length=nrep)
kgmnewbta          =     vector(mode=”list”,length=nrep)
kmedbta            =     vector(mode=”list”,length=nrep)
kmednewbta         =     vector(mode=”list”,length=nrep)
pccoefdel2         =     vector(mode=”list”,length=nrep)
rcoeflw            =     vector(mode=”list”,length=nrep)
rcoeflwnew         =     vector(mode=”list”,length=nrep)
transcoefols       =     vector(mode=”list”,length=nrep)
transpcdel0        =     vector(mode=”list”,length=nrep)
transpcdel1        =     vector(mode=”list”,length=nrep)
transpcdel2        =     vector(mode=”list”,length=nrep)
betastein          =     vector(mode=”list”,length=nrep)
kamtransformed     =     vector(mode=”list”,length=nrep)
kgmtransformed     =     vector(mode=”list”,length=nrep)
kmedtransformed    =     vector(mode=”list”,length=nrep)
kamnewtransformed  =     vector(mode=”list”,length=nrep)
kgmnewtransformed  =     vector(mode=”list”,length=nrep)
kmednewtransformed=      vector(mode=”list”,length=nrep)
pc0transformed     =     vector(mode=”list”,length=nrep)
pc1transformed     =     vector(mode=”list”,length=nrep)
pc2transformed     =     vector(mode=”list”,length=nrep)
steintransformed   =     vector(mode=”list”,length=nrep)
liummtransformed   =     vector(mode=”list”,length=nrep)
liucltransformed   =     vector(mode=”list”,length=nrep)
gliutransformed    =     vector(mode=”list”,length=nrep)
transformedhkb     =     vector(mode=”list”,length=nrep)
transformedhkbm    =     vector(mode=”list”,length=nrep)
transformedlw      =     vector(mode=”list”,length=nrep)
transformedlwstd   =     vector(mode=”list”,length=nrep)
transformedlwm     =     vector(mode=”list”,length=nrep)
gtransformed       =     vector(mode=”list”,length=nrep)
grttransformed     =     vector(mode=”list”,length=nrep)
```

```
gridge              =   vector(mode="list",length=nrep)
gridgecoef          =   vector(mode="list",length=nrep)
grxtxinv            =   vector(mode="list",length=nrep)
grtxtxinv           =   vector(mode="list",length=nrep)
grbta               =   vector(mode="list",length=nrep)
grtbta              =   vector(mode="list",length=nrep)
transformedlwnew    =   vector(mode="list",length=nrep)
transformedlwnewstd=    vector(mode="list",length=nrep)
transformedlwmnew   =   vector(mode="list",length=nrep)
gtransformednew     =   vector(mode="list",length=nrep)
grtnewtransformed   =   vector(mode="list",length=nrep)
gridgenew           =   vector(mode="list",length=nrep)
gridgenewcoef       =   vector(mode="list",length=nrep)
grnewxtxinv         =   vector(mode="list",length=nrep)
grtnewxtxinv        =   vector(mode="list",length=nrep)
grnewbta            =   vector(mode="list",length=nrep)
grtnewbta           =   vector(mode="list",length=nrep)
khk                 =   vector(mode="list",length=nrep)
ktroskie            =   vector(mode="list",length=nrep)
kam                 =   vector(mode="list",length=nrep)
kgm                 =   vector(mode="list",length=nrep)
kmed                =   vector(mode="list",length=nrep)
khknew              =   vector(mode="list",length=nrep)
ktroskienew         =   vector(mode="list",length=nrep)
kamnew              =   vector(mode="list",length=nrep)
kgmnew              =   vector(mode="list",length=nrep)
kmednew             =   vector(mode="list",length=nrep)
klwnew              =   0
klwnewstd           =   0
klwmnew             =   0
khkbnew             =   0
khkbmnew            =   0
klw                 =   0
klwstd              =   0
klwm                =   0
khkb                =   0
khkbm               =   0
```

```
dg                  =     vector(mode=”list”,length=nrep)
liuztz              =     solve(ztz + diag(1,p,p))
ybar                =     0
ssy                 =     0
sigma               =     0
sigmasq             =     0
sigma1              =     0
sigmasq1            =     0
c                   =     0
dmn                 =     0
dcl                 =     0

Set the MSEs, TMSEs and REs to zero
m                   =     0
mseols              =     0
msehkb              =     0
msehkbm             =     0
muselw              =     0
mselwm              =     0
msehkbnew           =     0
msehkbmnew          =     0
mselwnew            =     0
mselwmnew           =     0
msepcdel0           =     0
msepcdel1           =     0
msepcdel2           =     0
msegrtroskie        =     0
msegrhkb            =     0
msegrtroskienew     =     0
msegrhkbnew         =     0
msestein            =     0
mseliumm            =     0
mseliucl            =     0
msegliu             =     0
msekam              =     0
msekgm              =     0
msekmed             =     0
```

```
msekgmnew          =    0
msekmednew         =    0
tmseols            =    0
tmsehkb            =    0
tmsehkbm           =    0
tmselw             =    0
tmselwm            =    0
tmsehkbnew         =    0
tmsehkbmnew        =    0
tmselwnew          =    0
tmselwmnew         =    0
tmsepcdel0         =    0
tmsepcdel1         =    0
tmsepcdel2         =    0
tmsegrtroskie      =    0
tmsegrhkb          =    0
tmsegrtroskienew   =    0
tmsegrhkbnew       =    0
tmsestein          =    0
tmseliumm          =    0
tmseliucl          =    0
tmsegliu           =    0
tmsekam            =    0
tmsekgm            =    0
tmsekmed           =    0
tmsekamnew         =    0
tmsekgmnew         =    0
tmsekmednew        =    0
reols              =    0
rehkb              =    0
rehkbm             =    0
relw               =    0
relwm              =    0
rehkbnew           =    0
rehkbmnew          =    0
relwnew            =    0
relwmnew           =    0
```

```
repcdel0          =     0
repcdel1          =     0
repcdel2          =     0
regrtroskie       =     0
regrhkb           =     0
regrtroskienew    =     0
regrhkbnew        =     0
restein           =     0
reliumm           =     0
reliucl           =     0
regliu            =     0
rekam             =     0
rekgm             =     0
rekmed            =     0
rekamnew          =     0
rekgmnew          =     0
rekmednew         =     0
```

```
for (i in 1:nrep)
{
Yn2[[i]]=xm%*% beta+as.matrix(errn2[,i])

# Centre and Scale Yn2
ybar[i]=mean(Yn2[[i]])
ycntr[[i]]=Yn2[[i]]-ybar[i]
ssy[i]= sqrt(crossprod(ycntr[[i]]))
Y[[i]]=sweep(ycntr[[i]],2,ssy[i],”/”)

# the product matrix (Z Y correlation)
zty[[i]]=t(z)% * % Y[[i]]

# LEAST SQUARES
# NB Both Y and X are standardized

olsn2[[i]]=lm(Y[[i]] ˜ z)

# Extract the coefficients excluding the constant
```

```
olscoef[[i]]=(olsn2[[i]]$ coef[-1])

# transform back the coefficients. This
transcoefols[[i]]=olscoef[[i]]*ssy[i]/se

# Write the transformed coefficients on an Excel comma delimited file for reference.
write(transcoefols[[i]], file = ”C:/My documents/ Simulation/olsn2coef.csv”,
ncolumns =nrep, append = T)

# Calculate the differences between the transformed and true coefficients and write them to
an Excel comma delimited file.
diffols[[i]]=transcoefols[[i]]-beta[-1]
write(diffols[[i]],file=”C:/My documents/Simulation/ olsn2diff.csv”,
ncolumns=nrep,append=T)

# Calculate the sum of squared differences
mseols=mseols+diffols[[i]] ^ 2
tmseols=sum(mseols)

# Calculate the relative efficiency of least squares estimates.
reols=tmseols/tmseols

# Extract the least squares variance for computation of the biasing parameters
sigma[i]=summary(olsn2[[i]])$ sigma
sigmasq[i]=sigma[i] ^ 2
```

# PRINCIPAL COMPONENT REGRESSION

```
# PC delete 0 (zero roots deleted: same as OLS).
W=z%*% v
pcdel0[[i]]=lm(Y[[i]] ~ W-1)
alpha[[i]]=pcdel0[[i]]$ coef
transpcdel0[[i]]=v%*% as.matrix(alpha[[i]])
pc0transformed[[i]]=transpcdel0[[i]]*ssy[i]/se
write(pc0transformed[[i]], file = ”C:/My documents/Simulation/pcdel0n2coef.csv”,
ncolumns =nrep, append = T)
diffpcdel0[[i]]=pc0transformed[[i]]-beta[-1]
```

```
msepcdel0=msepcdel0+ diffpcdel0[[i]] ˆ 2
tmsepcdel0=sum(msepcdel0)
repcdel0=tmseols/tmsepcdel0
write(diffpcdel0[[i]],file=”C:/My documents/ Simulation/pcdel0n2diff.csv”,
ncolumns=nrep,append=T)

# Extracting sigma-squared
sigma1[i]=summary(pcdel0[[i]])$ sigma
sigmasq1[i]=sigma1[i] ˆ 2

# F statistic
F[[i]]= (eignv*alpha[[i]] ˆ 2)/sigmasq1[i]

# PC delete1
# d =diag(d,p,p)
d1=diag(d,p,p)
d1[p,p]=0
W1= u%*% d1
pcdel1[[i]]=lm(Y[[i]] ˜ W1)
alph1[[i]]=pcdel1[[i]]$ coef[-1]
alph1[[i]][p]=0
transpcdel1[[i]]=v%*% as.matrix(alph1[[i]])
pc1transformed[[i]]=transpcdel1[[i]]*ssy[i]/se
write(pc1transformed[[i]], file = ”C:/My documents/Simulation/pcdel1n2coef.csv”,
ncolumns =nrep, append = T)
diffpcdel1[[i]]=pc1transformed[[i]]-beta[-1]
msepcdel1=msepcdel1+ diffpcdel1[[i]] ˆ 2
tmsepcdel1=sum(msepcdel1)
repcdel1=tmseols/tmsepcdel1
write(diffpcdel1[[i]],file=”C:/My documents/ Simulation/pcdel1n2diff.csv”,
ncolumns=nrep,append=T)

# PC delete2
d2=d1
d2[p-1,p-1]=0
W2= u%*% d2
pcdel2[[i]]¡-lm(Y[[i]] ˜ W2)
```

```
alph2[[i]]=pcdel2[[i]]$ coef
alph2[[i]][p+1]=0
alph2[[i]][p]=0
transpcdel2[[i]]=v%*% as.matrix(alph2[[i]][-1])
pc2transformed[[i]]=transpcdel2[[i]]*ssy[i]/se
write(pc2transformed[[i]], file = "C:/My documents/Simulation/pcdel2n2.csv", ncolumns =nrep,
append = T)
diffpcdel2[[i]]=pc2transformed[[i]]-beta[-1]
msepcdel2=msepcdel2+ diffpcdel2[[i]] ^ 2
tmsepcdel2=sum(msepcdel2)
repcdel2=tmseols/tmsepcdel2
write(diffpcdel2[[i]],file="C:/My documents/Simulation/pcdel2n2diff.csv",
ncolumns=nrep,append=T)

# TRADITIONAL RIDGE AND GENERALIZED RIDGE PARAMETERS

khkb[i]=sigmasq[i]*(p)/(sum(olscoef[[i]] ^ 2))
khkbm[i] =sigmasq[i]*(p-2)/(sum(olscoef[[i]] ^ 2))
klw[i]=sigmasq[i]*(p)/(sum(olscoef[[i]] ^ 2*eignv))
klwm[i]=((p-2)*sigmasq[i]*sum(svd(Z)$ d ^ 2))/(p*t(olscoef[[i]])% * %(t(Z)% * % (Z))% *%
olscoef[[i]])
khk[[i]]= diag(sigmasq[i]/alpha[[i]] ^ 2,p,p)
ktroskie[[i]]= diag((eignv /(F[[i]]+1)),p,p)
kam[[i]]= diag((sum(sigmasq[i]/alpha[[i]] ^ 2))*(1/p),p,p)
kgm[[i]]=diag(sigmasq[i]/(prod(alpha[[i]] ^ 2)) ^ (1/p),p,p)
kmed[[i]]=diag(median(sigmasq[i]/((alpha[[i]] ^ 2))),p,p)


# NEW RIDGE AND GENERALIZED RIDGE PARAMETERS
alph11[[i]]=alph1[[i]][1:p-1]
khkbnew[i]=sigmasq1[i]*(p)/(sum(alph11[[i]] ^ 2))
khkbmnew[i] =sigmasq1[i]*((p-1)-2)/(sum(alph11[[i]] ^ 2))
klwnew[i]=sigmasq1[i]*(p)/(sum(alph1[[i]] ^ 2*eignv))
klwmnew[i]=(((p-1)-2)*sigmasq1[i]*sum(svd(z)$d ^ 2))/((p-1)*t (alph1[[i]])%*%(t(z)%*%
(z))%*%alph1[[i]])
kgrtnew[[i]]= diag(eignv /((eignv*(alph1[[i]]2̂)/sigmasq1[i])+1),p,p)
kmednew[[i]]=diag(median(sigmasq1[i]/((alph11[[i]]2̂))),p,p)
```

```
kamnew[[i]]= diag((sum(sigmasq1[i]/alph11[[i]] ^ 2))*(1/(p-1)),p,p)
kgmnew[[i]]=diag(sigmasq1[i]/(prod(alph11[[i]] ^ 2)) ^ (1/(p-1)),p,p)
kgrhknew[[i]]= diag(sigmasq1[i]/alph1[[i]] ^ 2,p,p)
kgrhknew[[i]]=diag(replace(diag(kgrhknew[[i]]),p,min(diag(kgrhknew[[i]])[-p])))
```

# RIDGE COEFFICIENTS, MSEs, TMSEs and REs

# khkb (Hoerl, Kennard and Baldwin, 1970)

```
# Estimates
ridgehkb[[i]]=(solve(ztz +diag( khkb[i],p,p))) % *% zty[[i]]

# transform the coefficients back into the unstandardized form and write them to an excel file.
transformedhkb[[i]]=ridgehkb[[i]]*ssy[i]/se
write(transformedhkb[[i]], file = "C:/My documents/ Final Simulation/hkbn2coef.csv",
ncolumns =nrep, append = T)

# Calculate the differences between the hkb ridge coefficients and true ones.
diffhkb[[i]]=transformedhkb[[i]]-beta[-1]
write(diffhkb[[i]],file="C:/My documents/Final Simulation/hkbn2diff.csv",
ncolumns=nrep,append=T)

Compute MSEs, TMSE and RE for ridgehkb. msehkb=msehkb+diffhkb[[i]] ^ 2
tmsehkb=sum(msehkb)
rehkb=tmseols/tmsehkb
```

# khkbnew)

```
# Estimates
ridgehkbnew[[i]]=(solve(ztz +diag( khkbnew[i],p,p))) % *% zty[[i]]

# transform the coefficients back into the unstandardized form and write them to an excel file.
transformedhkbnew[[i]]=ridgehkbnew[[i]]*ssy[i]/se
write(transformedhkb[[i]], file = "C:/My documents/ Final Simulation/hkbnewn2coef.csv",
ncolumns =nrep, append = T)

# Calculate the differences between the hkb ridge coefficients and true ones.
diffhkbnew[[i]]=transformedhkbnew[[i]]-beta[-1]
write(diffhkbnew[[i]],file="C:/My documents/Final Simulation/hkbnewn2diff.csv",
ncolumns=nrep,append=T)
```

Compute MSEs, TMSE and RE for ridgehkbnew. msehkbnew=msehkbnew+diffhkbnew[[i]] ^ 2

```
tmsehkbnew=sum(msehkbnew)
rehkbnew=tmseols/tmsehkbnew
```

```
# khkbm (Brown, 1993)
# Estimates
ridgehkbm[[i]]=(solve(ztz +diag( khkbm[i],p,p)))% * % zty[[i]]

# transform the coefficients back into the unstandardized form and write them to an excel file.
transformedhkbm[[i]]= ridgehkbm[[i]]*ssy[i]/se
write(transformedhkbm[[i]], file = "C:/My documents/Final Simulation/hkbmn2coef.csv",
ncolumns =nrep, append = T)

# Computing the deviations of the estimates from the true coefficients
diffhkbm[[i]]=transformedhkbm[[i]]-beta[-1]

# MSE, TMSE and RE
msehkbm=msehkbm+diffhkbm[[i]] ^ 2
tmsehkbm=sum(msehkbm)
rehkbm=tmseols/tmsehkbm
write(diffhkbm[[i]],file="C:/My documents/ Final Simulation/hkbmn2diff.csv",
ncolumns=nrep,append=T)
```

```
# khkbmnew
# Estimates
ridgehkbmnew[[i]]=(solve(ztz +diag( khkbmnew[i],p,p)))% * % zty[[i]]

# transform the coefficients back into the unstandardized form and write them to an excel file.
transformedhkbmnew[[i]]= ridgehkbmnew[[i]]*ssy[i]/se
write(transformedhkbmnew[[i]], file = "C:/My documents/Final Simulation/hkbmnewn2coef.csv",
ncolumns =nrep, append = T)

# Computing the deviations of the estimates from the true coefficients
diffhkbmnew[[i]]=transformedhkbmnew[[i]]-beta[-1]
```

```
# MSE. TMSE and RE
msehkbmnew=msehkbmnew+diffhkbmnew[[i]] ^ 2
tmsehkbmnew=sum(msehkbmnew)
rehkbmnew=tmseols/tmsehkbmnew
write(diffhkbmnew[[i]],file="C:/My documents/ Final Simulation/hkbmnewn2diff.csv",
ncolumns=nrep,append=T)

# kLW (Lawless and Wang, 1976)
# Estimates ridgelw[[i]]=(solve(ztz +diag( klw[i],p,p)))% * % zty[[i]]

# transforming back the coefficients
transformedlw[[i]]=ridgelw[[i]]*ssy[i]/se
write(transformedlw[[i]], file = "C:/My documents/Final Simulation/lwn2coef.csv",
ncolumns =nrep, append = T)

# Calculating the differences
difflw[[i]]=transformedlw[[i]]-beta[-1]

# MSE, TMSE and RE
mselw=mselw+difflw[[i]] ^ 2
tmselw=sum(mselw)
relw=tmseols/tmselw
write(difflw[[i]],file="C:/My documents/Final Simulation/lwn2diff.csv",
ncolumns=nrep,append=T)

# klwnew
# Estimates ridgelwnew[[i]]=(solve(ztz +diag( klwnew[i],p,p)))% * % zty[[i]]

# transforming back the coefficients
transformedlwnew[[i]]=ridgelwnew[[i]]*ssy[i]/se
write(transformedlwnew[[i]], file = "C:/My documents/Final Simulation/lwnewn2coef.csv",
ncolumns =nrep, append = T)

# Calculating the differences
difflwnew[[i]]=transformedlwnew[[i]]-beta[-1]

# MSE, TMSE and RE
```

```
mselwnew=mselwnew+difflwnew[[i]] ^ 2
tmselwnew=sum(mselwnew)
relwnew=tmseols/tmselwnew
write(difflwnew[[i]],file="C:/My documents/Final Simulation/lwnewn2diff.csv",
ncolumns=nrep,append=T)

# klwm (Brown, 1993)
# Estimates
ridgelwm[[i]]=(solve(ztz +diag( klwm[i],p,p)))% * % zty[[i]]

# transforming back the coefficients
transformedlwm[[i]]= ridgelwm[[i]]*ssy[i]/se
write(transformedlwm[[i]], file = "C:/My documents/Final Simulation/lwmn2coef.csv",
ncolumns =nrep, append = T)

# Computing the deviations
difflwm[[i]]=transformedlwm[[i]]-beta[-1]

# MSE, TMSE and RE
mselwm=mselwm+difflwm[[i]] ^ 2
tmselwm=sum(mselwm)
relwm=tmseols/tmselwm
write(difflwm[[i]],file="C:/ My documents/Final Simulation/lwmn2diff.csv",
ncolumns=nrep,append=T)

# klwmnew
# Estimates
ridgelwmnew[[i]]=(solve(ztz +diag( klwmnew[i],p,p)))% * % zty[[i]]

# transforming back the coefficients
transformedlwmnew[[i]]= ridgelwmnew[[i]]*ssy[i]/se
write(transformedlwmnew[[i]], file = "C:/My documents/Final Simulation/lwmnewn2coef.csv",
ncolumns =nrep, append = T)

# Computing the deviations
difflwmnew[[i]]=transformedlwmnew[[i]]-beta[-1]
```

```
# MSE, TMSE and RE
mselwmnew=mselwmnew+difflwmnew[[i]] ˆ 2
tmselwmnew=sum(mselwmnew)
relwmnew=tmseols/tmselwmnew
write(difflwmnew[[i]],file=”C:/ My documents/Final Simulation/lwmnewn2diff.csv”,
ncolumns=nrep,append=T)

# Kam (Kibria, 2003)
# Estimates
kamxtxinv[[i]]=solve(ztz + kam[[i]])
kambta[[i]]= kamxtxinv[[i]]%*% zty[[i]]

# transforming back the coefficients
kamtransformed[[i]]=kambta[[i]]*ssy[i]/se
write(kamtransformed[[i]], file = ”C:/My documents/Simulation/kamn2coef.csv”,
ncolumns =nrep, append = T)

# MSE, TMSE and RE
diffkam[[i]]= beta[-1]-kamtransformed[[i]]
msekam=msekam+diffkam[[i]] ˆ 2
tmsekam=sum(msekam)
rekam=tmseols/tmsekam
write(diffkam[[i]],file=”C:/My documents/Simulation/kamn2diff.csv”,
ncolumns=nrep,append=T)

# Kamnew
# Estimates
kamnewxtxinv[[i]]=solve(ztz + kamnew[[i]])
kamnewbta[[i]]= kamnewxtxinv[[i]]%*% zty[[i]]

# transforming back the coefficients
kamnewtransformed[[i]]=kamnewbta[[i]]*ssy[i]/se
write(kamnewtransformed[[i]], file = ”C:/My documents/Simulation/kamnewn2coef.csv”,
ncolumns =nrep, append = T)

# MSE, TMSE and RE
diffkamnew[[i]]= beta[-1]-kamnewtransformed[[i]]
```

```
msekamnew=msekam+diffkamnew[[i]] ^ 2
tmsekamnew=sum(msekamnew)
rekamnew=tmseols/tmsekamnew
write(diffkam[[i]],file=”C:/My documents/Simulation/kamnewn2diff.csv”,
ncolumns=nrep,append=T)

# Ridge Kibia (Kgm)
# The Estimates
kgmxtxinv[[i]]=solve(ztz + kgm[[i]])
kgmbta[[i]]= kgmxtxinv[[i]]%*% zty[[i]]

# The transformed estimates
kgmtransformed[[i]]=kgmbta[[i]]*ssy[i]/se
write(kgmtransformed[[i]], file = ”C:/My documents/Simulation/kgmn2coef.csv”,
ncolumns =nrep, append = T)

# MSE, TMSE and RE
diffkgm[[i]]= beta[-1]-kgmtransformed[[i]]
msekgm=msekgm+diffkgm[[i]] ^ 2
tmsekgm=sum(msekgm)
rekgm=tmseols/tmsekgm
write(diffkgm[[i]],file=”C:/My documents/Simulation/kgmn2diff.csv”,
ncolumns=nrep,append=T)

# (Kgmnew)
# The Estimates
kgmnewxtxinv[[i]]=solve(ztz + kgmnew[[i]])
kgmnewbta[[i]]= kgmnewxtxinv[[i]]%*% zty[[i]]

# The transformed estimates
kgmnewtransformed[[i]]=kgmnewbta[[i]]*ssy[i]/se
write(kgmnewtransformed[[i]], file = ”C:/My documents/Simulation/kgmnewn2coef.csv”,
ncolumns =nrep, append = T)

# MSE, TMSE and RE
diffkgmnew[[i]]= beta[-1]-kgmnewtransformed[[i]]
msekgmnew=msekgmnew+diffkgmnew[[i]] ^ 2
```

```
tmsekgmnew=sum(msekgmnew)
rekgmnew=tmseols/tmsekgmnew
write(diffkgmnew[[i]],file=”C:/My documents/Simulation/kgmnewn2diff.csv”,
ncolumns=nrep,append=T)

# Kmed(Kibria, 2003)
The estimates
kmedxtxinv[[i]]=solve(ztz + kmed[[i]])
kmedbta[[i]]= kmedxtxinv[[i]]%*% zty[[i]]
kmedtransformed[[i]]=kmedbta[[i]]*ssy[i]/se
write(kmedtransformed[[i]], file = ”C:/My documents/Simulation/kmedn2coef.csv”,
ncolumns =nrep, append = T)

# Computing the MSE, TMSE and RE values
diffkmed[[i]]= beta[-1]-kmedtransformed[[i]]
msekmed=msekmed+diffkmed[[i]] ^ 2
tmsekmed=sum(msekmed)
rekmed=tmseols/tmsekmed
write(diffkmed[[i]],file=”C:/My documents/Simulation/kmedn2diff.csv”,
ncolumns=nrep,append=T)

# kmednew
The estimates
kmednewxtxinv[[i]]=solve(ztz + kmednew[[i]])
kmednewbta[[i]]= kmednewxtxinv[[i]]%*% zty[[i]]
kmednewtransformed[[i]]=kmednewbta[[i]]*ssy[i]/se
write(kmednewtransformed[[i]], file = ”C:/My documents/Simulation/kmednewn2coef.csv”,
ncolumns =nrep, append = T)

# Computing the MSE, TMSE and RE values
diffkmednew[[i]]= beta[-1]-kmednewtransformed[[i]]
msekmednew=msekmednew+diffkmednew[[i]] ^ 2
tmsekmednew=sum(msekmednew)
rekmednew=tmseols/tmsekmednew
write(diffkmednew[[i]],file=”C:/My documents/Simulation/kmednewn2diff.csv”,
ncolumns=nrep,append=T)
```

```
# GENERALIZED RIDGE
# khk (Hoerl and Kennard, 1970a)

# Esimates
grxtxinv[[i]]=solve(ztz + khk[[i]])
grbta[[i]]= grxtxinv[[i]]% *% zty[[i]]

# transform the coefficients back into the unstandardized form and write them to an excel file.
gtransformed[[i]]=grbta[[i]]*ssy[i]/se
write(gtransformed[[i]], file = ”C:/My documents/Final Simulation/grhkbn2coef.csv”,
ncolumns =nrep, append = T)

# khknew

# Esimates
grnewxtxinv[[i]]=solve(ztz + khknew[[i]])
grnewbta[[i]]= grnewxtxinv[[i]]% *% zty[[i]]

# transform the coefficients back into the unstandardized form and write them to an excel file.
gnewtransformed[[i]]=grnewbta[[i]]*ssy[i]/se
write(gnewtransformed[[i]], file = ”C:/My documents/Final Simulation/grhkbnewn2coef.csv”,
ncolumns =nrep, append = T)

# MSE, TMSE and RE
diffgrhkbnew[[i]]= beta[-1]-gnewtransformed[[i]]
msegrhkbnew=msegrhkbnew+diffgrhkbnew[[i]] ^ 2
tmsegrhkbnew=sum(msegrhkbnew)
regrhkbnew=tmseols/tmsegrhkbnew
write(diffgrhkbnew[[i]],file=”C:/My documents/Final Simulation/grhkbnewn2diff.csv”,
ncolumns=nrep,append=T)

# ktroskie (Chalton and Troskie, 1996)
# Esimates
grtxtxinv[[i]]=solve(ztz + ktroskie[[i]])
grtbta[[i]]= grtxtxinv[[i]]% *% zty[[i]]
grttransformed[[i]]=grtbta[[i]]*ssy[i]/se
write(grttransformed[[i]], file = ”C:/ My documents/Final Simulation/grtn2coef.csv”,
ncolumns =nrep, append = T)

# MSE, TMSE and RE
diffgrtroskie[[i]]= beta[-1]-grttransformed[[i]]
msegrtroskie=msegrtroskie+diffgrtroskie[[i]] ˆ 2
tmsegrtroskie=sum(msegrtroskie)
regrtroskie=tmseols/tmsegrtroskie
write(diffgrtroskie[[i]],file=”C:/My documents/Final Simulation/grtn2diff.csv”,
ncolumns=nrep,append=T)

# ktroskienew
# Esimates
grtnewxtxinv[[i]]=solve(ztz + ktroskienew[[i]])
grtnewbta[[i]]= grtnewxtxinv[[i]]% *% zty[[i]]
grtnewtransformed[[i]]=grtnewbta[[i]]*ssy[i]/sc
write(grtnewtransformed[[i]], file = ”C:/ My documents/Final Simulation/grtnewn2coef.csv”,
ncolumns =nrep, append = T)

# MSE, TMSE and RE
diffgrtroskienew[[i]]= beta[-1]-grtnewtransformed[[i]]
msegrtroskienew=msegrtroskienew+diffgrtroskienew[[i]] ˆ 2
tmsegrtroskienew=sum(msegrtroskienew)
regrtroskienew=tmseols/tmsegrtroskienew
write(diffgrtroskienew[[i]],file=”C:/My documents/Final Simulation/grtnewn2diff.csv”,
ncolumns=nrep,append=T)

# STEIN ESTIMATION
# The stein constants (James and Stein, 1961)
c[i]=max(0,1-((p-2)*(n-p)*sigmasq/(n-p+2)*(sum(olscoef[[i]])) ˆ 2))

# Esimates
betastein[[i]]=c[i]*olscoef[[i]]
steintransformed[[i]]=betastein[[i]]*ssy[i]/sc
write(steintransformed[[i]], file = ”C:/My documents/ Simulation/steinn2coef.csv”,
ncolumns =nrep, append = T)

# MSE, TMSE and RE
```

```
diffstein[[i]]=steintransformed[[i]]-beta[-1]
msestein=msestein+diffstein[[i]] ^ 2
tmsestein=sum(msestein)
restein=tmseols/tmsestein
write(diffstein[[i]],file="C:/My documents/Simulation/ steinn2diff.csv",
ncolumns=nrep,append=T)

# LIU ESTIMATION
# The Liu constant (dmn; Liu, 1993)
dmn[i]=1-sigmasq[i]*(sum(1/eignv*(eignv+1))/sum(alpha[[i]] ^ 2/(eignv+1) ^ 2))

# Esimates
liuztz=solve(ztz + diag(1,p,p))
liumn[[i]]=zty[[i]]+(diag(dmn[i],p,p)%*% olscoef[[i]])
btaliudmn[[i]]=liuztz%*% liumn[[i]]
liumntransformed[[i]]=btaliudmn[[i]]*ssy[i]/se
write(liumntransformed[[i]], file = "C:/My documents/Simulation/ liumnn2coef.csv",
ncolumns =nrep, append = T)

# MSE, TMSE and RE
diffliumn[[i]]= beta[-1]- liumntransformed[[i]]
mseliumn=mseliumn+diffliumn[[i]] ^ 2
tmseliumn=sum(mseliumn)
reliumn=tmseols/tmseliumn
write(diffliumn[[i]],file="C:/My documents/Simulation/ liumnn2diff.csv",
ncolumns=nrep,append=T)

# dcl(cl criterion; Liu, 1993)
The constants
dcl[i]=1-sigmasq[i]*(sum(1/(eignv+1))/sum(eignv*alpha[[i]] ^ 2/(eignv+1) ^ 2))

# Esimates
liucl[[i]]=zty[[i]]+(diag(dcl[i],p,p)%*% olscoef[[i]])
btaliudcl[[i]]=liuztz%*% liucl[[i]]
liucltransformed[[i]]=btaliudcl[[i]]*ssy[i]/se
write(liucltransformed[[i]], file = "C:/My documents/Simulation/ liucln2coef.csv",
ncolumns =nrep, append = T)
```

```
# MSE, TMSE and RE
diffliucl[[i]]= beta[-1]- liucltransformed[[i]]
mseliucl=mseliucl+diffliucl[[i]] ^ 2
tmseliucl=sum(mseliucl)
reliucl=tmseols/tmseliucl
write(diffliucl[[i]],file=”C:/My documents/Simulation/liucln2diff.csv”,
ncolumns=nrep,append=T)

# GENERALIZED LIU
# The generalized Liu constant (Liu, 1993)
dg[[i]]=1-(sigmasq[i]*(eignv+1)/(eignv*alpha[[i]] ^ 2))
# The estimates
gliu[[i]]=zty[[i]]+(diag(dg[[i]],p,p)%*% alpha[[i]])
btagliu[[i]]=liuztz%*% gliu[[i]]
gliutransformed[[i]]=btagliu[[i]]*ssy[i]/se
write(gliutransformed[[i]], file = ”C:/My documents/ Simulation/gliun2coef.csv”, ncolumns
=nrep, append = T)

# MSE, TMSE and RE
diffgliu[[i]]= beta[-1]- gliutransformed[[i]]
msegliu=msegliu+diffgliu[[i]] ^ 2
tmsegliu=sum(msegliu)
regliu=tmseols/tmsegliu
write(diffgliu[[i]],file=”C:/My documents/Simulation/gliun2diff.csv”,
ncolumns=nrep,append=T)
}
mse=cbind((mseols),(msehkb),(msehkbm),(mselw),(mselwm),
(msegrtroskie),(msegrhkb),(msehkbnew),(msehkbmnew),(mselwnew),(mselwmnew),
(msegrtroskienew),(msegrhkbnew),(msepcdel1),(msepcdel2),
(mseliumm),(mseliucl),(msegliu),(msestein),(msekam),(msekgm),(msekmed), (msekamnew),(msekgmnev
mse=matrix(mse,p,25)

write(t(mse), file = ”C:/My documents/Simulation/msen2.csv”,
ncolumns =25, append =T)

tmse=cbind((tmseols),(tmsehkb),(tmsehkbm),(tmselw),(tmselwm),
(tmsegrtroskie),(tmsegrhkb),(tmsehkbnew),(tmsehkbmnew),(tmselwnew),(tmselwmnew),
(tmsegrtroskienew),(tmsegrhkbnew),(tmsepcdel1),(tmsepcdel2),
(tmseliumm),(tmseliucl),(tmsegliu),(tmsestein),(tmsekam),(tmsekgm),
(tmsekmed),(tmsekamnew),(tmsekgmnew),
(tmsekmednew))

write(t(tmse), file = ”C:/My documents/Simulation/tmsen2.csv”,
ncolumns =1, append =T)

re=cbind((reols),(rehkb),(rehkbm),(relw),(relwm),
(regrtroskie),(regrhkb),(rehkbnew),(rehkbmnew),(relwnew),(relwmnew),
(regrtroskienew),(regrhkbnew),(repcdel1),(repcdel2),
(reliumm),(reliucl),(regliu),(restein),(rekam),(rekgm),(rekmed), (rekamnew),(rekgmnew),(rekmednew))

write(t(re), file = ”C:/My documents/Simulation/ren2.csv”,
ncolumns =1, append =T)
```

Appendix B

Distributions

A summary of the general notion underlying the probability distributions used in the simulation study is presented. We first highlight the most important concepts in probability distribution theory and statistical inference and later describe each of the distributions used in the simulation study.

B.1 Fundamental concepts of probability distributions

Ideally, probability distributions are critically important in several practical problems; particularly in description of random variables and the corresponding patterns. The following concepts play a vital role in the theory of probability and its application.

B.1.1 Discrete and continuous probability distributions

The probability distributions associated with random variables that take on a countable number of values are discrete probability distributions and those that are associated with an uncountable number of values are continuous. We refer to the corresponding functions as mass and density functions, denoted by $p(x)$ and $f(x)$ for discrete and continuous probabilities respectively.

Importantly, for a discrete random variable X that assumes the values x_i,

- $0 \leq p(x_i) \leq 1 \qquad \forall \quad x_i$
- $\sum_{all \quad x_i} p(x_i) = 1$

Similarly, for a continuous random variable X that takes on values that range between a and b $\quad (a \leq x \leq b)$,

- $f(x) \geq 0$

- $\int_a^b f(x)d(x) = 1$

Probability distributions associated with one random variable are usually referred to as univariate whereas those that are associated with more than one random variable are known as multivariate distributions. A bivariate distribution is a special case of multivariate distribution where two random variables are considered simultaneously.

B.1.2 Distribution Function

If X is a continuous random variable with the density function $f(x)$ or a discrete random variable with the mass function $p(x)$, then the function that gives the probability that X takes on values less than or equal to x is called the distribution function, denoted by $F(x)$.

Suppose u and x are random variables such that

$$\begin{aligned} u &= F(x) \\ x &= G(u) \\ x &= G(F(x)) \\ u &= F(G(F(x))) \end{aligned}$$

then, in statistical terms, we say inverse distribution function $F(x)^{-1}$ maps inversely from u into x.

If the derivative exists, the distribution function associated with X may be differentiated once with respect to x to give the probability density function of x.

$$f(x) = \frac{\partial(F(x))}{\partial x} \tag{B.1}$$

In this thesis, we use some of the inverted distribution functions $F(x)^{-1}$ to generate error terms from some of the distributions.

B.1.3 Moments

Generally, the n^{th} moment taken about an arbitrary point b is defined by

$$\int (x-b)^n f(x)d(x) \qquad or \qquad \sum (x-b)^n p(x)$$

Special cases of which $b = 0$ and $b = \mu$ are referred to as raw and central moments respectively.

- **Raw moments**

  The n^{th} moment about zero (raw moment) is defined by

  $$\mu'_n = \int x^n f(x)d(x) \qquad or \qquad \mu'_n = \sum x^n p(x)$$

The first raw moment $\mu_1' = \sum xp(x)$ *or* $\int xf(x)d(x) = E(X)$ for discrete and continuous variables respectively.

- **Central moments**

  The n^{th} moment taken about the mean (central moment) is defined by

  $$\mu_n = \int (x-\mu)^n f(x)d(x) \qquad or \qquad \mu_n = \sum (x-\mu)^n p(x)$$

  The second central moment is also known as the variance

  $$\mu_2 = \int (x-\mu)^2 f(x)d(x) \qquad or \qquad \mu_2 = \sum (x-\mu)^2 p(x) = E[(x-\mu)^2]$$

B.1.4 Skewness

Skewness defines the degree of asymmetry of a distribution. If the distribution is long tailed to the left (has a long tail to the left of the maximum), the function has negative skewness otherwise it is positively skewed.

$$s = \frac{\mu_3}{\mu_2^{3/2}}$$

Where μ_2 and μ_3 are the second and third central moments respectively.

B.1.5 Kurtosis

Kurtosis is the degree of peakedness of a distribution, denoted by

$$k = \frac{\mu_4}{\mu_2^2}$$

The following bounds apply for k

- $k > 3$

  implies that the distribution is highly peaked (leptokurtic),

- $k < 3$

  reflects a flat-topped distribution (platykurtic) and

- $k = 3$

  shows a moderately peaked distribution (mesokurtic)

B.2 Distributions

In this section, we describe the distributions used in the simulation study. We learn from the statistical literature that the least squares estimator performs poorly when the error terms are not normal. Hence, we consider a selection of some of the skewed and symmetric distributions to observe whether or not the distribution of error terms plays a role in performance of estimators: Thiart (1994) observed constant efficiencies of estimators hence, no significant role played by the distributions except for one (slash), of which the variance does not exist.

We select four distributions, namely; Normal, Contaminated normal, Laplace and Exponential, of which variances exist and are defined in the summary table at the end of this chapter. Uniform distribution is used as an input to other distributions.

The corresponding moments are provided in the summary table.

B.2.1 Uniform Distribution

Uniform random numbers within the interval [0,1] are similar to random numbers between 0 and 1.

The uniform density function is denoted by

$$\mathbf{f}(x) = \left\{ \begin{array}{cc} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & x < a \quad or \quad x > b \end{array} \right\}$$

A uniformly distributed variable has mean $\mu = \frac{a+b}{2}$ and variance $\sigma^2 = \frac{(b-a)^2}{12}$. In this study, we use the uniformly distributed random variables to generate error terms from some of the distributions.

B.2.2 Normal Distribution

The Normal distribution (Gaussian) is one of the best known continuous distributions in statistics. The normal density function is defined as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \qquad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad 0 < \sigma \qquad \text{(B.2)}$$

We refer to the two parameters μ and σ as the mean and the standard deviation, used to specify location of the data and the spread of the distribution respectively. The Normal distribution has a bell-shape that flattens when σ increases.

A special case for a Normal distribution in which $\mu = 0$ and $\sigma = 1$ is known as the standard normal distribution, the density function of which is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad -\infty < x < \infty.$$

B.2.3 Contaminated Normal Distribution

The contaminated normal distributed variable is a sum of two weighted, normally distributed variables with independent values of μ and σ. For example; a variable $Y = w_1 Y_1 + w_2 Y_2$ has a Contaminated normal distribution if the following conditions hold.

- $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_2 \sim N(\mu_2, \sigma_2^2)$ and
- the weights sum to 1 $(w_1 + w_2 = 1)$; $\quad w_1, w_2 > 0$

The density function of Y is defined by

$$f(y) = \frac{w_1}{\sigma_1 \sqrt{2\pi}} \; e^{-\frac{(y_1-\mu_1)^2}{2\sigma_1^2}} + \frac{w_2}{\sigma_2 \sqrt{2\pi}} \; e^{-\frac{(y_2-\mu_2)^2}{2\sigma_2^2}} \tag{B.3}$$

The mean and variance of Y are $w_1\mu_1 + w_2\mu_2$ and $w_1\sigma_1^2 + w_2\sigma_2^2$ respectively.

B.2.4 Laplace Distribution

The Laplace distribution is the distribution of the difference between two independent variables with identical Exponential distributions. A variable X has the Laplace distribution if its density function is defined by

$$f(x) = \frac{1}{2c} \; e^{-\frac{|x-a|}{c}}, \qquad -\infty < x < \infty, \quad -\infty < a < \infty, \quad 0 < c \tag{B.4}$$

Where
a = the mean or the location parameter and
c = the scale parameter greater than zero, defined such that the variance of X is $2c^2$.

B.2.5 Exponential Distribution

The Exponential distribution is usually used to model the interval of time between events and the density depends on λ. If events are occurring randomly with an average rate of λ per unit of time, then the length of time is exponentially distributed with the density function denoted by

$$f(x) = \lambda e^{-\lambda x} \qquad x \geq 0, \quad \lambda > 0 \tag{B.5}$$

B.2.6 Moments, Skewness and Kurtosis

In summary, we tabulate the expressions for moments, skewness and kurtosis for the distributions under consideration. We use the following notation:

- μ'_1 = the first raw moment or the mean.
- μ_r = the r^{th} central moment.
- μ_2 = the 2^{nd} central moment or the variance.
- K = Kurtosis
- S = Skewness

	μ'_1	μ_r	μ_2	**K**	**S**
Uniform	$\frac{a+b}{2}$	$\frac{(b-a)^r}{2^r(r+1)}$ r even 0, r odd	$\frac{(b-a)^2}{12}$	$\frac{9}{5}$	0
Normal	μ	$\frac{r!\sigma^r}{(r/2)!2^{r/2}}$	σ^2	3	0
Cont. Normal	$w_1\mu_1 + w_2\mu_2$	$\frac{r!(w_1\sigma_1^r + w_2\sigma_2^r)}{(r/2)!2^{r/2}}$	$w_1\sigma_1^2 + w_2\sigma_2^2$	$\frac{3[w_1\sigma_1^4 + w_2\sigma_2^4]}{w_1\sigma_1^2 + w_2\sigma_2^2}$	0
Laplace	a	$r!c^r$ r even 0, r odd	$2c^2$	6	0
Exponential	$\frac{1}{\lambda}$	$-\frac{1}{\lambda^r} + \frac{r(\mu_{r-1})}{\lambda}$	$\frac{1}{\lambda^2}$	9	2

Table B.1: Moments

B.3 Standardizing the variables

We transform the variables to the standard form in the following manner.

Consider the following regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i$$

Centring

Subtract the mean of each variable from the corresponding variable. Be cautious to add and subtract the same terms to keep the model in its original form. That is

$$\begin{aligned} Y_i = \beta_0 + \beta_1\bar{X}_{i1} + \beta_2\bar{X}_{i2} + \beta_3\bar{X}_{i3} + \beta_4\bar{X}_{i4} + \beta_5\bar{X}_{i5} + \beta_1(X_{i1} - \bar{X}_{i1}) + \beta_2(X_{i2} - \bar{X}_{i2}) + \\ \beta_3(X_{i3} - \bar{X}_{i3}) + \beta_4(X_{i4} - \bar{X}_{i4}) + \beta_5(X_{i5} - \bar{X}_{i5}) + \epsilon_i \end{aligned}$$

Let

$$\grave{\beta}_0 = \bar{Y} = \beta_0 + \beta_1\bar{X}_{i1} + \beta_2\bar{X}_{i2} + \beta_3\bar{X}_{i3} + \beta_4\bar{X}_{i4} + \beta_5\bar{X}_{i5}$$

Then

$$Y_i = \grave{\beta}_0 + \beta_1(X_{i1} - \bar{X}_{i1}) + \beta_2(X_{i2} - \bar{X}_{i2}) + \beta_3(X_{i3} - \bar{X}_{i3}) + \beta_4(X_{i4} - \bar{X}_{i4}) + \beta_5(X_{i5} - \bar{X}_{i5}) + \epsilon_i$$

$$Y_i - \bar{Y} = \beta_1(X_{i1} - \bar{X}_{i1}) + \beta_2(X_{i2} - \bar{X}_{i2}) + \beta_3(X_{i3} - \bar{X}_{i3}) + \beta_4(X_{i4} - \bar{X}_{i4}) + \beta_5(X_{i5} - \bar{X}_{i5}) + \epsilon_i.$$

Scaling

Divide and multiply each factor by the square root of the sum of squares of the corresponding centred variable as indicated below.

$$\begin{aligned} \frac{(Y_i - \bar{Y})\sqrt{\sum_i^n (Y_i - \bar{Y})^2}}{\sqrt{\sum_i^n (Y_i - \bar{Y})^2}} &= \frac{\beta_1(X_{i1} - \bar{X}_{i1})\sqrt{\sum_i^n (X_{i1} - \bar{X}_{i1})^2}}{\sqrt{\sum_i^n (X_{i1} - \bar{X}_{i1})^2}} + \frac{\beta_2(X_{i2} - \bar{X}_{i2})\sqrt{\sum_i^n (X_{i2} - \bar{X}_{i2})^2}}{\sqrt{\sum_i^n (X_{i2} - \bar{X}_{i2})^2}} + \\ &\quad \frac{\beta_3(X_{i3} - \bar{X}_{i3})\sqrt{\sum_i^n (X_{i3} - \bar{X}_{i3})^2}}{\sqrt{\sum_i^n (X_{i3} - \bar{X}_{i3})^2}} + \frac{\beta_4(X_{i4} - \bar{X}_{i4})\sqrt{\sum_i^n (X_{i4} - \bar{X}_{i4})^2}}{\sqrt{\sum_i^n (X_{i4} - \bar{X}_{i4})^2}} + \\ &\quad \frac{\beta_5(X_{i1} - \bar{X}_{i5})\sqrt{\sum_i^n (X_{i5} - \bar{X}_{i5})^2}}{\sqrt{\sum_i^n (X_{i5} - \bar{X}_{i5})^2}} \end{aligned}$$

We can simplify the above expressions as follows;

Let

$$Y^* = \frac{(Y_i - \bar{Y})}{\sqrt{\sum_i^n (Y_i - \bar{Y})^2}}$$

$$X_{ij}^* = \frac{(X_{ij} - \bar{X}_{ij})}{\sqrt{\sum_i^n (X_{ij} - \bar{X}_{ij})^2}}$$

$$s_{xi} = \sqrt{\sum_i^n (X_{ij} - \bar{X}_{ij})^2}$$

$$s_{yi} = \sqrt{\sum_i^n (Y_i - \bar{Y})^2}$$

For $j = 1, \ldots, 5$ and $i = 1, \ldots n$

Let
$\beta_j^* = \frac{\beta_j (s_{xi})}{(s_{yi})}$

Then, the regression model in which Y and X are standardized may be expressed as

$$Y_i^* = \beta_j^* X_{ij}^* + \epsilon^* \qquad j = 1, \ldots, 5 \quad and \quad i = 1 \ldots, n$$

Appendix C

Estimators and estimation methods considered

Name	**Expression**	**Symbol**
Least Squares	$(X'X)^{-1}X'Y$	$\hat{\beta}$
General Shrinkage	$d_{sh}\hat{\beta}$	$\hat{\beta}_{sh}$
Stein	$c\hat{\beta}$	$\hat{\beta}_{s}$
Ridge	$(X'X + kI)^{-1}X'Y$	$\hat{\beta}_{R}$
Generalized Ridge	$(X'X + K)^{-1}X'Y$	$\hat{\beta}_{GR}$
Liu	$(X'X + I)^{-1}(X'Y + d\hat{\beta})$	$\hat{\beta}_{L}$
Generalized Liu	$(X'X + I)^{-1}(X'Y + D\hat{\beta})$	$\hat{\beta}_{GL}$
Van Houwelingen and Le Cessie	$\hat{c}\hat{\beta}$	$\hat{\beta}_{vHlC}$
Principal components	$\sum_{i=1}^{p-m} \frac{v_i c_i}{\lambda_i}$	$\hat{\beta}_{pc}$

Table C.1: Notation

OLS	Ordinary Least Squares
Rhkb	Ridge regression with $k = k_{hkb}$
Rhkb-new	Ridge regression with $k = k_{hkb-new}$
Rhkbm-new	Ridge regression with $k = k_{hkbm-new}$
Rhkbm	Ridge regression with $k = k_{hkbm}$
Rlw-new	Ridge regression with $k = k_{lw-new}$
Rlw	Ridge regression with $k = k_{lw}$
Rlwm	Ridge regression with $k = k_{lwm}$
Rlwm-new	Ridge regression with $k = k_{lwm-new}$
Rkam-new	Ridge regression with $k = k_{am-new}$
Rkgm-new	Ridge regression with $k = k_{gm-new}$
Rkmed-new	Ridge regression with $k = k_{med-new}$
Rkam	Ridge regression with $k = k_{am}$
Rkgm	Ridge regression with $k = k_{gm}$
Rkmed	Ridge regression with $k = k_{med}$
GRhk	Ridge regression with $K = K_{hk}$
GRhk-new	Ridge regression with $K = K_{hk-new}$
GRtc	Ridge regression with $K = K_{tc}$
GRtc-new	Ridge regression with $K = K_{tc-new}$
Liumm	Liu estimation with $d = d_{mm}$
Liucl	Liu estimation with $d = d_{cl}$
Gliu	Generalized Liu estimation
Stein	Stein estimation by James and Stein (1961)
PCdel1	Principal components regression deleting one root
PCdel2	Principal components regression deleting two roots

Table C.2: Estimation methods and abbreviations

Parameter	**Estimation Method**
$k_{hkb} = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$	Rhkb
$k_{hkbm} = \frac{(r-2)\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$	Rkhkbm
$k_{lw} = p\hat{\sigma}^2 / \sum \hat{\alpha}_i^2 \lambda_i$	Rklw
$k_{lwm} = \frac{(r-2)\hat{\sigma}^2 \sum \lambda_i}{r\hat{\beta}' X' X \hat{\beta}}$	Rklwm
$k_{am} = \frac{1}{p} \sum_{i=1}^{p} \left(\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2} \right)$	Rkam
$k_{gm} = \frac{\hat{\sigma}^2}{(\prod_{i=1}^{p} \hat{\alpha}_i^2)^{\frac{1}{p}}}$	Rkgm
$k_{med} = median\left(\frac{\hat{\sigma}^2}{\hat{\alpha}_1^2, \ldots, \hat{\alpha}_p^2} \right)$	Rkmed
$k_{hk_i} = \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}$	GRtc
$k_{tc_i} = \frac{(\lambda_i)}{((\lambda_i \hat{\alpha}_i^2 / \hat{\sigma}^2) + 1)}$	GRhk
$d_{mm} = 1 - \hat{\sigma}^2 \frac{\sum_{i=1}^{p} 1/\lambda_i(\lambda_i+1)}{\sum_{i=1}^{p} \hat{\alpha}_i^2/(\lambda_i+1)^2}$	Liumm
$\hat{d}_{cl} = 1 - \hat{\sigma}^2 \frac{\sum_{i=1}^{p} 1/(\lambda_i+1)}{\sum_{i=1}^{p} \lambda_i \hat{\alpha}_i^2/(\lambda_i+1)^2}$	Liucl
$\hat{d}_i = d_i = 1 - \hat{\sigma}^2 \frac{(\lambda_i+1)}{\lambda_i \hat{\alpha}_i^2}$	GLiu
$c = max\left(0, \left[1 - \frac{(p-2)(n-p)\hat{\sigma}^2}{(n-p+2)\hat{\beta}'\hat{\beta})} \right] \right)$	Stein

Table C.3: Traditional parameters

Parameter	**Estimation Method**
$k_{hkb-new} = \frac{p\hat{\sigma}^2_{pcdel1}}{\hat{\beta}'_{pcdel1}\hat{\beta}_{pcdel1}}$	Rhkb-new
$k_{hkbm-new} = \frac{(r-2)\hat{\sigma}^2_{pcdel1}}{\hat{\beta}'_{pcdel1}\hat{\beta}_{pcdel1}}$	Rkhkbm-new
$k_{lw-new} = \frac{p\hat{\sigma}^2_{pcdel1}}{\sum_{i=1}^{p-1}\hat{\alpha}^2_{pcdel1_i}\lambda_i}$	Rklw-new
$k_{lwm-new} = \frac{(r-2)\hat{\sigma}^2_{pcdel1}\sum_{i=1}^{p-1}\lambda_i}{r\hat{\beta}'_{pcdel1}X'X\hat{\beta}_{pcdel1}}$	Rklwm-new
$k_{am-new} = \frac{1}{(p-m)}\sum_{i=1}^{p-m}\left(\frac{\hat{\sigma}^2_{pcdel1}}{\hat{\alpha}^2_{pcdel1_i}}\right)$	Rkam-new
$k_{gm} = \frac{\hat{\sigma}^2_{pcdel1}}{(\prod_{i=1}^{p-m}\hat{\alpha}^2_{pcdel1_i})^{\frac{1}{p-m}}}$	Rkgm-new
$k_{med} = median\left(\frac{\hat{\sigma}^2_{pcdel1}}{\hat{\alpha}^2_{pcdel1_1},\ldots,\hat{\alpha}^2_{pcdel1_{p-1}}}\right)$	Rkmed-new
$k_{hk-new_i} = \frac{\hat{\sigma}^2_{pcdel1}}{\hat{\alpha}^2_{pcdel1_i}}$	GRhk-new
$k_{tc_i} = \frac{(\lambda_i)}{((\lambda_i\hat{\alpha}^2_{pcdel1_i}/\hat{\sigma}^2_{pcdel1})+1)}$	GRtc-new

Table C.4: New parameters

Appendix D

Past simulation studies

Author	**Measure**	**Methods compared**	**Superior**
Hoerl et al. (1975)	TMSE	- Ridge (k_{hkb}) - OLS	- Ridge
Marquardt and Snee (1975)	Residual prediction error	- Ridge, - OLS, - Generalized inverse (Marquardt (1970), - All possible subsets	- Generalized inverse - Ridge
Guilkey and Murphy (1975)	TMSE	- Generalized ridge, - OLS, - Directed ridge	- Directed ridge
Lawless and Wang (1976)	MSE	-Ridge, - Generalized ridge, - Principal components, - OLS	-Ridge
Hoerl and Kennard (1976)	TMSE	- Ridge, - OLS	-Ridge
Hocking (1976)	MSE	-Ridge, - Best subset selection, - Principal components, - OLS	- Ridge, - Principal components

Gunst and Mason (1977)	TMSE	- Ridge, - Latent root (Hawkins.1973), - Principal components, - OLS	- Ridge, - Principal components
Winchen and Churchill (1978)	TMSE	- OLS, - Ridge (k_{lw}, k_{hk}, k_{hkb}, Mcdonald and Galarneau (1975))	- All ridge except k_{hk}
Thiart et al. (1993)	Relative Efficiencies	- Ridge, - Principal components, - Generalized ridge, - Jackknife (Quenouille 1956; Tukey, 1958), - Fractional principal components (Mayer and Willke, 1973)	- All biased estimators; no outstanding estimator
Thiart (1994)	Relative Efficiencies	- Ridge, - Principal components, - Generalized ridge, - Jackknife (Quenouille 1956; Tukey, 1958), - Fractional principal components (Mayer and Willke, 1973), - L_p norm	-All biased estimators; no outstanding estimator

Breiman (1995)	Prediction error	- nonnegative garrote (Breiman, 1995), - OLS, - Ridge, - Subset selection	-Ridge
Aldrin (1997)	Prediction error	- ridge - Stein, - Partial least squares, - Variable selection, - Length modified ridge (Aldrin, 1997)	-Ridge,
Fu (1998)	MSE	- Bridge estimation (Frank and Friedman, 1993), - Least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), - OLS and - Ridge regression	-Bridge
Kaciranlar and Sakallioglu (2001)	MSE	- r-d class (Kaciranlar and Sakallioglu, 2001), - Liu (Liu, 1993), - OLS, - Principal components,	- r-d class
Wencheko (2001)	Pitman measure of nearness (Pitman, 1937)	- Principal components, - Ridge, - OLS, - Shrinkage	- Ridge
Liu (2003)	MSE	- Principal components, - Ridge, - OLS, - Liu-type (Liu, 2003)	-Liu-type

Table D.1: case studies