LAM, Chun Choon, 1944-
A CLASS OF BIASED ESTIMATORS FOR GENERAL LINEAR MODELS.

The Ohio State University, Ph.D., 1973
Statistics

University Microfilms, A XEROX Company, Ann Arbor, Michigan
A CLASS OF BIASED ESTIMATORS
FOR GENERAL LINEAR MODELS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

by

Chun Choon Lam, B.Sc, M.S.

* * * * * *

The Ohio State University
1973

Reading Committee:
Professor T.A. Willke
Professor J.S. Rustagi
Professor J. Singh

Approved by

Adviser
Division of Statistics
ACKNOWLEDGMENTS

First of all I wish to express my heartiest thanks to Professor T.A. Willke, my adviser and the Chairman of my dissertation research committee, for his help, advice and those valuable conversations I had with him during the past few years. My thanks also go to Professor J.S. Rustagi and Professor P.O. Anderson for the useful discussions and suggestions during the preparation of this dissertation; to the faculty, the staff and everyone in the Division of Statistics, The Ohio State University, who made my stay at Ohio State a pleasant one. Finally, I am grateful to the Fulbright-Hayes program and the Division of Statistics of the Ohio State University for making this work possible.
VITA

January 6, 1944 . . . . . Born, Ipoh, Malaysia

May, 1969 . . . . . . . . . . B.Sc. in Mathematics, Nanyang University, Singapore

1969-1972 . . . . . . . . Teaching Assistant, Division of Statistics, Ohio State University

September, 1971 . . . . . M.S., Ohio State University

1972-1973 . . . . . . . . Research Associate, Department of Computer and Information Sciences, Ohio State University

FIELDS OF STUDY

MAJOR FIELD: Statistics

ANALYSIS: Professor R. Bojanic

DECISION THEORY: Professor P.O. Anderson

INFEREN CE: Professor R.C. Srivastava
TABLE OF CONTENTS

ACKNOWLEDGEMENTS ............................................... ii
VITA ................................................................. iii

Chapter

1. INTRODUCTION AND REVIEW OF LITERATURE

   1. Introduction 1
   2. Review of Literature 2
   3. The Shrunken Estimators 3
   4. The Ridge Estimators 3
   5. The Generalized Inverse Estimators 4
   6. Other Biased Estimators 6
      6-1. A Modification of the Ridge Estimator 6
      6-2. The Linear Transform Estimators 7

2. THE PROPOSED CLASS OF ESTIMATORS

   1. The Proposed Class of Estimators 8
   2. Geometrical Interpretation 9
   3. Properties of the Proposed Class of Estimators 10

3. A SUBCLASS OF THE PROPOSED CLASS OF ESTIMATORS

   1. A Subclass of $C_5$ 15
   2. Properties of $\hat{\beta}[D]$ 17
   3. Minimum Mean Square Error for $C_6$ 19
   4. Minimum Mean Square Error for $C_1$ and $C_4$ 21
   5. Minimum Generalized Variance on the Ellipsoid
      $\xi(\beta) = \xi_o$ for $C_6$ 24
   6. Minimum Generalized Variance on the Ellipsoid
      $\xi(\beta) = \xi_o$ for $C_1$ and $C_4$ 25
TABLE OF CONTENTS (cont.)

4. A DECISION THEORETIC APPROACH
   1. A Bayesian Interpretation of $\hat{\beta}[A'A]$ .......................... 28
   2. Admissibility of $\hat{\beta}[A'A]$ ......................................... 31
   3. Essentially Complete Class of Decision Rules Based on
      $\hat{\beta}[A'A]$ ............................................................. 33

5. COMPARISONS BETWEEN CLASSES OF ESTIMATORS
   1. Geometrical Interpretation of Estimators ................................. 34
   2. Optimality Properties of Estimators .................................... 35
   3. Decision Theoretic Interpretation of Estimators ....................... 38
   4. Interpretation of Estimators as Least Squares Estimators ............ 38

6. AN EXAMPLE ................................................................................. 42

7. CONCLUSION ................................................................................. 51

APPENDIX ......................................................................................... 54
REFERENCES .................................................................................... 56
CHAPTER 1

Introduction and Review of Literature

1. Introduction

The theory of general linear model is one of the most widely used statistical theories in the analysis of data. It is well known that the least squares estimators of the general linear model are minimum variance unbiased estimators under the usual assumption of normal theory and minimum variance linear unbiased without normality assumption \([3]\). However, it is also well known that there are situations such that the variance of the minimum variance unbiased (linear) estimator is too large for any practical purposes. For example, when some of the input variables are highly correlated the variances of the minimum variance unbiased (linear) estimator of the unknown coefficients in the general linear model will be too large for estimation and prediction purposes \([6]\) \([8]\). This problem has recently been studied by a number of people \([5]\) \([7]\) \([8]\) \([10]\) \([14]\). Their idea is to sacrifice the notion of unbiasedness; a small amount of bias is introduced into the estimator to obtain a substantial reduction in the variance of the estimator of the unknown coefficients, and hence improves the mean square error of the estimator. The present research is along this line. The proposed class of estimators not only is a generalization of some of the existing biased estimators, it also has a much richer statistical structure than any other biased estimators discussed in the reference above.
2. Review of Literature

Consider the general linear model

\[ Y = X\beta + \epsilon \]

where \( Y \) is an \( n \times 1 \) random vector of observed values, \( X \) is an \( n \times p \) known input matrix, \( \beta \) is a \( p \times 1 \) vector of unknown coefficients, and the \( n \times 1 \) error vector is assumed to be distributed with mean \( E\epsilon = 0 \) and variance-covariance matrix \( E\epsilon\epsilon' = \sigma^2 I \). The primary aim is to estimate \( \beta \) when the matrix \( X'X \) is ill-conditioned. The method proposed allows estimation when \( X'X \) is singular as well.

For any estimator \( B \), let

\[ \hat{\phi}(B) = (Y - XB)'(Y - XB) \]
\[ \text{Var}(B) = E(B - E(B))(B - E(B))' \]
\[ V(B) = E(B - E(B))'(B - E(B)) \]
\[ D(B) = (E(B) - \beta)'(E(B) - \beta) \text{ and} \]
\[ G(B) = E(B - \beta)'(B - \beta) \]

be the sum of squares of residuals, the variance-covariance matrix, the generalized variance, the squared bias, and the mean square error of \( B \) respectively.

It is easy to verify that

\[ V(B) = \text{tr} \text{Var}(B), \quad G(B) = V(B) + D(B), \]

and for the least squares estimator \( \hat{\beta} \) of \( \beta \), if \( X'X \) is nonsingular,

\[ \hat{\beta} = (X'X)^{-1}X'Y \]
\[ \hat{\phi}(\hat{\beta}) = \min_\beta \hat{\phi}(\beta) = \hat{\phi}_{\text{min}} \]
\[ \text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1} \quad \text{and} \]
\[ V(\hat{\beta}) = G(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \quad \text{where} \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0 \]

are eigenvalues of \( X'X \).
Since \( V(\hat{\beta}) \) is bounded below by \( \sigma^2/\lambda_p \), it will be large if some of the eigenvalues of \( X'X \) are close to 0. Hence, if the input data are such that the resulting \( X'X \) matrix is ill-conditioned the least squares estimator \( \hat{\beta} \) may be too unreliable for prediction and estimation, even though it is the minimum variance unbiased (linear) estimator. To remedy this undesirable situation, several biased procedures have been proposed and studied. We shall discuss briefly these procedures in the following sections.

Before we proceed, it should be remembered that we are not only dealing with the case when \( X'X \) is ill-conditioned, we are also interested in the case when \( X'X \) is singular. Because of this reason the following rule is adopted in this research. Whenever we write \( A^{-1} \) we assume that the matrix \( A \) is nonsingular, otherwise no assumption on the singularity of \( A \) is made unless it is stated in the context.

3. The Shrunken Estimators

In a series of papers James and Stein [7] [14] [15] suggested the class of shrunken estimators

\[
C_\perp = \{ \hat{\beta}(c) | \hat{\beta}(c) = c\hat{\beta}, \ 0 \leq c \leq 1 \},
\]

where \( \hat{\beta} = (X'X)^{-1}X'Y \) is the least squares estimator of \( \beta \). They studied the mean square error of the estimator \( \hat{\beta}(c) \) and showed how to construct a \( c_0 \) such that the mean square error of \( \hat{\beta}(c_0) \) is less than that of \( \hat{\beta} \).

4. The Ridge Estimators

Instead of shortening the length of the vector \( \hat{\beta} \), as James and Stein did, Hoerl and Kennard [5] advanced the idea of augmenting the
X'X matrix by adding a small positive quantity to the diagonal of X'X. Thus, they proposed the estimator \( \hat{\beta}[kI] = (X'X + kI)^{-1}X'Y \) with \( k \geq 0 \). Denote this class of estimators by \( C_2 = \{ \hat{\beta}[kI] \mid \hat{\beta}[kI] = (X'X + kI)^{-1}X'Y, k \geq 0 \} \).

This is the class of so-called ridge estimators. Geometrically, for given \( k \geq 0 \), \( \hat{\beta}[kI] \) is the point on the \( p \)-dimensional ellipsoid having \( \hat{\beta} = \text{constant} \) that minimizes \( \beta'\beta \). Although they fail to construct the optimal \( k^* \) that minimizes \( G(\hat{\beta}[kI]) \), they are able to prove the existence of a \( k^*_0 > 0 \) such that \( G(\hat{\beta}[k^*_0 I]) < G(\hat{\beta}) \). In [6] they illustrate how one might choose a \( k \) in order to arrive at a stabilized system by making use of the concept of ridge trace.

Note that the addition of a small positive quantity to the diagonal of X'X not only introduces bias into the estimator, it also has the effect of removing the severe nonorthogonality of the system under consideration. This is the main idea behind this research. We will demonstrate this point in later chapters.

5. The Generalized Inverse Estimators

The recent development in generalized inverse of matrices and its applications in statistics [4] [12] suggests a legitimate procedure in estimating \( \beta \) when X'X is ill-conditioned or singular. In [8], Marquardt utilized the generalized inverse of X'X to propose a biased procedure to estimate \( \beta \). His procedure is to assign a rank \( r, 0 < r \leq \text{rank} \ X'X \), to the matrix X'X and study the properties of the resulting estimator. The procedure is as follows:

Let \( P \) be an orthogonal \( p \times p \) matrix that diagonalizes X'X;
that is,

\[ P' X'X P = (\lambda_i \delta_{ij}) = D_{\lambda}, \text{ say,} \]

where \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \) are eigenvalues of \( X'X \) and \( \delta_{ij} = 1 \) or 0 according to \( i = j \) or \( i \neq j \) respectively. Assume that the rank of \( X'X \) is \( \ell \) and \( \ell < p \), then \( \lambda_{\ell+1} = \lambda_{\ell+2} = \cdots = \lambda_p = 0 \). Partition \( P \) and \( D_{\lambda} \) as follows.

\[ P = (P_{\ell} : P_{p-\ell}) \quad \text{and} \quad D_{\lambda} = \begin{pmatrix} D_{\ell} & 0 \\ 0 & D_{p-\ell} \end{pmatrix}, \]

where \( P_{\ell} \) is \( p \times \ell \), \( P_{p-\ell} \) is \( p \times (p-\ell) \) and \( D_{k} \) is \( k \times k \) for \( k = \ell, p-\ell \).

Denote the generalized inverse of \( X'X \) by \( (X'X)_{\ell}^{-1} \), then it is easy to show that

\[ (X'X)_{\ell}^{-1} = P_{\ell} D_{\ell}^{-1} P_{\ell}' = \sum_{i=1}^{\ell} \frac{1}{\lambda_i} P_i P_i' \]

where \( P_i \) is the eigenvector of \( X'X \) corresponding to \( \lambda_i \). The generalized inverse estimator is then defined to be

\[ \hat{\beta}(\ell) = (X'X)_{\ell}^{-1} X'X \hat{\beta}. \]

Now suppose that a rank \( r, 0 < r \leq \text{rank} \ (X'X) \), is assigned to the matrix \( X'X \) and estimate \( \beta \) by

\[ \hat{\beta}(r) = (X'X)_{r}^{-1} X'X \hat{\beta} \]

where

\[ (2) \ (X'X)_{r}^{-1} = \sum_{i=1}^{r} \frac{1}{\lambda_i} P_i P_i' \]

Since \( r \) can take any value between 0 and rank \( (X'X) \) the summation in (2) is taken to include all terms less than or equal to the integral part of \( r \), plus that fraction of the next term in (2) by which \( r \) exceeds its integral part. For the sake of convenience, denote this
class of estimators by
\[ C_3 = \{ \hat{\beta}^{-}(r) | \hat{\beta}^{-}(r) = (X'X)^{-} X'X \hat{\beta}, 0 < r \leq \text{rank} (X'X) \}. \]

In his paper the author shows that if \[ \sum_{i=r+1}^{p} \frac{1}{\lambda_i} > \frac{1}{\sigma^2} \beta'\beta \] then \[ G(\hat{\beta}^{-}(r)) < G(\hat{\beta}). \] Note that the same remark applies to the summation here as that given in (2). If \( X'X \) is singular the above inequality holds true in the limiting case.

6. Other Biased Estimators

Recently, Mayer and Willke [10] proposed two different classes of estimators. One of them is a modification of the ridge estimator and the other is a general class of estimators, which includes \( C_1, C_2 \) and \( C_3 \) as subclasses, and the proposed class of estimators in this research as well.

6-1. A Modification of the Ridge Estimator

Suppose the input variables can be partitioned into two sets such that the inter-dependence between these two sets is very large and the dependence within the set is small. We express this in terms of \( X'X \) as follows:

\[
(3) \quad X'X = \begin{pmatrix}
X_{11} & X_{12} \\
X_{21} & X_{22}
\end{pmatrix}
\]

where \( X_{11} \) is \( p_1 \times p_1 \) and \( X_{22} \) is \( (p-p_1) \times (p-p_1) \). Note that \( |X'X| \) is small because of the large entries in \( X_{12} \). Also note the assumption that \( X'X \) is nonsingular in this section.

For the ridge estimator a small positive quantity is added to the diagonal of \( X'X \); that is

\[ \hat{\beta}[kI] = (X'X + kI)^{-1} X'X \hat{\beta}. \]

However, for the estimator discuss in this section the elements of the
submatrices $X_{12}$ and $X_{21}$ in (3) are decreased instead. The estimator is

$$\hat{\beta}(\alpha X_{1j}) = \left( \begin{array}{cc} X_{11} & \alpha X_{12} \\ \alpha X_{21} & X_{22} \end{array} \right)^{-1} \ X' \hat{\beta}$$

where $0 \leq \alpha \leq 1$.

The estimator is a modification of the ridge estimator. For example, it can be easily seen that when $X'X$ is a 2 x 2 matrix, they are related. Write $X'X$ in correlation form

$$X'X = \left( \begin{array}{cc} -1 & \rho \\ \rho & 1 \end{array} \right)$$

and for $0 < \alpha < 1$, (and usually close to 1)

$$\hat{\beta}(\alpha \rho) = \left( \begin{array}{cc} 1 & \alpha \rho \\ \alpha \rho & 1 \end{array} \right)^{-1} \ X' \hat{\beta} = \alpha(X'X + \frac{1-\alpha}{\alpha} \ I)^{-1} X' \hat{\beta} = \alpha \beta \left( \frac{1-\alpha}{\alpha} \ I \right)$$.

6-2. The Linear Transform Estimators

Let $\hat{\beta} = (X'X)^{-1} X'Y$ be the least squares estimator of $\beta$, in [10] Mayer and Willke studied the following class of linear transform estimators, in full detail.

$$C_4 = \{ \hat{\beta}(A) | \hat{\beta}(A) = A \hat{\beta}, A \text{ is a } p \times p \text{ matrix} \}.$$  

They show that the estimator that minimizes $\hat{\beta}'A' \hat{\beta}$ on the ellipsoid $\varphi(A\hat{\beta}) = \text{constant}$ is a ridge estimator. They further show that $\hat{\beta}(A_0)$ minimizes $G(A\hat{\beta})$ over the class $C_4$, where $A_0 = \beta \beta'(\sigma^2(X'X)^{-1} + \beta \beta')^{-1}$. Another contribution of their work is that they have found the member in $C_4$ that minimizes $V(A\hat{\beta})$ on the ellipsoid $\varphi(A\hat{\beta}) = \text{constant}$. 
CHAPTER 2

The Proposed Class of Estimators

We study and investigate the properties of the proposed class of estimators in this chapter, and give a similar study for a subclass of the proposed class of estimators in the next chapter. A decision theoretic interpretation of the proposed class of estimators is given in Chapter 4. Comparisons between various classes of estimators discussed in this research are made in Chapter 5 and in Chapter 6 an example is given to show their performance.

1. The Proposed Class of Estimators

In the general linear model (1) let \( \hat{\beta} \) be a solution of the normal equation

\[
\begin{align*}
\hat{\beta} &= \hat{\beta} [A' A] \hat{\beta} [A' A] = (X'X + A'A)^{-1}X'Y, A is a \, r \times p \, matrix, r \geq 1 
\end{align*}
\]

It is shown in [13] that \( \hat{\beta} \) is a least squares estimator of \( \beta \) regardless whether \( X'X \) is singular or not. If \( X'X \) is nonsingular a unique \( \hat{\beta} \) is obtained, otherwise \( \hat{\beta} \) is not unique.

Consider the following class of estimators

\[
\hat{\beta} = [\beta [A' A] \hat{\beta} [A' A] = (X'X + A'A)^{-1}X'X\hat{\beta}, A is a \, r \times p \, matrix, r \geq 1 \}
\]

where \( \hat{\beta} \) is a solution of (4).

The class \( \hat{\beta} \) will be called the class of general ridge estimators.

Hoerl and Kennard [5] proposed a "generalized ridge" estimator at the end of their paper which was a minor generalization of the ridge
estimator. Instead of adding the same number, \( k \), to all diagonal elements of \( X'X \), their "generalized ridge" allowed possibly different values to be added to different diagonal elements. The proposed class, \( C_3 \), carries the ridge idea to complete generalization by adding \( A'A \) to \( X'X \). Hence, it contains not only the ridge estimators, but also Hoerl and Kennard's "generalized ridge" as special case. Thus we justify calling it the class of general ridge estimators.

The fact that \( C_5 \) contains \( C_1 \) and \( C_2 \) as subclass can be easily seen by setting \( A'A = \alpha X'X \) (\( \alpha \geq 0 \)) and \( A'A = kI \) (\( k \geq 0 \)) respectively. Although it is clear that \( C_5 \) is a subclass of \( C_4 \), we feel that it is worthwhile to pursue the present research for the following reasons. First of all, \( C_5 \) has many properties that are not shared by all members of \( C_4 \): for example, for every \( A \) such that \( A'A \) is non-singular, then \( \hat{\beta}[A'A] \) is a Bayes estimator and is admissible under certain conditions. Secondly, results obtained in this research are simpler and easier to apply than those obtained in \( C_4 \).

2. Geometrical Interpretation

Geometrically, \( \hat{\beta}[A'A] \) can be thought of as the \( \beta \) in the \( p \)-dimensional space that minimizes \( \delta(\beta) = \beta'A'\beta \) subject to a constant loss, \( \phi(\beta) = \) constant, or equivalently, it is that \( \beta \) which minimizes the loss, \( \phi(\beta) \), on the ellipsoid \( \beta'A'\beta = \) constant (See Figures 1 and 2). This is shown in

\[
\hat{\beta} \left[ \frac{1}{Y} A' A \right] = (X'X + \frac{1}{Y} A' A)^{-1} X'X\hat{\beta}
\]

minimizes \( \delta(\beta) = \beta'A'\beta \) on the ellipsoid
\[ \hat{s}(\beta) = (Y - X\beta)'(Y - X\beta) = \hat{s}_0, \text{ a constant,} \]

where \( \gamma \) is chosen to satisfy \( \hat{s}(\beta[A'\hat{A}]) = \hat{s}_0 \).

Proof: The theorem is proved by the Lagrangian method of multipliers. Let

\[ F(\beta) = \beta'A'\beta + \gamma(\hat{s}(\beta) - \hat{s}_0) \]

where \( \gamma \) is the Lagrangian constant. Rewrite \( F(\beta) \) as

\[ F(\beta) = \beta'A'\beta + \gamma(Y'Y - 2\beta'X'Y + \beta'X'X\beta - \hat{s}_0) . \]

Differentiate \( F \) with respect to \( \beta \), we get

\[ \frac{dF(\beta)}{d\beta} = 2\beta'A'\beta + \gamma(-2X'Y + 2X'X\beta) . \]

Equating \( \frac{dF(\beta)}{d\beta} \) to 0 and solving for \( \beta \), gives

\[ \hat{\beta}[\frac{1}{\gamma} A'\hat{A}] = (X'X + \frac{1}{\gamma} A'\hat{A})^{-1} X'X\hat{s}_0 , \]

as desired.

Note that for given matrix \( A \), \( \hat{\beta}[\frac{1}{\gamma} A'\hat{A}] \) depends on \( \gamma \). To determine \( \gamma \), we regard \( \hat{s}(\beta[\frac{1}{\gamma} A'\hat{A}] \) as a function of \( \gamma \) and solve the equation \( \hat{s}(\beta[\frac{1}{\gamma} A'\hat{A}]) = \hat{s}_0 \).

The following remarks concern the singularity of the matrix \( A'\hat{A} \).

Remark 1. If \( A \) is such that \( A'\hat{A} \) is nonsingular then \( \hat{\beta}[\frac{1}{\gamma} A'\hat{A}] \) is unique. In this case, given the fixed ellipsoid \( \hat{s}(\beta) = \hat{s}_0 \) we think of increasing \( \hat{s}_0 = \delta(\beta) \) until the two ellipsoids touch as in Figure 1 to determine \( \hat{\beta}[\frac{1}{\gamma} A'\hat{A}] \).

Remark 2. If \( A \) is such that \( A'\hat{A} \) is singular then \( \beta'A'\beta = \text{constant} \) is a degenerated ellipsoid, and \( \hat{\beta}[\frac{1}{\gamma} A'\hat{A}] \) might not be uniquely determined. Various possible cases about this situation are illustrated in Figure 2.

3. Properties of the Proposed Class of Estimators

Some basic properties of the estimator \( \hat{\beta}[A'\hat{A}] \) are given below.
Figure 1 - Geometrical Interpretation of an Optimal Estimator in $C_5$ when $A'A$ is non-singular

Figure 2a.
Figure 2b. \[ \beta \]

\[ \hat{\beta}_{\frac{1}{2}A'A} \]

\[ \phi(\beta) = \phi_0 \]

\[ \delta(\beta) = 0 \]

Figure 2c. \[ \beta \]

\[ \hat{\beta}_{\frac{1}{2}A'A} \]

\[ \delta(\beta) = \delta_0 \]

\[ \delta(\beta) = \delta_0 \]

Figure 2 - Geometrical Interpretation of an Optimal Estimator in $C_J$ when $A' A$ is singular
For any $r \times p$ matrix $A$

(i) $E(\hat{\beta}[A'A]) = (X'X + A'A)^{-1} X'X\beta$,

(5) (ii) $\text{Var}(\hat{\beta}[A'A]) = (X'X + A'A)^{-1} X'X(X'X + A'A)^{-1} \sigma^2$

\leq \text{Var}(\hat{\beta}), \text{if} \ X'X \text{ is nonsingular.}

where $C \leq D$ means that $D - C$ is positive semi-definite; that is, $x'(D - C)x \geq 0$ for all $x$, and $x'(D - C)x = 0$ for some $x \neq 0$.

(6) (iii) $\hat{\phi}(\hat{\beta}[A'A]) = \hat{\phi}(\hat{\beta}) + \hat{\beta}((X'X + A'A)^{-1} X'X - I)'X'(X'X + A'A)^{-1} X'X - I)\hat{\beta}$.

Proof: Since $\hat{\beta}$ satisfies the normal equation $X'X\hat{\beta} = X'Y$, and $EY = X\beta$ and $\text{Var} Y = \sigma^2 I$.

(i) $E(\hat{\beta}[A'A]) = E((X'X + A'A)^{-1} X'Y) = (X'X + A'A)^{-1} X'X\beta$

(ii) $\text{Var}(\hat{\beta}[A'A]) = (X'X + A'A)^{-1} X'X(\text{Var}(Y))(X'X + A'A)^{-1}$

\[= (X'X + A'A)^{-1} X'X(X'X + A'A)^{-1} \sigma^2\]

To prove that $\text{Var}(\hat{\beta}[A'A]) \leq \text{Var}(\hat{\beta})$ we see that if $X'X$ is nonsingular,

$(X'X + A'A)(X'X)^{-1}(X'X + A'A) = X'X + 2A'A + A'A(X'X)^{-1}A'A \geq X'X,$

because both of the terms $2A'A$ and $A'A(X'X)^{-1}A'A$ are positive semidefinite or positive definite.

Hence, $\text{Var}(\hat{\beta}[A'A]) \leq \sigma^2(X'X)^{-1} = \text{Var}(\hat{\beta}).$

(iii) $\hat{\phi}(\hat{\beta}[A'A]) = (Y - X\hat{\beta}[A'A])'(Y - X\hat{\beta}[A'A])$

\[= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (\hat{\beta}[A'A] - \hat{\beta})'X'X(\hat{\beta}[A'A] - \hat{\beta})

\[= \hat{\phi}(\hat{\beta}) + \hat{\beta}((X'X + A'A)^{-1} X'X - I)'X'(X'X + A'A)^{-1} X'X - I)\hat{\beta}.

From (i) it is clear that $\hat{\beta}[A'A]$ is a biased estimator unless $A'A = 0$.

As we pointed out in Chapter 1 if $X'X$ is ill-conditioned, the least squares estimator $\hat{\beta}$ of $\beta$ may be unreliable for any practical purposes. Under such circumstances one naturally seeks to look for an estimator which is as close to the true value, $\beta$, as possible. To achieve this a price has to be paid, in our case it is unbiasedness.
We give up the notion of unbiasedness in order to achieve our goal. For this end, let

\[ \ell^2(\hat{\beta}[A'A]) = (\hat{\beta}[A'A] - \beta)'(\hat{\beta}[A'A] - \beta) \]

denote the squared distance from \(\hat{\beta}[A'A]\) to \(\beta\), and consider the mean square error of \(\hat{\beta}[A'A]\). We now show that

(iv) \(G(\hat{\beta}[A'A]) = V(\hat{\beta}[A'A]) + D(\hat{\beta}[A'A])\)

\[ = \sigma^2 \text{tr}(X'X + A'A)^{-1} X'X(X'X + A'A)^{-1} \]

\[ + \beta'((X'X + A'A)^{-1}X'X - I)'((X'X + A'A)^{-1}X'X - I)\beta \]

since

(7) \(V(\hat{\beta}[A'A]) = \sigma^2 \text{tr} (X'X + A'A)^{-1} X'X(X'X + A'A)^{-1}\) and

(8) \(D(\hat{\beta}[A'A]) = \beta'((X'X + A'A)^{-1}X'X - I)'((X'X + A'A)^{-1}X'X - I)\beta\).

Proof: (7) Since

\[ V(\hat{\beta}[A'A]) = \text{tr Var} (\hat{\beta}[A'A]) \]

\[ = \text{tr} [(X'X + A'A)^{-1}X'X(X'X + A'A)^{-1} \sigma^2] \quad \text{(From (5))} \]

\[ = \sigma^2 \text{tr} (X'X + A'A)^{-1} X'X(X'X + A'A)^{-1}. \]

(8) By definition

\[ D(\hat{\beta}[A'A]) = (E(\hat{\beta}[A'A]) - \beta)'(E(\hat{\beta}[A'A]) - \beta) \]

\[ = ((X'X + A'A)^{-1}X'X\beta - \beta)'((X'X + A'A)^{-1}X'X\beta - \beta) \]

\[ = \beta'((X'X + A'A)^{-1}X'X - I)'((X'X + A'A)^{-1}X'X - I)\beta. \]

In \(C_i, \ i = 1, 2, 3, 4\), it has been shown that there always exists a member in the class which has mean square error less than that of \(\hat{\beta}\). We are going to see, in the next chapter, how a member in the proposed class estimators can be constructed to have mean square error less than the mean square error of the least squares estimator \(\hat{\beta}\).
CHAPTER 3
A Subclass of the Proposed Class of Estimators

In statistics one assumption frequently made is the commutativity of two square matrices. Geometrically, this assumption says that the two matrices have the same eigenvectors. Thus, if A and B are matrices associated with the ellipsoids $x'Ax = \text{constant}$ and $x'Bx = \text{constant}$ then $AB = BA$ implies that these two ellipsoids have the same (parallel) axes, and hence we could find an orthogonal matrix to diagonalize both matrices A and B simultaneously. In this chapter we are going to study $C_5^*$ under this assumption.

1. A Subclass of $C_5^*$

Consider the following subclass of $C_5^*$

$$C_6^* = \{ \hat{\beta}[A'A][\hat{\beta}[A'A] = (X'X + A'AX)'X^2X^2, A \text{ is } r \times p \text{ and } A'AX'X = X'XA'AX \}.$$ 

From now on we will call this class of estimators the class of orthogonal ridge estimators.

It should be noted that we are really dealing with equivalence class of matrices. Two matrices $A_1$ and $A_2$ are said to be equivalent if

$$P'A_1'A_1P = P'A_2'A_2P = (\delta_{ij} d_i) = D$$

where $P$ is an $p \times p$ orthogonal matrix and $d_i \geq 0, i = 1, 2, \ldots, p$ are the common eigenvalues of $A_1'A_1$ and $A_2'A_2$. Indeed, if $A_1$ and
\( A_2 \) belong to the same equivalence class by the above definition, then both \( A_1 \) and \( A_2 \) will yield the same estimator in \( C_0 \), hence for our purpose they are equivalent. As a matter of fact, it is true that (9) holds if and only if there exists an orthogonal matrix \( Q \) such that \( A_1 = QA_2 \) and hence \( A_1 \) and \( A_2 \) are equivalent in the usual sense of the term. Since the equivalence class is determined by the diagonal matrix \( D \), we henceforth index the estimator in \( C_0 \) by \( D \) and denote the estimator by \( \hat{\beta}[D] \).

\[
\hat{\beta}[D] = (X'X + A'A)^{-1} X'X\beta
\]

where \( D = (d_{ij} \sqrt{d_i}) \) and \( d_i \geq 0, i = 1,2, \ldots, p \) are eigenvalues of \( A'A \).

The lemma given below, though simple, gives us a procedure to construct \( A \) from a given set \( \{d_1, d_2, \ldots, d_p\} \) of nonnegative numbers. **Lemma 1.** Let \( \{d_1, d_2, \ldots, d_p\} \) be a set of nonnegative real numbers, let \( P \) be any \( p \times p \) orthogonal matrix, then the \( r \times p \) (\( r > 1 \)) matrix \( A \) defined below has the property that \( A'A \) has eigenvalues \( d_1, d_2, \ldots, d_p \).

\[
A = (\delta_{ij} \sqrt{d_i})P' \quad \text{where} \quad (\delta_{ij} \sqrt{d_i}) \text{ is } p \times p, \text{ if } r = p
\]

\[
= \left(\begin{array}{c}
(\delta_{ij} \sqrt{d_i}) \\
0
\end{array}\right)P' \quad \text{where} \quad (\delta_{ij} \sqrt{d_i}) \text{ is } p \times p \text{ and } 0 \text{ is}
\]

\[
(r - p) \times p, \text{ if } r > p,
\]

\[
= (\delta_{ij} \sqrt{d_i}, 0)P' \quad \text{where} \quad (\delta_{ij} \sqrt{d_i}) \text{ is } r \times r \text{ and } 0 \text{ is}
\]

\[
r \times (p - r), \text{ if } r < p. \text{ Here we assume that } d_{r+1} = d_{r+2} = \cdots = d_p = 0.
\]

The importance of this lemma is that if we have a set of optimal \( d_i \)'s, we can construct an optimal \( A \) just by using a \( P \) that diagonalizes \( X'X \).
The following lemma is also of interest.

**Lemma 2.** Let $P$ be an orthogonal matrix that diagonalizes $X'X$ and let $\{d_1, d_2, \ldots, d_p\}$ be a set of nonnegative real numbers. If $A$ is defined as in Lemma 1 above, then the eigenvectors of $A'A$ are identical to those of $X'X$.

Lemma 2 tells us that any ellipsoids associated with the matrices $X'X$ and $A'A$ will have the same (parallel) axes. The length of the axes is inversely proportional to the squared root of their respective eigenvalues.

2. Properties of $\hat{\beta}[D]$

In this section we express some basic properties of $\hat{\beta}[D]$ in terms of the $d_i$'s.

For any given $r \times p$ matrix $A$, let $P$ be an orthogonal matrix that diagonalizes both $X'X$ and $A'A$ simultaneously, then

\begin{equation}
\text{Var} (\hat{\beta}[D]) = \sigma^2 P (d_i C_i) P' \quad \text{where} \quad C_i = \frac{\lambda_i}{(\lambda_i + d_i)^2}, \quad i = 1, 2, \ldots, p.
\end{equation}

\begin{equation}
\text{Cov} (\hat{\beta}[D]) = \frac{2}{1} \frac{\alpha^2 d_i^2}{(\lambda_i + d_i)^2} \quad \text{where} \quad \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)' = P'\beta
\end{equation}

\begin{equation}
\text{Var} (\hat{\beta}[D]) = \sigma^2 \frac{\lambda_i}{1 (\lambda_i + d_i)^2}
\end{equation}

\begin{equation}
\text{Cov} (\hat{\beta}[D]) = \sigma^2 \frac{\alpha_i d_i}{\lambda_i (\lambda_i + d_i)} \quad \text{where} \quad \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)' = P'\beta
\end{equation}

\begin{equation}
\text{Var} (\hat{\beta}[D]) = \sigma^2 \frac{\lambda_i}{1 (\lambda_i + d_i)^2} + \frac{\lambda_i}{1} \left( \frac{\alpha_i d_i}{\lambda_i + d_i} \right)^2
\end{equation}

In (i) - (v) if $\lambda_i = d_i = 0$, we set the quantity involving both $\lambda_i$
and $d_i$ equal to zero.

Proof:  

(i) From (5)

$$\text{Var} \left( \hat{\beta}[D] \right) = \sigma^2 (X'X + A'A)^{-1} X'X (X'X + A'A)^{-1} \left( \sigma^2 PP' (X'X + A'A)^{-1} PP'X'XPP' (X'X + A'A)^{-1} PP' \right)$$

$$= \sigma^2 P (\delta_{ij} \lambda_i) + (\delta_{ij} d_i)^{-1} (\delta_{ij} \lambda_i) (\delta_{ij} \lambda_i + (\delta_{ij} d_i))^{-1} P' $$

$$= \sigma^2 P (\delta_{ij} c_i) P' .$$

(ii) From (6)

$$\hat{\sigma} (\hat{\beta}[D]) = \hat{\sigma} (\hat{\beta}) + \beta' ((X'X + A'A)^{-1} X'X - I)' X' ((X'X + A'A)^{-1} X'X - I) \beta$$

$$= \hat{\sigma} (\hat{\beta}) + \beta' PP' ((X'X + A'A)^{-1} X'X - I)' PP'X'XPP' ((X'X + A'A)^{-1} X'X - I) PP' \beta$$

$$= \hat{\sigma} (\hat{\beta}) + \beta' P (P' (X'X + A'A)^{-1} PP'X'XP - I)' P'X'XP (P' (X'X +$$

$$+ A'A)^{-1} PP'X'XP - I) P' \beta$$

$$= \hat{\sigma} (\hat{\beta}) + \beta' P ((\delta_{ij} \lambda_i) + (\delta_{ij} d_i))^{-1} (\delta_{ij} \lambda_i) - I)' (\delta_{ij} \lambda_i) (\delta_{ij} \lambda_i)$$

$$+ (\delta_{ij} d_i))^{-1} (\delta_{ij} \lambda_i) - I) P' \beta$$

$$= \hat{\sigma} (\hat{\beta}) + (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_p) (\delta_{ij} \lambda_i d_i^2 / (\lambda_i + d_i)^2) (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_p)'$$

$$= \hat{\sigma} (\hat{\beta}) + \frac{p}{\Sigma} \frac{\alpha^2_i \lambda_i d_i^2}{(\lambda_i + d_i)^2} .$$

(iii) $\text{V} (\hat{\beta}[D]) = \sigma^2 \text{tr} \text{Var} (\hat{\beta}[D]) = \sigma^2 \frac{p}{\Sigma} \frac{\lambda_i}{(\lambda_i + d_i)^2} . $

(iv) From (8),

$$D (\hat{\beta}[D]) = \beta' ((X'X + A'A)^{-1} X'X - I)' ((X'X + A'A)^{-1} X'X - I) \beta$$

$$= \beta' PP' ((X'X + A'A)^{-1} X'X - I)' PP' ((X'X + A'A)^{-1} X'X - I) PP' \beta$$

$$= \beta' P (P' (X'X + A'A)^{-1} PP'X'XP - I)' (P' (X'X + A'A)^{-1} PP'X'XP -$$

$$- I) P' \beta$$

$$= \beta' P ((\delta_{ij} \lambda_i) + (\delta_{ij} d_i))^{-1} (\delta_{ij} \lambda_i) - I)' ((\delta_{ij} \lambda_i) +$$

$$+ (\delta_{ij} d_i))^{-1} (\delta_{ij} \lambda_i) - I) P' \beta$$
\((\alpha_1, \alpha_2, \ldots, \alpha_p)(\delta_{ij}(d_i/(\lambda_1 + d_i))^2)(\alpha_1, \alpha_2, \ldots, \alpha_p)\)  
\[= \frac{p}{\Sigma} \left( \frac{\alpha_i d_i}{\lambda_1 + d_i} \right)^2 \]

(v) This follows immediately from the fact that \(G(\hat{\beta}[D]) = V(\hat{\beta}[D]) + D(\hat{\beta}[D]).\)

**Remark 3.** For any given set \(\{d_1, d_2, \ldots, d_p\}\) of nonnegative numbers the squared bias has the property \(D(\hat{\beta}[D]) \leq \beta'\beta,\) since

\[D(\hat{\beta}[D]) = \frac{p}{\Sigma} \frac{\alpha_i^2}{(1 + \frac{\lambda_1}{d_i})^2} \leq \Sigma \frac{\alpha_i^2}{\Sigma \alpha_i^2} = \beta_1^2 = \beta_1^2, \text{ and } \alpha_1 = \beta_1^2.\]

**Remark 4.** \(D(\hat{\beta}[D]) \to \beta'\beta\) as \(d_i \to \infty, \) \(i = 1, 2, \ldots, p.\) This is so because \(\hat{\beta}[D]\) approaches 0 as all \(d_i \to \infty,\) thus in this sense the limit of the bias is just the length of \(\beta.\)

3. Minimum Mean Square Error For \(C_6\)

The minimum mean square error for \(C_6\) will be established in this section. Similar results which are available in other classes of estimators discussed in this research will be presented in the following section for the purpose of comparisons in later chapters.

**Theorem 3.1.** Let the eigenvalues of the matrix \(X'X\) be such that
\[\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{\ell} > 0 \quad (1 \leq \ell \leq p) \quad \text{and} \quad \lambda_{\ell+1} = \lambda_{\ell+2} = \cdots = \lambda_p = 0,\]
let \(d_{oi} = \frac{\alpha_i^2}{\alpha_i}\) for \(i = 1, 2, \ldots, \ell\) and \(d_{oi}\) be any arbitrary nonnegative real number for \(i = \ell + 1, \ell + 2, \ldots, p.\) Then \(\hat{\beta}[D_0]\) minimizes \(G(\hat{\beta}[D])\) within the class \(C_6\) for given \(\beta\) where \(D_0 = (\delta_{ij}d_{oi});\) that is
\[G(\hat{\beta}[D_0]) = \min_{\hat{\beta}[D] \in C_6} G(\hat{\beta}[D]).\] Moreover,
(15) \[ G(\beta(D)) = \sigma^2 \sum_{i} \frac{1}{\lambda_i + \frac{\sigma^2}{\alpha_i^2}} + \frac{p}{p+1} \alpha_i^2. \]

Proof: From (14)

(16) \[ G(\beta[D]) = \sum_{i} \frac{\sigma^2 \lambda_i + (\alpha_i d_i)^2}{(\lambda_i + d_i)^2} = \sum_{i} \frac{\sigma^2 \lambda_i + (\alpha_i d_i)^2}{(\lambda_i + d_i)^2} + \frac{p}{p+1} \alpha_i^2 \]

Since \( G(\beta[D]) \) is a sum of nonnegative terms, each of which being functionally independent of the others. \( G(\beta[D]) \) will be minimized if each term attains its minimum.

We see that for each \( i = 1, 2, \ldots, l \),

\[
\frac{d}{dd_i} \left( \frac{\sigma^2 \lambda_i + (\alpha_i d_i)^2}{(\lambda_i + d_i)^2} \right) = \frac{-2\lambda_i \sigma^2 + 2\alpha_i^2 \lambda_i d_i}{(\lambda_i + d_i)^3} = 0
\]

gives \( d_{oi} = \frac{\sigma^2}{\alpha_i^2} \).

\[
\frac{d^2}{dd_i^2} \left( \frac{\sigma^2 \lambda_i + (\alpha_i d_i)^2}{(\lambda_i + d_i)^2} \right) \bigg|_{d_i = d_{oi}} = \frac{6\lambda_i \sigma^2 + 2\alpha_i^2 \lambda_i^2 - 4\alpha_i^2 \lambda_i d_i}{(\lambda_i + d_i)^4} \bigg|_{d_i = d_{oi}}
\]

\[
= \frac{6\lambda_i \sigma^2 + 2\alpha_i^2 \lambda_i^2 - 4\alpha_i^2 \lambda_i (\sigma^2/\alpha_i^2)}{(\lambda_i + \frac{\sigma^2}{\alpha_i^2})^4} > 0
\]

implies that \( d_{oi} = \frac{\sigma^2}{\alpha_i^2} \) minimizes \( \frac{\sigma^2 \lambda_i + (\alpha_i d_i)^2}{(\lambda_i + d_i)^2} \) for each \( i = 1, 2, \ldots, l \).

For \( i = l + 1, l + 2, \ldots, p \) since \( \alpha_i^2 \) is independent of \( d_i \),

\( d_i \) can be assigned to have any real nonnegative value. This proves the first part of the theorem.
To prove (15), the minimizing values of $d_i$ are substituted into (16)

$$G(\hat{\beta}[D]) = \sum_{1}^{l} \frac{\sigma^2 \lambda_1 + \alpha_1^2 (\sigma^2/\alpha_1^2)}{(\lambda_1 + \sigma^2/\alpha_1^2)^2} + \sum_{l+1}^{p} \alpha_1^2$$

This completes the proof of the theorem.

The following two corollaries are immediate consequences of the above theorem.

**Corollary 1.** If $X'X$ is nonsingular then $d_{oi} = (\sigma^2/\alpha_1^2)$ for $i = 1, 2, \ldots, p$ and

$$G(\hat{\beta}[D]) = \min_{\hat{\beta}[D] \in C_6} G(\hat{\beta}[D])$$

**Corollary 2.** If $X'X$ is nonsingular then

$$G(\hat{\beta}[D]) = \sigma^2 \sum_{1}^{l} \frac{1}{\lambda_1 + (\sigma^2/\alpha_1^2)} < \sigma^2 \sum_{1}^{l} \frac{1}{\lambda_1} = G(\hat{\beta}).$$

Hence the estimator in $C_6$ with minimum mean square error has mean square error less than that of $\hat{\beta}$, the least squares estimator.

4. Minimum Mean Square Error For $C_1$ and $C_4$

In this section we will state and prove results similar to Theorem 3.1 for the class of shrunken estimators $C_1$, and for the class of general linear transform estimators $C_4$.

**Theorem 3.2.** If the matrix $X'X$ is nonsingular then $\hat{\beta}(c_o)$ minimizes $G(c_\hat{\beta})$ within the class $C_1$ where
\[ c_o = \frac{\beta' \beta}{\beta' \beta + \sigma^2 \sum \frac{1}{\lambda_i}} \]

Proof: Let \( c \) be any number between 0 and 1, consider the estimator \( c\beta \), where \( \beta = (X'X)^{-1}X'Y \) is the least squares estimator of \( \beta \),

\[ G(c\beta) = E((c\beta - \beta)'(c\beta - \beta)) \]
\[ = E(c\beta - \beta + \beta - \beta)'(c\beta - \beta + \beta - \beta) \]
\[ = c^2 E((\beta - \beta)'(\beta - \beta)) + (1 - c)^2 \beta' \beta \]
\[ = c^2 \sigma^2 \sum \frac{1}{\lambda_i} + (1 - c)^2 \beta' \beta \]

Equating the derivative of \( G(c\beta) \), with respect to \( c \), to zero and solving for \( c \) yields the desired result.

**Corollary 3.** If \( X'X \) is nonsingular, then

\[ G(c_o) = \frac{\beta' \beta}{\beta' \beta + \sigma^2 \sum \frac{1}{\lambda_i}} \sigma^2 \sum \frac{1}{\lambda_i} \leq \sigma^2 \sum \frac{1}{\lambda_i} = G(\beta) \]

Thus the shrunken estimator with minimum mean square error has mean square error less than that of the least squares estimator \( \hat{\beta} \).

Next theorem gives a similar result for the class \( C_4 \).

**Theorem 3.** If the matrix \( X'X \) is nonsingular, then \( \beta(A_o) \) minimizes \( G(\hat{A} \beta) \) within the class \( C_4 \) where

\[ A_o = \beta\beta' (\sigma^2 (X'X)^{-1} + \beta\beta')^{-1}. \]

Proof: Since \( \hat{\beta}(A) = \hat{\beta} \), we have

\[ G(\hat{\beta}(A)) = \sigma^2 \text{tr} A(X'X)^{-1}A' + \beta'(A - I)'(A - I)\beta \]
\[ = \sigma^2 \text{tr} A(X'X)^{-1}A' + \beta'A'\beta - 2\beta'A'\beta + \beta'\beta. \]

Then

\[ \frac{dG(\hat{\beta}(A))}{dA} = 2\sigma^2 A(X'X)^{-1} + 2A\beta\beta' - 2\beta\beta'. \]
Equating \( \frac{da(\hat{\beta}(A))}{dA} \) to zero and using the fact that \( \sigma^2(X'X)^{-1} + \beta' \) is nonsingular, gives
\[
A_0 = \beta'(\sigma^2(X'X)^{-1} + \beta')^{-1}.
\]

It is interesting to note that the optimal member in \( C_4 \) can be written as \( \hat{\beta}(A_0) = c\beta \) where \( c = \beta'(\sigma^2(X'X)^{-1} + \beta')^{-1} \). Thus \( \hat{\beta}(A_0) \) is some sort of projection from the vector \( \hat{\beta} \) onto the parameter vector \( \beta \).

Since \( C_6 \) contains \( C_1 \) and \( C_2 \) as subclasses, we have
\[
G(\hat{\beta}(D_0)) \leq \min_{B \in C_1 \cup C_2} G(B),
\]
and a (possibly) lower mean square error is attained within \( C_6 \) than either \( C_1 \) or \( C_2 \). The minimum will be attained in \( C_1 \) if
\[
\frac{\sigma^2}{\lambda_i^2} \frac{\beta'^2}{\beta' \beta + \sigma^2 \sum \frac{1}{\lambda_i}}
\]
for each \( i = 1, 2, \ldots, p \), by Theorem 3.1 and Theorem 3.2. It will be attained in \( C_2 \) if \( \frac{\sigma^2}{\alpha_i} = k \) for each \( i = 1, 2, \ldots, p \), since in that case the minimizing matrix \( D \) in \( C_6 \) is just a diagonal matrix with equal entries on the diagonal, and \( \hat{\beta}(D) \) is in \( C_2 \).

Although optimal estimators are obtained in the classes \( C_1, C_4 \) and \( C_6 \), a disadvantage of all these optimal estimators is that they depend on the unknown parameter vector \( \beta \). Thus, we cannot specify the optimal estimator exactly unless \( \beta \) itself is known, but, of course, knowledge of \( \beta \) is the prime purpose of the study. Hence, in order to apply the theorems we have to estimate \( \beta \) first. Some possible way in handling this situation will be discussed in Chapter 5.
5. Minimum Generalized Variance on the Ellipsoid \( \hat{\beta}(\beta) = \hat{\phi} \) For \( C_6 \)

In estimating the parameter vector \( \beta \) in a linear model one optimality criterion often used is that of minimizing \( V(\cdot) \), the generalized variance, subject to the constraint that the squared loss function, \( \hat{\phi}(\beta) \), is constant. The following theorem is thus of interest.

**Theorem 3.4.** Let the eigenvalues of the matrix \( X'X \) be such that 
\[
\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_\ell > 0 \quad (1 \leq \ell \leq p) \quad \text{and} \quad \lambda_{\ell+1} = \lambda_{\ell+2} = \cdots = \lambda_p = 0,
\]
let
\[
d_{li} = \frac{\sigma^2}{\gamma_1 \lambda_i \alpha_i^2} \quad \text{for } i = 1, 2, \ldots, \ell,
\]
and \( d_{li} \) be any arbitrary nonnegative real number for \( i = \ell+1, \ell+2, \ldots, p \), where \( \gamma_1 \) is chosen to satisfy \( \hat{\phi}(\beta[D_1]) = \hat{\phi}_0 \), a given constant, with \( D_1 = (\delta_1, d_{li}) \). Then \( \hat{\beta}[D_1] \) minimizes \( V(\hat{\beta}[D]) \), \( \hat{\beta}[D] \in C_6 \), on the ellipsoid \( \hat{\phi}(\hat{\beta}[D]) = \hat{\phi}_0 \).

**Proof:** The proof of this theorem follows the same spirit as that given in the proof of Theorem 3.1. By the method of Lagrangian's multiplier we want to minimize
\[
F(d_1, d_2, \ldots, d_p) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + d_i)^2} + \gamma_1 (\hat{\phi}(\beta) + \frac{\sigma^2 \lambda_i \alpha_i^2}{(\lambda_i + d_i)^2} - \hat{\phi}_0)
\]
\[
= \sum_{i=1}^{\ell} \frac{\sigma^2 \lambda_i + \gamma_1 \alpha_i^2 \lambda_i d_i^2}{(\lambda_i + d_i)^2} + \gamma_1 (\hat{\phi}(\beta) - \hat{\phi}_0)
\]

For each \( i = 1, 2, \ldots, \ell \)
\[
\frac{\partial}{\partial d_i} \left( \frac{\sigma^2 \lambda_i + \gamma_1 \alpha_i^2 \lambda_i d_i^2}{(\lambda_i + d_i)^2} \right) = \frac{-2\sigma^2 \lambda_i + 2\gamma_1 \alpha_i^2 \lambda_i d_i^2}{(\lambda_i + d_i)^3} = 0
\]
gives
\[
d_{li} = \frac{\sigma^2}{\gamma_1 \lambda_i \alpha_i^2}.
\]
Now
\[ \frac{d^2}{dd_1^2} \left( \frac{\sigma^2 \lambda_1 + \gamma \alpha_1 \lambda_1 d_1^2}{(\lambda_1 + d_1)^2} \right) \bigg|_{d_1 = d_{11}} = \frac{6\sigma^2 \lambda_1 + 2\gamma \alpha_1 \lambda_1^2 (\lambda_1 - 2d_1)}{(\lambda_1 + d_1)^4} \bigg|_{d_1 = d_{11}} \]
\[ = \frac{2\sigma^2 \lambda_1 + 2\gamma \alpha_1 \lambda_1^3}{(\lambda_1 + \sigma^2/(\gamma \lambda_1 \alpha_1))^4} > 0 \]
says that \( d_{11} = \frac{\sigma^2}{\gamma \lambda_1 \alpha_1} \) minimizes
\[ \frac{\sigma^2 \lambda_1 + \gamma \alpha_1 \lambda_1 d_1^2}{(\lambda_1 + d_1)^2} \]
for each \( i = 1, 2, \ldots, \ell \).

For \( i = \ell+1, \ell+2, \ldots, p \), \( F(d_1, d_2, \ldots, d_p) \) is independent of \( d_1 \), so for \( i = \ell+1, \ell+2, \ldots, p \), \( d_1 \) can be assigned to any nonnegative value arbitrarily. Hence \( \{d_{11}, d_{12}, \ldots, d_{1p}\} \) minimizes \( F(d_1, d_2, \ldots, d_p) \). Therefore, \( \hat{\beta}(D) \) minimizes \( V(\hat{\beta}(D)) \) subject to \( \hat{\beta}(D) = \hat{\beta}_0 \).

As we pointed out at the end of last section the optimal estimator obtained in Theorem 3.1 depends on the unknown parameter vector \( \beta \). But, the one obtained in the above theorem is completely specified by the data. Hence Theorem 3.4 will be more applicable than Theorem 3.1.

6. Minimum Generalized Variance on the Ellipsoid \( \hat{\beta}(\beta) = \hat{\beta}_0 \) For \( C_1 \) and \( C_4 \)

In this section we will prove results similar to Theorem 3.4 for the class \( C_1 \) and \( C_4 \). The first theorem is pertinent to the class of shrunken estimators which is a straightforward result, so we just state the result without proof.
Theorem 3.5. If \( X'X \) is nonsingular, then \( \hat{\beta}(c) \) minimizes \( V(\hat{\beta}(c)) \), \( \hat{\beta}(c) \in C_1 \) on the ellipsoid \( \hat{\beta}(c) = \hat{\beta}_0 \), a given constant, where

\[
c_1 = 1 - \frac{\hat{\beta}_0 - \hat{\beta}(c)}{\hat{\beta}'X'X\hat{\beta}} \quad \text{if} \quad 1 - \frac{\hat{\beta}_0 - \hat{\beta}(c)}{\hat{\beta}'X'X\hat{\beta}} > 0
\]

\[
c_1 = 0 \quad \text{otherwise.}
\]

The following theorem gives us the optimal estimator for the class of general linear transform estimators.

Theorem 3.6. If the matrix \( X'X \) is nonsingular, then minimizing \( V(\hat{A}\hat{\beta}) \), \( \hat{A}\hat{\beta} \in C_4 \), on the ellipsoid \( \hat{\beta}(\hat{A}\hat{\beta}) = \hat{\beta}_0 \), a given constant, is attained at \( \hat{A}\hat{\beta} \) where

\[
\hat{A}_1 = (a(1), a(2), \ldots, a(p))
\]

and \( a(1), a(2), \ldots, a(p) \) are obtained from

\[
\begin{pmatrix}
a(1) \\
a(2) \\
\vdots \\
a(p)
\end{pmatrix}
= \gamma(\sigma^2(X'X)^{-1} \otimes I + \hat{\beta}'\hat{\beta} X'X)^{-1}
\begin{pmatrix}
b(1) \\
b(2) \\
\vdots \\
b(p)
\end{pmatrix}
\]

where \( b(1), b(2), \ldots, b(p) \) are column vectors of \( X'X\hat{\beta}' \), \( \gamma \) is chosen to satisfy \( \hat{\beta}(\hat{A}_1\hat{\beta}) = \hat{\beta}_0 \), and \( D \otimes E \) means the Kronecker product of \( D \) and \( E \) (see [4] [9] and [12]).

Proof: We prove the theorem by the method of Lagrange multipliers.

Consider the function

\[
F(A) = \sigma^2 \text{tr} A(X'X)^{-1}A' + \gamma((Y - X\hat{\beta})'(Y - X\hat{\beta}) - \hat{\phi}_0)
\]

\[
= \sigma^2 \text{tr} A(X'X)^{-1}A' + \gamma(\hat{\beta}'\hat{\beta} + \beta'X'X\hat{\beta} - 2\hat{\phi}_0)
\]

Differentiating \( F(A) \) with respect to \( A \) gives

\[
\frac{dF(A)}{dA} = 2\sigma^2 A(X'X)^{-1} + \gamma(2X'X\hat{\beta}' - 2X'X\hat{\beta}')
\]
\[ \frac{dF(A)}{dA} = 0 \] gives

(17) \[ \sigma^2 A(X'X)^{-1} + \gamma X'X \beta \beta' = \gamma X'X \beta \beta'. \]

To solve the above matrix equation, write

\[ A = (a^{(1)}, a^{(2)}, \ldots, a^{(p)}) \]

and suppose that \[ \gamma X'X \beta \beta' = \gamma (b^{(1)}, b^{(2)}, \ldots, b^{(p)}) \], then (17) can be expressed as (see page 9 of [9])

(18) \[ \sigma^2 (X'X)^{-1} \otimes I_p + \gamma \hat{\beta} \otimes X'X \]

\[ \left( \begin{array}{c} a^{(1)} \\ a^{(2)} \\ \vdots \\ a^{(p)} \end{array} \right) = \gamma \left( \begin{array}{c} b^{(1)} \\ b^{(2)} \\ \vdots \\ b^{(p)} \end{array} \right) \]

Since \( (X'X)^{-1} \) and \( I_p \) are nonsingular, so as \( (X'X)^{-1} \otimes I_p \).

Therefore, \( \sigma^2 (X'X)^{-1} \otimes I_p + \gamma \hat{\beta} \otimes X'X \) is nonsingular, and the solution to (18) is

\[ \left( \begin{array}{c} a^{(1)} \\ a^{(2)} \\ \vdots \\ a^{(p)} \end{array} \right) = \gamma \left( \begin{array}{c} b^{(1)} \\ b^{(2)} \\ \vdots \\ b^{(p)} \end{array} \right) \]

Hence, the matrix \( A_1 \) given in the theorem is a solution to \( (17) \).

This completes the proof of the theorem.

As in the case for the class \( C_6 \) of orthogonal ridge estimators the optimal estimator given by this theorem is completely known, although it is difficult to determine \( \gamma \). For this reason in practice Theorem 3.4 is more preferred than Theorem 3.6.
CHAPTER 4

A Decision Theoretic Approach

This Chapter will be devoted to the decision theoretic interpretation of the class of general ridge estimators \( \hat{\beta}_0 \). The notation we use will be coincident with that of [2]. Throughout this Chapter we assume that \( \epsilon \sim N_n(0, \sigma^2 I) \) and the loss function is the quadratic loss \( L \) defined as follows:

\[
L(\beta, B) = (B - \beta)'(B - \beta), \text{ where } B \text{ is an estimator of } \beta.
\]

1. A Bayesian Interpretation of \( \hat{\beta}[A'A] \)

One most desired property of an estimator is that of being Bayes. In this section we are going to establish the fact that for every \( A \) such that \( A'A \) is nonsingular \( \hat{\beta}[A'A] = (X'X + A'A)^{-1}X'X \hat{\beta} \) is the Bayes estimator of \( \beta \) when the prior distribution of \( \beta \) is \( N_p(0, \sigma^2(A'A)^{-1}) \).

Let \( F(Y|\beta) \) denote the distribution function of the observed random vector \( Y \) given \( \beta \) and \( \tau(\beta) \) the prior distribution function of \( \beta \). For any estimator \( B = B(Y) \), the Bayes risk is

\[
r(\tau, B(Y)) = \int \left\{ \int L(\beta, B(Y)) \, dF(Y|\beta) \right\} d\tau(\beta)
\]

where the integral inside the brace is to be integrated over the \( n \)-dimensional Euclidean space \( \mathbb{E}^{(n)} \) and the one outside over the \( p \)-dimensional Euclidean space \( \mathbb{E}^{(p)} \).

By interchanging the order of integration, the Bayes risk of \( B(Y) \) can be expressed as
\[ r(\tau, B(Y)) = \int \left\{ \int L(\beta, B(Y)) \, d\tau(\beta|Y) \right\} \, dF(Y) \]

where \( \tau(\beta|Y) \) is the posterior distribution function of \( \beta \) given \( Y \) and \( F(Y) \) is the distribution function of \( Y \).

By making use of the fact that whenever there is a Bayes decision rule with respect to a prior distribution \( \tau \), there exists a nonrandomized Bayes rule (See page 43 of [2]), we can restrict ourselves to the class of nonrandomized decision rules.

To find a \( B(Y) \) that minimizes (20) we may find a nonrandomized \( B(Y) \) that minimizes
\[ \int L(\beta, B(Y)) \, d\tau(\beta|Y) \]
separately for each \( Y \). Hence the Bayes decision rule minimizes \( \mathcal{L}(\beta, B(Y)) \), the posterior conditional expected loss, given the observation \( Y \).

We see that
\[ \mathcal{L}(\beta, B(Y)) = E(B(Y) - \beta)'(B(Y) - \beta) \]
\[ = E(B(Y) - E(\beta|Y) + E(\beta|Y) - \beta)'(B(Y) - E(\beta|Y) + E(\beta|Y) - \beta) \]
\[ = E(\beta - E(\beta|Y))'(\beta - E(\beta|Y)) + (E(\beta|Y) - B(Y))'(E(\beta|Y) - B(Y)). \]

Since the first term on the right hand side of the equality above does not involve \( B(Y) \) and the second term is always nonnegative for all values of \( B(Y) \), \( \mathcal{L}(\beta, B(Y)) \) will be minimized if and only if
\[ (E(\beta|Y) - B(Y))'(E(\beta|Y) - B(Y)) = 0. \]
Hence the unique Bayes decision rule for \( \beta \) is \( B(Y) = E(\beta|Y) \).

Now we are in a position to prove the following.

**Theorem 4.1.** Let \( A \) be a given matrix such that \( A'A \) in nonsingular, the estimator \( \hat{\beta}[A'A] = (X'X + A'A)^{-1}X'X\hat{\beta} \) is the unique Bayes estimator of \( \beta \) with respect to the squared loss function if the prior
distribution of \( \beta \) is \( \mathcal{N}_p(0, \sigma^2(A'A)^{-1}) \).

Proof: Since the distribution of \( Y \) given \( \beta \) is \( \mathcal{N}(X\beta, \sigma^2I) \) and the prior distribution of \( \beta \) is \( \mathcal{N}_p(0, \sigma^2(A'A)^{-1}) \), the joint probability density function of \( Y \) and \( \beta \) is

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

By direct computation, it can be shown that the posterior probability density function of \( \beta \) given \( Y \) is

\[
f(\beta | Y) = \frac{1}{\sigma^{n/2}} \exp\left[-\frac{1}{2\sigma^2} (\beta - (X'X + A' A)^{-1} X' Y)'(X' X + A' A)^{-1} (\beta - (X'X + A' A)^{-1} X' Y)\right]
\]

This is the density of \( \mathcal{N}_p((X'X + A' A)^{-1} X' Y, \sigma^2(X'X + A' A)^{-1}) \).

Therefore, the Bayes estimator of \( \beta \) is

\[
\hat{\beta}[A'A] = (X'X + A' A)^{-1} X' Y = (X'X + A' A)^{-1} X' X \hat{\beta}.
\]

This proves the theorem.

Corollary 4. The Bayes risk of the Bayes estimator in Theorem 4.1 is

\[
r(\tau, \hat{\beta}[A'A]) = \sigma^2 \text{tr} (X'X + A' A)^{-1}.
\]

Proof: The Bayes risk of \( \hat{\beta}[A'A] \) is

\[
r(\tau, \hat{\beta}[A'A]) = \mathbb{E}(\hat{\beta}[A'A] - \beta)'(\hat{\beta}[A'A] - \beta)
= \mathbb{E}[(\hat{\beta}[A'A] - \beta)'(\hat{\beta}[A'A] - \beta)]
= \mathbb{E}[(\text{tr} \hat{\beta}[A'A] - \beta)'(\hat{\beta}[A'A] - \beta)]
= \mathbb{E}[(\text{tr} \mathbb{E}(\hat{\beta}[A'A] - \beta)'(\hat{\beta}[A'A] - \beta)]
= \mathbb{E}[(\text{tr} \mathbb{E}(\hat{\beta}[A'A] - \beta)(\hat{\beta}[A'A] - \beta))]
\]

But the posterior distribution of \( \beta \) given \( Y \) from Theorem 4.1 is
normal with variance-covariance matrix \( \sigma^2(X'X + A'A)^{-1} \). Hence

\[
\rho(\tau, \hat{\beta}[A'A]) = E[\sigma^2 \text{tr} (X'X + A'A)^{-1}] = \sigma^2 \text{tr} (X'X + A'A)^{-1}.
\]

In Theorem 4.1 if the mean of the prior distribution is not located at 0, instead it is at \( \beta_0 \), that is, if the prior distribution of \( \beta \) is \( N_p(\beta_0, \sigma^2(A'A)^{-1}) \), then it can be shown that the Bayes estimator of \( \beta \) with respect to the loss function (19) is

\[
\tilde{\beta}[A'A] = (X'X + A'A)^{-1}(X'X\beta + A'A\beta_0).
\]

Hence, if the prior distribution of \( \beta \) is \( N_p(\beta_0, \sigma^2(A'A)^{-1}) \) with \( \beta_0 \neq 0 \) a Bayesian statistician might prefer \( \tilde{\beta}[A'A] \) rather than \( \hat{\beta}[A'A] \) in estimating \( \beta \).

2. Admissibility of \( \hat{\beta}[A'A] \)

In the last section we showed that if the loss function (19) is used and if the prior distribution is \( N_p(0, \sigma^2(A'A)^{-1}) \), then \( \hat{\beta}[A'A] \) is the Bayes estimator of \( \beta \). In the present section we will show that \( \hat{\beta}[A'A] \) is admissible (see page 54 of [2] for definition).

**Theorem 4.2.** Let \( A \) be any given matrix such that \( A'A \) is nonsingular. The estimator \( \hat{\beta}[A'A] = (X'X + A'A)^{-1}X'X\beta \) is admissible with respect to the loss function (19).

**Proof:** Suppose that \( \hat{\beta}[A'A] \) is inadmissible, then there exists an estimator \( \hat{\beta} \) of \( \beta \) such that \( \hat{\beta} \neq \hat{\beta}[A'A] \), and

\[
R(\hat{\beta}, \beta) < R(\hat{\beta}[A'A], \beta) \quad \text{for} \quad \beta \in S, \quad \text{where} \quad S \text{ is a subset of the } \quad \mathbb{E}^p \quad \text{p-dimensional Euclidean space}.
\]

And

\[
R(\hat{\beta}, \beta) = R(\hat{\beta}[A'A], \beta) \quad \text{for} \quad \beta \in \mathbb{E}^p - S = S^C.
\]

Note that \( R(\hat{\beta}, \beta) \) is the risk function of \( \hat{\beta} \) for given \( \beta \), the same definition applies to \( R(\hat{\beta}[A'A], \beta) \).
To prove the theorem, we distinguish the following cases: (i) \( \int_S \tau(\beta) > 0 \) and (ii) \( \int_S \tau(\beta) = 0 \), where \( \tau(\beta) \) is the prior distribution function of \( \beta \), which is \( N_p(0,\sigma^2(A'A)^{-1}) \).

For case (i) we see that
\[
r(B,\beta) = \int_{B(\beta)} R(B,\beta) \, d\tau(\beta)
= \int_S R(B,\beta) \, d\tau(\beta) + \int_{S^c} R(B,\beta) \, d\tau(\beta)
\leq \int_S R(\beta[A'\beta],\beta) \, d\tau(\beta) + \int_{S^c} R(\beta[A'\beta],\beta) \, d\tau(\beta)
= r(\beta[A'\beta],\tau).
\]
But Theorem 4.1 tells us that \( \hat{B}[A'\beta] \) is the Bayes estimator of \( \beta \), \( \hat{\beta}[A'\beta] \) minimizes \( r(B,\tau) \) for all nonrandomized \( B \). So \( r(\hat{B},\beta) < r(\hat{\beta}[A'\beta],\beta) \) is impossible. Therefore, \( \hat{\beta}[A'\beta] \) must be admissible.

For case (ii) it can be easily shown that \( r(\hat{\beta}[A'\beta],\beta) = r(B,\beta) \).

By definition \( \hat{B} \) is also a Bayes estimator of \( \beta \). This contradicts to the fact that \( \hat{\beta}[A'\beta] \) is the unique Bayes estimator of \( \beta \). Thus \( \hat{\beta}[A'\beta] \) is an admissible Bayes estimator of \( B \).

With Theorem 4.2 in hand we can now construct a class of admissible estimators of \( \beta \) easily. If \( X'X \) is nonsingular, then the class \( C_1 \) of shrunken estimators and the class \( C_2 \) of ridge estimators are both admissible.

Before we conclude this section we would like to point out the following. Stein [14] and recently Barnachik [1] have shown that \( \hat{\beta}(c) = c\hat{\beta} \) is in general inadmissible, it seems to contradict what we have shown above. However, it should be noted that the loss function
they used is different from the quadratic loss function used in this Chapter. In their papers a normalized mean square error is used as the loss function, which is mainly suitable for the purpose of prediction only.

3. Essentially Complete Class of Decision Rules Based on $\hat{\beta}[A'A]$

Let $A$ be a given matrix, and consider the estimator

$$\hat{\beta}[A'A] = (X'X + A'A)^{-1} X'X\hat{\beta}. $$

Since $\hat{\beta}[A'A]$ is a function of sufficient statistic $\hat{\beta}$ of $\beta$, it is therefore sufficient for $\beta$. Hence by Theorem 1 on Pg. 120 of [2], the collection of decision rules based on $\hat{\beta}[A'A]$ forms an essentially complete class. Thus to look for a good decision rule we only have to consider rules that are functions of $\hat{\beta}[A'A]$.

It should be noted that the above argument applies to any estimator of any other classes of estimators discussed in this research.

In this chapter we have been able to show that the proposed estimator $\hat{\beta}[A'A]$ possesses three of the most desired optimal properties in the decision theoretic framework of estimation. However, we fail to prove that $\hat{\beta}[A'A]$ is minimax, indeed we are quite convinced that it is not a minimax estimator of $\beta$. 
CHAPTER 5
Comparisons Between Classes of Estimators

The main purpose of this chapter is to compare various properties of estimators among different classes discussed in this research. We will discuss these properties separately.

1. Geometrical Interpretation of Estimators

The geometric meaning of various biased estimators can be best visualized with the help of Figure 3. For a given value of $\hat{\xi}(\beta)$, say $\hat{\xi}_0$, the member in $C_\perp$, the class of shrunken estimators, satisfying $\hat{\xi}(\beta) = \hat{\xi}_0$, a constant loss, is the intersection, $\hat{\beta}(c)$, of the straight line $O\hat{\beta}$ and the ellipsoid $\hat{\xi}(\beta) = \hat{\xi}_0$. The ridge procedure searches for the point on the ellipsoid $\hat{\xi}(\beta) = \hat{\xi}_0$ such that the distance from the origin is shortest: this is the intersection, $\hat{\beta}[kI]$, of the ellipsoid $\hat{\xi}(\beta) = \hat{\xi}_0$ and the circle $\beta'\beta = C$ centered on the origin, of smallest possible radius. On the other hand, the general procedure proposed by Mayer and Willke searches for the linear transformation $A$ which transforms $\hat{\beta}$ to a point on the ellipsoid $\hat{\xi}(\beta) = \hat{\xi}_0$ such that $\hat{\beta}(A)$ is shortest, that is, $(A\hat{\beta})'(A\hat{\beta})$ is smallest. Hence the estimator $\hat{\beta}(A)$ proposed by Mayer and Willke coincides with the ridge estimator $\hat{\beta}[kI]$. For any given matrix $A$ such that $A'A$ is nonsingular, $\hat{\beta}[A'A]$ is the point on the ellipsoid $\hat{\xi}(\beta) = \hat{\xi}_0$ such that $\beta'A'\beta$ is minimum. If $A'A = I$, then $\hat{\beta}[A'A]$ and $\hat{\beta}[kI]$ coincide, which says that the ridge estimator is a special case of the proposed estimator.
If the matrices $X'X$ and $A'A$ commute, that is, $X'X A' = A' X' X$, then the ellipsoids $\hat{\Psi}(\hat{\beta}) = \hat{\Psi}_0$ and $\beta' A' A \beta = \text{constant}$ have same (parallel) axes. The location of various estimators on the ellipsoid $\hat{\Psi}(\hat{\beta}) = \hat{\Psi}_o$ are shown in Figure 3. In general, among the estimators discussed above, the ridge estimator $\hat{\beta}[A'A]$ which is identical with $\hat{\beta}(A)$, in this geometrical interpretation, has smallest length.

Note that for a given value $\hat{\Psi}_o$ of $\hat{\Psi}(\beta)$, the optimal points given by various procedures are points on $\hat{\Psi}(\beta) = \hat{\Psi}_o$. They are all fixed once the value of $\hat{\Psi}(\beta)$ is fixed. However, the point $\hat{\beta}[A'A]$ on $\hat{\Psi}(\beta) = \hat{\Psi}_o$ depends on the matrix $A$, so $\hat{\beta}[A'A]$ may lie anywhere on the curve CFD. This is one of the advantages of the estimator $\hat{\beta}[A'A]$ over other biased estimators discussed in this research. If prior knowledge of $\beta$ is available, other biased estimators cannot make use of this extra information so directly. However, prior knowledge can easily be incorporated into the estimator $\hat{\beta}[A'A]$, just by choosing the proper matrix $A$ (see Theorem 4.1 and Theorem 2.1).

2. Optimality Properties of Estimators

In Chapter 3, we see that the optimal member minimizing $G(\cdot)$, the mean square error, or minimizing $V(\cdot)$, the generalized variance, subject to $\hat{\Psi}(\cdot) = \hat{\Psi}_o$ can always be found in the classes $C_1, C_4$ and $C_6$. But the corresponding members in $C_2$ and $C_3$ are difficult to find. Indeed, the generalized inverse procedure proposed by Marquardt is merely a numerical algorithm. Although there exist a member in $C_2$ which minimizes $G(\cdot)$ and a member minimizes $V(\cdot)$ subject to $\hat{\Psi}(\cdot) = \hat{\Psi}_o$, they can only be obtained through the use of iterative procedure if
Figure 3 - Location of Various Optimal Estimators on the Ellipsoid $\hat{\beta}(\beta) = \text{constant}$
the dimensionality of $\beta$ is small, otherwise it is practically impossible to locate these optimal estimators.

Since the shrunken estimator and the ridge estimator are special cases of the proposed estimator $\hat{\beta}[A'A]$, for the purpose of this section we can narrow our discussion to the classes $C_4$, $C_5$, and $C_6$.

As we pointed out earlier, the estimator minimizing $G(\cdot)$ depends on $\sigma^2$ and $\beta$. To obtain the optimal estimator it is necessary to estimate $\sigma^2$ and $\beta$ before the results in the theorems could be applied. One suggestion is given as follows. Denote the optimal estimator minimizing $G(\cdot)$ in $C_4$, $C_5$, and $C_6$ by $\hat{\beta}(\sigma^2, \beta)$. First we estimate $\beta$ by $\hat{\beta}$, the least square estimator of $\beta$ and $\sigma^2$ by $\hat{\sigma}^2 = (Y-X\hat{\beta})' (Y-X\hat{\beta})/(n - \text{rank}(X))$. Then we estimate $\hat{\beta}(\hat{\sigma}^2, \hat{\beta})$ by $\hat{\beta}(\hat{\sigma}^2, \hat{\beta})$. After obtaining $\hat{\beta}(\hat{\sigma}^2, \hat{\beta})$, re-estimate it by $\hat{\beta}(\hat{\sigma}^2, \hat{\beta}(\hat{\sigma}^2, \hat{\beta}))$ where

$$\hat{\sigma}^2 = (Y - X\hat{\beta}(\hat{\sigma}^2, \hat{\beta}))' (Y - X\hat{\beta}(\hat{\sigma}^2, \hat{\beta}))/ (n - \text{rank}(X)).$$

This process can be continued in the obvious fashion as many times as desired. However, it should be noted that, at this stage, we still do not know whether the resulting estimator of $\hat{\beta}(\hat{\sigma}^2, \hat{\beta})$ converges in probability to the estimator itself or not. Studies on convergent procedures of estimating $\hat{\beta}(\hat{\sigma}^2, \hat{\beta})$ will be desired.

Theorems 3.4 and 3.6 give the member in $C_6$ and $C_4$ that minimizes $V(\cdot)$ on the ellipsoid $\hat{\phi}(\cdot) = \hat{\phi}_0$. For $C_6$, this member is $\hat{\beta}[D_1]$ where

$$d_{1i} = \frac{\sigma}{\sqrt{\gamma_{1i}^2 - \lambda_1}} \quad \text{if} \quad \lambda_1 \neq 0, \quad d_{1i} = \text{arbitrary if} \quad \lambda_1 = 0, \quad \text{and}$$

$$D_1 = (e_{ij} d_{1j}).$$

We note that the smaller the eigenvalues $\lambda_1$ the larger the $d_{1i}$
will be. Thus, in order to counter-act the severe nonorthogonality of the matrix $X'X$ we should choose $A$ in such a way that the eigenvalues $d_1$ of $A'A$ are inversely proportional to the eigenvalues $\lambda_1$ of $X'X$, and ordered as $d_1 \leq d_2 \leq \cdots \leq d_p$ if the $\lambda_i$'s are ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Although Theorem 3.6 gives the optimal $A_1$, in general, it is difficult to compute $A_1$. Hence if the dimensionality of $\beta$ is large Theorem 3.6 would be difficult to apply.

3. Decision Theoretic Interpretation of Estimators

In Chapter 4 we studied some decision theoretic properties of the proposed class of estimators $C_5$. Since the class, $C_2$, of ridge estimators is a subclass of $C_5$, it has all the decision theoretic properties that pertain to $C_5$. If $X'X$ is nonsingular the class of shrunken estimators will also have a share of all the optimal properties of $C_5$ studied in Chapter 4. However, estimators in $C_3$ and $C_4$ have no such decision theoretic properties attached to them. This is one of the marked differences between the class of linear transform estimators studied by Mayer and Willke, and the class of estimators proposed in this research.

4. Interpretation of Estimators as Least Squares Estimators

As we have pointed out at the beginning of Chapter 2 although $C_5$ is a subclass of $C_4$, we still think that $C_5$ is worthwhile studying. In this section we will give another reason to support our previous assertion. We are going to show that for every member $\hat{\beta}[A'A]$ in $C_5$, and hence every member in $C_1$ and $C_2$, $\beta[A'A]$ can be regarded as a least squares estimator of $\beta$.

**Theorem 5.1.** For any given matrix $A$, the estimator
\[ \hat{\beta}[A'A] = (X'X + A'A)^{-1} X'\hat{\beta} \]

is equivalent to a least squares estimator when the actual data are supplemented by a fictitious set of data points taken according to an experiment \( A \); the response \( Y \) being set to zero for each of these supplementary data points.

Proof: The supplemented model is

\[
E \begin{pmatrix} Y \\ 0 \end{pmatrix} = \begin{pmatrix} X \\ A \end{pmatrix} \beta,
\]

and the least squares estimator of \( \beta \) based on this supplemented model is

\[
\hat{\beta} = \left( \begin{pmatrix} X \\ A \end{pmatrix}' \begin{pmatrix} X \\ A \end{pmatrix} \right)^{-1} \begin{pmatrix} X \\ A \end{pmatrix}' \begin{pmatrix} Y \\ 0 \end{pmatrix} = (X'X + A'A)^{-1} X'Y = (X'X + A'A)^{-1} X' \hat{\beta} = \hat{\beta}[A'A].
\]

This completes the proof of the theorem.

Remark 5. If the matrix \( A \) of the supplementary experiment is orthogonal, then \( \hat{\beta}[A'A] \) reduces to the ridge estimator \( \hat{\beta}[kI] \).

Remark 6. If the experiment is duplicated several times, while the response \( Y \) of the duplicated data are arbitrarily set to zero, then \( \hat{\beta}[A'A] \) becomes the shrunken estimator \( \hat{\beta}(c) \).

According to the theorems in Chapter 3, if the original experiment \( X \) suffers from severe nonorthogonality, then by augmenting the experiment by another experiment \( A \) which is orthogonal to the original experiment, and the eigenvalues \( d_i \)'s of the \( A'A \) are such that \( d_i \) is inversely proportional to \( \lambda_i \), where the \( \lambda_i \)'s are eigenvalues of \( X'X \). The design matrix of the augmented experiment will then be more orthogonal.

In [8] Marquardt gives the following least squares estimator
interpretation to the generalized inverse estimator $\hat{\beta}(r)$.

**Theorem 5.2.** [8]. The generalized inverse estimator $\hat{\beta}(r)$ is equivalent to a least squares estimator according to one of the following circumstances:

1. $D_{p-r}$ is a null matrix. In this case, the columns of $P_{p-r}$ are imposed constraints and the estimator is immediately seen to be a constrained least squares estimator and is then said to be conditionally unbiased, or

2. $D_{p-r}$ is not precisely a null matrix. In this case, the generalized inverse estimator is equivalent to a least squares estimator when the actual data are supplemented by a fictitious set of data points taken according to an experiment $H_r$; the response $Y$ being set to zero for each of these supplementary data points. The matrix of supplementary data points may be defined as

$$H_r = P_{p-r} D_{p-r}^{1/2} \sqrt{-1}.$$  

The supplementary data are thus unrealizable, having imaginary coordinates.

The proof of this theorem appears in [8], so we will not repeat it here.

Among all the classes of biased estimators we discussed in this research, we see that $C_5$ and $C_6$ have a richer statistical structure than any other classes. Although $C_4$ is the most general class of all, it does not have much decision theoretic properties. This is logical, because as a procedure becomes more and more general, it will lose more and more of the elegant nature it possesses. Also, results about $C_4$ are rather complicated and not so easy to apply as those
pertaining to \( C_6 \). Another problem about the class \( C_4 \) is that if the matrix \( X'X \) is singular, then most of the properties and results about \( C_4 \) are no longer valid, and suitable modification has to be made. This is another advantage of the proposed class of estimator over the general class \( C_4 \); all the properties and results about the proposed class of estimator remain valid under no assumption on the singularity of \( X'X \).
CHAPTER 6

An Example

In this chapter an example is given to show the performance of various optimal properties of different estimators discussed in this research.

Consider the following linear model [8]

\[ y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad i = 1, 2, 3 \]

or in matrix form,

\[ Y = X\beta + \varepsilon \]

with

\[
Y = \begin{pmatrix}
1 \\
2 \\
3
\end{pmatrix}
\quad \text{and} \quad
X = \begin{pmatrix}
\frac{3}{5 \sqrt{2}} & \frac{4}{5 \sqrt{2}} \\
\frac{4}{5 \sqrt{2}} & \frac{3}{5 \sqrt{2}} \\
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}
\end{pmatrix}
\]

The eigenvalues of the matrix

\[ X'X = \begin{pmatrix}
1 & .98 \\
.98 & 1
\end{pmatrix} \]

are \( \lambda_1 = 1.98 \) and \( \lambda_2 = .02 \), and the least squares estimator of \( \beta = (\beta_1, \beta_2)' \) is \( \hat{\beta} = (5.3569, -1.7142)' \).

A close and careful examination of the data shows that \( y \) increases with both \( x_1 \) and \( x_2 \), although not so consistently; hence one would expect both \( \beta_1 \) and \( \beta_2 \) are positive. However, the near singularity of \( X'X \) destroys this intrinsic property of the data structure, as is shown by the least squares estimator \( \hat{\beta} \).

Figure 4 shows the location of various estimator on the ellipse...
Figure 4 - Optimal Estimators of $C_1$ on the Ellipse $\hat{\psi}(\beta) = .8636$. 
\( \psi(\beta) = .8636. \) Note that the generalized inverse estimator \( \beta^{-1}(r) \) corresponds to \( r = 1, \) while \( \beta[D] \to (d_1, d_2) = (.01, 1) \) and \( \gamma = .253. \)

The coordinates, the length and the generalized variance of these estimators in the \( \beta_1 - \beta_2 \) space are given below.

\[
\begin{align*}
\hat{\beta} &= (5.357, -1.714)' \quad |\hat{\beta}| = 5.624 \quad V(\hat{\beta}) = 50.505\sigma^2 \\
\hat{\beta}(c) &= (5.160, -1.651)' \quad |\hat{\beta}(c)| = 5.418 \quad V(\hat{\beta}(c)) = 46.869\sigma^2 \\
\hat{\beta}[kI] &= (2.015, 1.320)' \quad |\hat{\beta}[kI]| = 2.409 \quad V(\hat{\beta}[kI]) = .910\sigma^2 \\
\hat{\beta}^*(r) &= (1.822, 1.822)' \quad |\hat{\beta}^*(r)| = 2.577 \quad V(\hat{\beta}^*(r)) = .505\sigma^2 \\
\hat{\beta}[D] &= (1.803, 1.768)' \quad |\hat{\beta}[D]| = 2.525 \quad V(\hat{\beta}[D]) = .665\sigma^2 \\
\end{align*}
\]

From the above we observe the following:

1. It is surprising that \( \hat{\beta} \) and \( \hat{\beta}(c) \) both lie in the fourth quadrant. This is due to the severe nonorthogonality of the input matrix.

2. Out of all the estimators, the length of \( \hat{\beta}[kI] \) is shortest. This agrees with the geometrical interpretation of various estimators: \( \hat{\beta}[kI] \) is the point on \( \psi(\beta) = .8636 \) that minimizes \( \beta'\beta. \)

3. Comparison of the generalized variance shows that \( \hat{\beta}^*(r) \) and \( \hat{\beta}[D] \) have a \( V(\cdot) \) much smaller than that of other estimators, especially that of the least squares estimator \( \hat{\beta}. \)

It is interesting to note that \( \hat{\beta}^*(r) \in C_\gamma \) and it corresponds to

\[
A = \begin{pmatrix}
0 & 0 \\
\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\
\end{pmatrix}
\]

with \( A'A \) singular.

As we know the location of \( \hat{\beta}[A'A] \) (see Theorem 2.1) on the ellipsoid \( \psi(\beta) = \text{constant} \) depends on the \( A \) we chose. Figure 5
Figure 5 - Location of Optimal Estimators in $C_5$ for different $A$
shows this effect of $A$. In Figure 5, $\beta[\Lambda'A] \overline{\gamma_1}$ corresponds to

$$A_1 = \begin{pmatrix} \sqrt{.01} & 0 \\ 0 & 1 \end{pmatrix} P_i, \ i = 1,2,3,$$

where $P_1$, $P_2$, and $P_3$ are orthogonal transformations which rotate the coordinate axes through an angle of $45^\circ$, $10^\circ$ and $-30^\circ$ respectively.

In Section 2 of Chapter 5 we remarked that the member minimizes $V(\cdot)$ subject to $\hat{s}(\cdot) = \text{constant}$ is difficult to find in all the classes, except $C_6$, of estimators discussed in this research. This optimal member in $C_6$ can be determined as follows.

By Theorem 3.4

$$d_{1i} = \frac{\sigma^2}{\gamma_1 \lambda_i \alpha_i}, \ i = 1,2,\ldots,\ell$$

satisfy the equation

$$\hat{s}(\beta[D])) = \hat{s}(\beta) + \frac{\sum \alpha_i \lambda_i d_{1i}^2}{\sum (\lambda_i + \alpha_i)^2}$$

Thus

$$\hat{s}(\beta[D])) = \hat{s}(\beta) + \frac{\sum \alpha_i \lambda_i d_{1i}^2}{\sum (\lambda_i + \alpha_i)^2}$$

Regarding $\hat{s}(\beta[D]))$ as a function of $\gamma_1$, assuming that $\sigma^2$ is known, the graph of $\hat{s}(\beta[D]))$ can be drawn. Since $\hat{s}(\beta[D]))$ is a decreasing function in $\gamma_1$, for a given value $\tilde{s}$ of $\hat{s}(\beta[D]))$ the unique $\gamma_1$ can be determined easily.

For our example, if we take $\sigma^2 = 1$, and $\hat{s}_o = .8636$, we obtain $\gamma_1 = 1.23$ and $V(\hat{\beta}[D]) = .4311$ where $\hat{\beta}[D] = (.8092, 1.7230)$ with

$d_{11} = .06188$ and $d_{12} = 1.626$.

A study of the concentration ellipses of various biased estimators on the ellipse $\hat{s}(\beta) = .8636$ have been performed and the results are summarized in Figure 6 and Figure 7. Figure 6 corresponds to a given
value of $\beta = (1,5)'$ and Figure 7 to $\beta = (4,3)'$.

A word of caution should be made here. When the distribution of the estimator is not a degenerated one, the usual definition of concentration ellipsoid is used. Otherwise the following definition for the term will be adopted. Let the $p$-dimensional random variable $Z$ be distributed with mean $\mu$ and variance-covariance matrix $\Sigma$. Assume that rank $(\Sigma) = q < p$. The concentration ellipsoid of $Z$ is defined to be $(Z - \mu)' \Sigma^{-1}(Z - \mu) = q + 2$, where $\Sigma^{-1}$ is a generalized inverse of $\Sigma$.

The concentration ellipses are those of the estimators given on page 44 with the exception that $\hat{\beta}[D_1]$ is the member in $C_6$ that minimizes $V(\cdot)$ subject to $\hat{\delta}(\cdot) = .3636$. The concentration ellipse of a particular estimator is indexed by that particular estimator.

From Figures 6 and 7 we note that:

1. The shrunken estimator $\hat{\beta}(c)$ although does better than the least squares estimator $\hat{\beta}$, its improvement over the least squares estimator is not very great because of the severe nonorthogonality of $X'X$.

2. The ridge estimator $\hat{\beta}[k\mathbf{I}]$, obtained by adding a small amount of positive quantity to the diagonal of $X'X$, does a much better job than both $\hat{\beta}$ and $\hat{\beta}(c)$. However, it still suffers from nonorthogonality of the input variables. It is perhaps due to the fact that addition of equal amount to the diagonal of $X'X$ has not been able to remove much of the effect of nonorthogonality because of the large ratio of $\lambda_1$ and $\lambda_2$.

3. The effect of the estimator $\hat{\beta}[D_1]$ is really surprising.
Figure 6 - Concentration Ellipses of Various Estimators for $\beta_0 = (1, 5)^T$
Figure 7 - Concentration Ellipses of Various Estimators for $\beta_0 = (4,3)'$
It not only does much better than another estimators, especially $\hat{\beta}$ and $\hat{\beta}(c)$, it also has the ability of counteracting the severe nonorthogonality effect of the input variables. This is mainly due to the fact that the ratio of $\frac{d_{11}}{d_{12}}$ and $\frac{d_{22}}{d_{12}}$ are inversely proportional to that of $\lambda_1$ and $\lambda_2$, and hence brings in this counteraction.

(4) The concentration ellipse of $\hat{\beta}(r)$ is degenerated, it is just a straight line along the direction of the eigenvector associated with $\lambda_1 = 1.98$ on the plane. It is so because $\hat{\beta}(r)$ corresponds to an assigned rank $r = 1$.

Assuming that $\beta = \beta_0$, a known value, is the true value of the parameter vector $\beta$, the generalized variance and the squared bias, and hence the mean square error, of various estimators are computed and given below.

1) For $\beta_0 = (1,5)'$,

\[
\begin{align*}
G(\beta) &= V(\beta) + D(\beta) \\
G(\hat{\beta}(c)) &= 46.860 \sigma^2 + .035 \\
G(\hat{\beta}(r)) &= .505 \sigma^2 + 8 \\
G(\hat{\beta}[D_1]) &= .482 \sigma^2 + 7.824
\end{align*}
\]

2) For $\beta_0 = (4,3)'$

\[
\begin{align*}
G(\hat{\beta}) &= 50.505 \sigma^2 \\
G(\hat{\beta}(c)) &= 46.860 \sigma^2 + .034 \\
G(\hat{\beta}[D_1]) &= .910 \sigma^2 + .581 \\
G(\hat{\beta}[D_1]) &= .487 \sigma^2 + .504
\end{align*}
\]

We see that in both cases, all the estimators, except the shrunken estimator, are much better than the least squares estimator.

In this example we see that the performance of the proposed class of estimators is, in general, much better than that of any other classes of estimators discussed in this research, which agrees to the theory we have developed in this research.
CHAPTER 7

Conclusion

In this research we see that in the general linear model if the product of the input matrix and its transpose is ill-conditioned, then the least squares estimator of the parameter vector will be too unreliable for many purposes. Several procedures have been proposed and studied in order to untangle this situation. Although all these procedures do a much better job than the least squares procedure, it seems that none of them is doing exactly what it should be done.

The shrunken estimator, proposed by James and Stein, merely shortens the length of the least squares estimator and is obviously not especially attractive. Hoerl and Kennard proposed the ridge procedure by adding a small positive quantity to the diagonal of the product of the input matrix and its transpose. This procedure is much superior to the least squares and that proposed by James and Stein. However, adding equal amount of positive quantity to each of the eigenvalues of the product of the input matrix and its transpose does not recognize the non-uniformity of the distribution of the eigenvalues of the product matrix, it shifts all of the eigenvalues away from zero the same amount whether it is necessary or not. The effect of the procedure proposed in this research is different. It not only shifts the eigenvalues away from zero, it augments each eigenvalue only to the amount that is necessary; optimally, it adds to each eigenvalue a positive quantity, in some way, inversely proportional to the eigenvalue itself. In fact, the proposed procedure is equivalent to augmenting the original experiment by
by another experiment which is orthogonal to the original one. Thetheoretically, corresponding to every general linear model there is a true experiment, existing somewhere in the universe, that gives us the true value of the parameter vector. But the problem is that we do not know where is this true experiment, that is why we perform experiment and collect data to estimate the unknown parameter vector in the model. If it turns out that the experiment is badly nonorthogonal, then augmenting it by another experiment which is orthogonal to the original one seems reasonable when we have no information about the true experiment.

Marquardt's procedure and that developed by Rao and others (see [11] [12] for example) rely heavily on the notion of generalized inverse of matrices. However, Rao's generalized inverse procedure is different from Marquardt's generalized inverse procedure. Marquardt's procedure assigns a rank to the product of the input matrix and its transpose, and the estimator he proposed is then based on the generalized inverse of the resulting product matrix. Further investigation of Marquardt's generalized inverse estimator would be desired. But Rao's procedure based on the generalized inverse of the product matrix, thus gives a unified theory in estimating the parameter vector in the general linear model. It would be interesting to compare the results obtained in this research with that obtained by Rao's generalized inverse procedure. Also it might be of interest to develop theory similar to that of [12] for the proposed class of estimators.

The procedure proposed by Mayer and Willke generalizes all the procedures discussed above. Because of the generalization of the procedure, not every member in that class has good statistical meanings attached to it, as it does for the members of the proposed class.
Note that all other biased procedures discussed in this research assume that the product of the input matrix and its transpose is nonsingular, though it may be ill-conditioned. For the proposed estimator no assumption on singularity is made. Since the shrunken estimator and the ridge estimator are special cases of the proposed estimator, the assumption that the product of the input matrix and its transpose is nonsingular can be dropped. Also for the linear transform estimator proposed by Mayer and Willke, slight modification can be made in order to include the singular case (see Appendix).
Appendix

As we have pointed out in Chapter 7 that for the class $C_4$ the assumption that $X'X$ is nonsingular is necessary. In order to include the case that $X'X$ is singular, we consider the following class of estimators.

$$C_7 = \{ \hat{\beta}(A) | \hat{\beta}(A) = AX'X\hat{\beta}, \text{ where } A \text{ is a } p \times p \text{ matrix and } \hat{\beta} \text{ is a solution to } X'X\hat{\beta} = X'Y \}.$$ 

First we note that if $X'X$ is nonsingular, we have the obvious relationship between members of $C_4$ and $C_7$.

$$\hat{\beta}'(A) = \hat{\beta}(AX'X).$$

Therefore $C_4 \subset C_7$. Secondly, if $X'X$ is singular we can write

$$\hat{\beta}(A) = AX'Y$$

and hence the following results are immediate.

(i) $E(\hat{\beta}(A)) = AX'X\beta$.

(ii) $\text{Var}(\hat{\beta}(A)) = AX'XA'\sigma^2$

(iii) $\delta(\hat{\beta}(A)) = \delta(\hat{\beta}) + \delta'(AX'X - I)'X'X(AX'X - I)\hat{\beta}$

(iv) $V(\hat{\beta}(A)) = \delta^2 \text{tr } AX'XA'$

(v) $D(\hat{\beta}(A)) = \delta'(AX'X - I)'(AX'X - I)\beta$.

The following theorem is a generalization of Theorem 3.3 of Chapter 3.

**Theorem.** Let $A_0$ be an $p \times p$ matrix such that

$$A_0X'X(I + \frac{1}{\sigma^2} \beta\beta'X'X) = \frac{1}{\sigma^2} \beta\beta'X'X,$$

then $\hat{\beta}(A_0)$ minimizes $G(\hat{\beta}(A))$ within the class $C_7$.

**Proof:** From (iv) and (v) above we have

$$G(\hat{\beta}(A)) = V(\hat{\beta}(A)) + D(\hat{\beta}(A))$$

$$= \delta^2 \text{tr } AX'XA' + \delta'(AX'X - I)'(AX'X - I)\beta$$
\[ = \sigma^2 \text{tr } AX'XA' + \beta'X'XA'X + \beta'X'X - 2\beta'X'XA' + \beta' \beta \]

and

\[ \frac{dG(\hat{\beta}(A))}{dA} = 2(\sigma^2 AX'X + AX'X\beta'X'X - \beta\beta'X'X), \]

setting \( \frac{dG(\hat{\beta}(A))}{dA} = 0 \) gives

(21)

\[ AX'X(I + \frac{1}{\sigma^2} \beta\beta'X'X) = \frac{1}{\sigma^2} \beta\beta'X'X. \]

Thus \( \hat{\beta}(A_o) = A_o X'X^\wedge = \frac{1}{\sigma^2} \beta\beta'X'X(I + \frac{1}{\sigma^2} \beta\beta'X'X)^{-1} \) minimizes

\( G(\hat{\beta}(A)) \) with the class \( C_7 \).

Note that \( A_o \) may not be unique, but \( \hat{\beta}(A_o) \) is unique for any \( A_o \) satisfying (21). Thus we always get a unique member in \( C_7 \) that minimizes \( G(\hat{\beta}(A)) \) no matter \( X'X \) is singular or not.

Although \( C_7 \) is the most general class of biased estimators among all the classes we discussed in this research, it is regretted that we fail to establish results similar to Theorem 2.1 and Theorem 3.6 because of the complication of the mathematics involved.
References


