DIRECTIONALLY MINIMAX MEAN SQUARE ERROR ESTIMATION IN LINEAR MODELS

by

Guillermo Pedro Zarate de Lara

Institute of Statistics
Mimeograph Series No. 1102
Raleigh, N.C.

ABSTRACT

ZARATE DE LARA, GUILLERMO PEDRO. Directionally Minimax Mean Square Error Estimation In Linear Models. (Under the direction of THOMAS M. GERIG and ROBERT J. MONROE.)

A minimax criterion, using the trace of the mean squared error matrix, for estimating regression parameters in a linear model is proposed. It is shown that with respect to this criterion non-homogeneous linear estimators are inadmissible. Estimators are derived using the criterion for (i) the class of ordinary least squares estimators computed subject to false restrictions, (ii) the class of shrinkage estimators and (iii) the class of general linear estimators.

The criterion proposed above is then generalized by applying the minimax argument directly to the mean square error matrix. A matrix ordering, and a least upper bound with respect to this ordering, is defined and a set of sufficient conditions for its existence are presented. It is shown that the estimator obtained by using this criterion can be seen as a generalization of the Ridge Regression Estimator of Hoerl and Kennard. The form of the estimator is used to propose a joint estimator for the regression coefficients of a set of dependent regression equations describing different variables but sharing the same design matrix. The estimator is compared with a naive Ridge Regression type and some asymptotic results for both estimators are presented.

BIOGRAPHY

The author was born on August 1, 1946, in Mexico City, Mexico. He was reared primarily in Oaxaca, Mexico, where he received his elementary and secondary education.

He received the Bachelor of Science degree with a major in Soil Science from the National School of Agriculture in 1969. Under the support of the "Consejo Nacional de Ciencia y Tecnología" (CONACYT), he then received the Master of Science Degree in Statistics from the "Centro de Estadistica y Cálculo (CEC), del Colegio de Postgraduados" of the National School of Agriculture in 1972, and has been a professor there since 1972.

In 1973, a scholarship was granted to him from the CONACYT to pursue his studies in statistics towards a Doctorate at North Carolina State University.

The author is married to the former Graciela González Kauffmann and they have a daughter - Graciela.

ACKNOWLEDGEMENTS

No man knows fully all the forces and significant people who shaped his thinking and his work. Yet, upon the completion of one of the most demanding undertakings, a dissertation, one cannot escape a feeling of tremendous debt owed to others. Therefore of the many persons who actively contributed to its development, I must acknowledge the following.

Professor Thomas M. Gerig, who from the onset of our relationship as student and advisor, had a genuine interest in my graduate program and supported me with his invaluable counsel and guidance throughout the preparation of this thesis. In addition to showing me the rigors of the statistical thought, he showed me the virtue of being a person without pretence, challenged me to learn and gave me his friendship.

Professor Robert G. D. Steel, who from the first hours of my arrival on campus and throughout my entire graduate program, has supported me with his counsel and friendship.

Professor Robert J. Monroe, for serving on my committee.

Professor Stephen L. Campbell, who always gave me advice with patience, friendship and an enthusiastic attitude.

Professor John W. Bishir, for his counsel and his sincere friendship throughout these years of study.

Professor Ignacio Mendez R., for introducing me to the study of statistics.

Professors Alfonso Carrillo L. and Eduardo Casas D., for their encouragement and help in studying statistics.

My wife, who encouraged me and assisted me through the difficult times, for the belief that I could do it, and for her waiting until it was finished.

My daughter, who always brightened my day.

To all my fellow graduate students, who have been an exemplary model of intellectual and academic excellence, in addition to being true friends.

To Linda Bielawski, who spent many hours typing this thesis.

Finally an eternal thanks to my parents, without whose motivation and devotion this entire educational program would never have been made possible.

A grant from the "Consejo Nacional de Ciencia y Tecnología" has made possible my graduate program. I want to express my gratitude to this institution.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

1. INTRODUCTION AND SUMMARY

The most commonly used estimators of the regression coefficients in general linear models are the best linear unbiased estimators (or ordinary least squares estimators). By dropping the unbiasedness criterion, linear estimators can be obtained that have smaller variances. In this situation, it is common to adopt some type of mean squared error criterion which balances increased bias against reduced variance. Unfortunately, estimators derived from this type of criterion invariably lead to expressions that involve the parameters themselves and, as such, are useless in practice.

Several authors have proposed classes of biased regression estimators which can be used as alternatives to the ordinary least squares (OLS) estimator. These estimators have been studied and compared to OLS estimators with respect to mean squared error. Typically, they are not uniformly better in this respect but in some situations can be dramatically superior. Some of the common methods of biased regression estimation include Hoerl and Kennard's [9] ridge regression, Marquardt's [13] generalized inverse regression, Mayer and Willke's [14] shrunken OLS regression, and Toro and Wallace's [18] false restrictions regression. Many others have proposed modifications and extensions to these.

The purpose of this work is to develop meaningful criteria and use them to derive biased estimators which will be superior to OLS with respect to mean squared error.

In Chapter 4 a criterion is proposed which is based on a minimax argument. In the formula for the trace of the mean squared error

matrix, the value of the regression parameters which are, in a sense, the least favorable to estimation (that is, which result in the greatest bias) are substituted for the unknown parameters. By this, it is hoped that the derived estimators will protect the user from the worst case that nature can produce. It is shown that, with respect to this criterion, non-homogeneous linear estimators are inadmissible.

The criterion is applied to three classes of (homogeneous) linear estimators. First, it is applied to the class of estimators obtained by computing OLS estimators subject to false constraints. In this case it is found that the generalized inverse estimator of Marquardt [13] is optimum. Secondly, the criterion is applied to the class of shrinkage estimators defined by Goldstein and Smith [7]. Here the optimum is found to be a member of the subclass of shrunken OLS estimators defined by Mayer and Willke [14]. Finally, the general class of linear estimators is studied. For this, the optimum turns out to be the same as that for the shrinkage class. It is shown as well that the estimator of a linear combination of the parameters is not in general equal to the same linear combination of the estimators of the parameters.

In Chapter 5, the criterion proposed before is generalized, in the sense that the same minimax argument is applied to the mean square error matrix. In order to do so, a matrix ordering and a supremum or least upper bound with respect to this ordering is defined, and a set of sufficient conditions for the existence of the supremum is presented. When this criterion is minimized, the resulting estimator is shown to be a generalization of the ridge regression estimator of Hoerl and Kennard [9]. It is shown also that, in general, for this

criterion, the estimator of a linear combination of the parameters is equal to the same linear combination of the estimators of the parameters.

In Chapter 6, the form of the estimator obtained in Chapter 5 is used to propose a joint estimator for the regression coefficients of a set of dependent regression equations describing different variables but sharing the same design matrix. These estimators are similar to the "seemingly unrelated regression" estimators of Zellner [21]. The estimator is compared with a naive ridge regression type competitor and some asymptotic results for both estimators are presented.

Finally, in Chapter 7, some ideas and problems that arose in the course of this research are proposed for further studies.

2. NOTATION

Consider the linear model

$$y = X\beta + e \ , \tag{2.1}$$

where y is an $(n \times 1)$ vector of observation, X is an $(n \times m)$ matrix of rank $m(\leq n)$ of known constants, β is an $(m \times 1)$ vector of unknown regression parameters, and e is an $(n \times 1)$ vector of unobservable random variables with $E(e) = 0$ and $E(ee') = \Sigma\sigma^2$. $|\Sigma| \neq 0$. For convenience, we shall refer to this model by $(y, X\beta, \Sigma\sigma^2)$.

Define

$$S = X'X \tag{2.2}$$

and

$$\hat{\beta} = S^{-1}X'y \tag{2.3}$$

and write the singular value decomposition of X as

$$X = U\Lambda^{1/2}V' \ , \tag{2.4}$$

where U is $(n \times m)$, $\Lambda^{1/2} = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_m^{1/2})$, $\lambda_1^{1/2} \geq \ldots \geq \lambda_m^{1/2} > 0$, V is $(m \times m)$, $U'U = I_m$, $V'V = VV' = I_m$, and I_m is the $(m \times m)$ identity matrix. From this we write

$$S = V\Lambda V' \ . \tag{2.5}$$

Define the partitioned matrices

$$U = (U_1 : U_2) \ , \tag{2.6}$$

and

$$V = (V_1 : V_2) \ , \tag{2.7}$$

where V_1 is $(m \times m-u)$, V_2 is $(m \times u)$, U_1 is $(n \times m-u)$, and U_2 is $(n \times u)$ for some u . Similarly, define

$$\Lambda^{1/2} = \begin{vmatrix} \Lambda_1^{1/2} & 0 \\ 0 & \Lambda_2^{1/2} \end{vmatrix} \ , \tag{2.8}$$

where $\Lambda_1^{1/2}$ is $[(m-u) \times (m-u)]$ and $\Lambda_2^{1/2}$ is $(u \times u)$ for some u .

We shall be concerned with the simultaneous estimation of the regression parameters, β , using linear functions of the observations $Ay + b$, where A is some $(m \times m)$ matrix and b an $(m \times 1)$ vector. For this we shall need the mean squared error matrix:

$$M = M(Ay + b,\beta) = M(Ay + b,\beta;\beta,\Sigma\sigma^2)$$

$$= E[(Ay + b - \beta)(Ay + b - \beta)']$$

$$= A\Sigma A'\sigma^2 + [b + (AX - I)\beta][b + (AX - I)\beta]' \ , \tag{2.9}$$

and we shall need the trace of M ,

$$T = T(Ay + b,\beta) = T(Ay + b,\beta;\beta,\Sigma\sigma^2)$$

$$= tr[M(Ay + b,\beta)] \ . \tag{2.10}$$

Furthermore,

$$M(Ay, P'\beta; \beta, \Sigma\sigma^2) = A\Sigma A'\sigma^2 + (P' - AX)\beta\beta'(P' - AX)' \qquad (2.11)$$

will be used as the MSE matrix for the estimator Ay of parameter $P'\beta$.

The arguments β and $\Sigma\sigma^2$ are included to emphasize the dependence of the above expressions on them. They will be omitted from the notation when there is no danger of confusion.

Finally, define

(i) $\|x\|^2 = x'x$, (2.12)

(ii) $Ch_M(A)$ is the largest characteristic root of the matrix A, (2.13)

$Ch_m(A)$ is the smallest characteristic root of the matrix A, (2.14)

(iii) $\mathcal{C}(A)$ is the vector space generated by the columns of A, (2.15)

(iv) $\mathcal{R}(A)$ is the vector space generated by the rows of A, (2.16)

(v) $\mathcal{R}^{\perp}(A)$ is the vector space orthogonal to the vector space generated by the rows of A, (2.17)

(vi) $\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$, where A is a $(p \times q)$ matrix. (2.18)

3. LITERATURE REVIEW

A problem that can occur in regression analysis is that, when the model $(y, X\beta, \Sigma\sigma^2)$, $|\Sigma| \neq 0$ is correctly specified, some columns of the X matrix come quite close to being linearly dependent. This results in some very small but positive eigenvalues of the X'X matrix, and, therefore, in large variances for the least squares (LS) estimates of some linear combinations of the regression coefficients. Some comments about how this problem arises in econometrics are given by Chipman [3] who says that it can arise in two different ways:

> (1) It is very common that a "true" specification would usually result in the number of variables exceeding the number of observations. In demand analysis, from the point of view of pure theory all prices should, in principle, be included as determining variables. This fact is usually obscured by the practice of neglecting some variables that do not seem individually to be of great importance, or by combining some variables linearly in order to obtain aggregate variables. Such practices are usually ad hoc, finding justification neither in economics nor statistical theory.
>
> (2) Even if the number of observations exceeds the number of variables, a number of columns of the observation matrix may be linearly dependent, or nearly so. This occurs in econometric time series analysis when some variables (such as prices) move up and down together or remain stationary; or in cross section analysis when different consumers face the same set of market prices.

The two key properties of the OLS estimator

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

are that it is unbiased:

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

and that it has minimum variance among all linear unbiased estimators. The variance is

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

By dropping the unbiasedness criterion, linear estimators can be obtained that have smaller variances. In this situation it is common to adopt some type of mean square error (MSE) criterion which balances increased bias against reduced variance. Some of the earlier work done in the area of estimation using a MSE criterion includes that of Chipman [3]. He uses the concept of minimum MSE procedure to obtain estimates of the β parameter. The approach given in his paper is delineated in the following lines.

Let the $(m \times 1)$ vector β have a prior probability distribution with

$$E\beta = \bar{\beta} ,$$

$$E(\beta - \bar{\beta})(\beta - \bar{\beta})' = C = \tau^2 \Phi .$$

Similarly, let the $(n \times 1)$ random vector e have mean and variance

$$E(e) = 0 ,$$

$$E(ee') = \sigma^2 \Sigma = D .$$

Assume further that β and e are uncorrelated; i.e.,

$$E(\beta - \bar{\beta})e' = 0 .$$

He defines the deviation of β from its mean as

$$\bar{\bar{\beta}} = \bar{\beta} - \beta \ ,$$

thus,

$$E \begin{bmatrix} \bar{\bar{\beta}} \\ e \end{bmatrix} = 0 \ ; \quad \text{Var} \begin{bmatrix} \bar{\bar{\beta}} \\ e \end{bmatrix} = \begin{bmatrix} C & 0 \\ 0 & D \end{bmatrix} \ .$$

From the probability distribution of β and e, Chipman derives the conditional distribution of $y = X\beta + e$ given, respectively, β and e, with mean and variance

$$E(y|\beta) = X\beta \ ; \quad \text{Var}(y|\beta) = D$$

and

$$E(y|e) = X\bar{\beta} + e \ ; \quad \text{Var}(y|e) = XCX' \ .$$

Thus, the unconditional distribution of y has mean and variance

$$Ey = X\bar{\beta} \ ; \quad \text{Var}(y) = XCX' + D \equiv W$$

which defines W.

Under this setup, he introduces the following definition:

Definition 3.1

A linear estimator $\tilde{\beta} = Ay + b$ is said to be a minimum mean squared error estimator of β if A and b are such that the matrix

$$R = R(A,b) = E(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'$$

is minimum. R may be called the matrix of MSE, or more briefly the risk matrix.

The precise meaning of minimum is stated in Section 5.1. He then proceeds, as in Foster [6, p. 388], first to minimize R with respect to b and then with respect to A , obtaining that R(A,b) is minimized with respect to b when

$$b = (I - AX)\bar{\beta}$$

so that the problem is reduced to finding a matrix A such that

$$R(A) = (I - AX)C(I - AX)' + ADA'$$

is a minimum.

The choice of the matrix A is then given by the following theorem:

Theorem 3.1

Let X be an (n x m) matrix, and let C , D be positive definite matrices of order m and n , respectively. Then there is a unique (m x n) matrix $A = A_0$ (called the optimum inverse of X) which minimizes

$$R = (I - AX)C(I - AX)' + ADA'$$

and it is equal to

$$A_0 = CX'(XCX' + D)^{-1} = (C^{-1} + X'DX)^{-1}X'D^{-1} .$$

the minimum risk then becomes

$$R(A_0) = (I - A_0X)C ,$$

so the estimator becomes

$$\tilde{\beta} = CX'(XCX' + D)^{-1}y = (C^{-1} + X'DX)^{-1}X'y . \qquad (3.1)$$

Toro and Wallace [18] consider using LS estimators in linear models, $(y,X\beta,I\sigma^2)$, calculated subject to false restrictions in hopes of reducing the MSE of estimation. They propose a uniformly most powerful test to check for a given data set whether a particular set of restrictions will reduce MSE. Since the MSE criterion suggests a framework for thinking about the problem of multicollinearity in a linear model, they present some examples to illustrate the linkage of the MSE criterion with multicollinearity.

Toro and Wallace's work consists first in obtaining a class of biased estimators by imposing a set of false restrictions on the parameter space, of the form

$$R\beta = h ,$$

where R is a $(u \times m)$ matrix of known constants with rank $u \leq m$ and h a vector of known constants. Under this setup, the estimator is of the form

$$\tilde{\beta} = \hat{\beta} - S^{-1}R'(RS^{-1}R')^{-1}(R\hat{\beta} - h) ,$$

where $\hat{\beta}$ is the OLS defined in (2.3) and S^{-1} is given in (2.2). The distribution of $\tilde{\beta}$ is

$$\tilde{\beta} \sim N[\beta - S^{-1}R'(RS^{-1}R')^{-1}(R\beta - h), \Sigma_{\tilde{\beta}\tilde{\beta}}] ,$$

where

$$\Sigma_{\tilde{\beta}\tilde{\beta}} = \sigma^2[I - S^{-1}R'(RS^{-1}R')^{-1}R\,S^{-1}].$$

Toro and Wallace make the remark "the restrictions come into play in reducing variances and one should also note that the restriction reduce variances no matter whether it is true or not".

The problem that arises now is to determine whether a specific set of restrictions leads to better estimates, where the criterion "betterness" is taken to be the MSE criterion. To be precise, we have:

> "For $\hat{\theta}$ to be better than θ^* in MSE it is required that for every $(n \times 1)$ vector $d \neq 0$
> $$\text{MSE } d'\hat{\theta} \leq \text{MSE } d'\theta^* \text{ ."}$$

Under this definition, Toro and Wallace found that

$$M(\hat{\beta},\beta) - M(\tilde{\beta},\beta) = \sigma^2 S^{-1}R'(RS^{-1}R')^{-1}[RS^{-1}R' - \frac{1}{\sigma^2}(R\beta - h)(R\beta - h)'](RS^{-1}R')^{-1}RS^{-1}, \qquad (3.2)$$

where $M(\hat{\beta},\beta)$ and $M(\tilde{\beta},\beta)$ are the MSE matrices of $\hat{\beta}$ and $\tilde{\beta}$ respectively, and M is defined in Chapter 2. Moreover, they found that (3.2) is a positive-semidefinite matrix if, and only if,

$$\ell'[RS^{-1}R' - \frac{1}{\sigma^2}(R\beta - h)(R\beta - h)']\ell \geq 0, \quad \text{for all } \ell \neq 0$$

and this occurs if, and only if,

$$\lambda = \frac{(R\beta - h)'(RS^{-1}R')^{-1}(R\beta - h)}{2\sigma^2} \leq \frac{1}{2}. \qquad (3.3)$$

Next they point out:

> Getting the criterion into the form $\lambda \leq 1/2$ is an important step, because λ is a parameter in a well-defined probability density function. Thus we can make a test of the hypotheses that a particular set of restrictions yields better structural estimates than OLS estimator according to the MSE criterion.

In order to develop the statistical test, let

$$Q_1 = SSE(\hat{\beta}) = y'y - \hat{\beta}X'y = y'[I - XS^{-1}X']y$$

$$= y'My \cap \sigma^2\chi^2_{(n-m)} ,$$

$$\frac{Q_0}{\sigma^2} = \frac{SSE(\tilde{\beta}) - SSE(\hat{\beta})}{\sigma^2}$$

$$= \frac{[y - XS^{-1}R'(RS^{-1}R')^{-1}h]'M^*[y - XS^{-1}R'(RS^{-1}R')^{-1}]}{\sigma^2}$$

where

$$M^* = XS^{-1}R'(RS^{-1}R')^{-1}RS^{-1}X' \cap \chi^2(u,\lambda) .$$

Toro and Wallace have shown that Q_1 and Q_0 are independent and therefore the ratio $(n-m)Q_0/mQ_1$ has the non-central F distribution with parameters $(u, n-m, \lambda)$. So the hypothesis that a set of constrained estimators $\tilde{\beta}$ is better than the unconstrained estimators $\hat{\beta}$ according to the generalized MSE criterion, can be written as

$$H_0: \lambda \leq 1/2 \qquad H_1: \lambda > 1/2 .$$

We will accept H_0, if $W \geq W_\alpha$, where

$$W = \frac{(n-m)Q_0}{uQ_1} .$$

Some critical points and power computations are given in [18].

Sclove [19] has obtained point estimators for the coefficients in orthogonal linear regression which are "better" than OLS estimators when at least three coefficients are to be estimated. The measure of goodness of an estimator is the sum of weighted sum of the componentwise MSE. The extention of the results to the general case of non-orthogonal regression is given, where the measure of goodness of an estimator is the mean of a quadratic form in the componentwise error.

The precise meaning of "better" that will be used subsequently is the following: "One estimator is better than another if the sum of its componentwise MSE's is smaller than that of the other, for all parameter values."

Sclove considers first the model for the regression on orthogonal independent variables

$$y_j = Z_j'\beta + e_j , \quad j = 1, \ldots, n$$

where the e_i are i.i.d. $N(0,\sigma^2)$ random variables, and

$$\sum_{j=1}^{n} Z_j Z_j' = I .$$

He defines

$$\beta' = (\beta_1, \ldots, \beta_m) ,$$

$$y' = (y_1, \ldots, y_n) ,$$

$$Z_j' = (Z_{1j}, \ldots, Z_{mj}) ,$$

and

$$Z' = (Z_1', \ldots, Z_n') .$$

The OLS estimator for β is

$$\hat{\beta} = Z'y$$

since $Z'Z = I$. The residual sum of squares is

$$SSE = v = y'y - \hat{\beta}'\hat{\beta} .$$

It is well-known that the statistics $\hat{\beta}$ and v are independent, $\hat{\beta} \wedge N(\beta, \sigma^2 I)$ and $v \wedge \sigma^2 \chi^2_{n-m}$. The measure of goodness of an estimator $\hat{\beta}$ will be a weighted sum of componentwise MSE

$$\gamma_w(\hat{\beta};\beta) = \sum_{i=1}^{m} w_i E[\hat{\beta}_i - \beta_i]^2$$

where $w_i > 0$, $i = 1, \ldots, m$ are given weights. A special case is

$$\gamma_1(\hat{\beta};\beta) = \sum_{i=1}^{m} E(\hat{\beta}_i - \beta_i)^2 .$$

Thus, $\tilde{\beta}$ is "better" than $\hat{\beta}$ if

$$\gamma_w(\tilde{\beta};\beta) < \gamma_w(\hat{\beta};\beta)$$

for all β .

An important series of theorems and corollaries, given by James and Stein [10] and by Baranchik [1], that are used intensively by Sclove and subsequent authors are the following:

Theorem 3.2 (James-Stein [10])

For $m > 3$, the estimator

$$\tilde{\beta} = (1 - \frac{cv}{\hat{\beta}'\hat{\beta}})\hat{\beta}$$

where $0 \leq c \leq \frac{2(m-2)}{(n-m-2)}$, has the property that

$$\gamma_1(\tilde{\beta};\beta) < \gamma_1(\hat{\beta};\beta) ,$$

for all β. For any β, $\gamma_1(\tilde{\beta};\beta)$ is smallest when $c = \frac{(m-2)}{(n-m+2)}$.

Theorem 3.3 (Baranchik [1])

Let $F = \frac{\hat{\beta}'\hat{\beta}}{v}$ and let $a(.)$ be any function such that $a(.)$ is monotone nondecreasing and $0 < a(.) < \frac{2(m-1)}{(n-m+2)}$. Let

$$\psi(\hat{\beta},v) = (1 - \frac{a(F)}{F})\hat{\beta} .$$

Then $\gamma_1(\psi;\beta) < \gamma_1(\hat{\beta};\beta)$, for all β.

Corollary 3.1

The estimator

$$\tilde{\beta}^{+} = (1 - \frac{cv}{\hat{\beta}'\hat{\beta}})^{+}\hat{\beta}$$

where $0 \leq c \leq \frac{2(m-2)}{(n-m+2)}$, is better than $\hat{\beta}$ in the sense that $\gamma_1(\tilde{\beta}^{+};\beta) < \gamma_1(\hat{\beta};\beta)$ for all β.

It should be stressed that all these results are based upon the assumption that the errors are normal.

Sclove applies the results obtained above to the case when we can partition β as $\beta' = (\beta'_{(1)},\beta'_{(2)})$, where $\beta_{(1)}$ is an m vector

and $\beta_{(2)}$ is a q vector $(n = m + q)$. Similarly, partition $\hat{\beta}$ as $\hat{\beta}' = (\hat{\beta}'_{(1)}, \hat{\beta}'_{(2)})$. Let

$$\tilde{\beta}_{(2)} = (1 - \frac{cv}{\hat{\beta}'_{(2)}\hat{\beta}'_{(2)}})\hat{\beta}_{(2)}$$

where $0 < c < \frac{2(q-2)}{(n-m+2)}$. Then by Theorem 3.2,

$$\gamma_1(\tilde{\beta}_{(2)};\beta_{(2)}) < \gamma_1(\hat{\beta}_{(2)};\beta_{(2)})$$

for all $\beta_{(2)}$ and from Corollary 3.1 we have:

Corollary 3.2

The estimator

$$\beta^* = \begin{Bmatrix} \hat{\beta}_{(1)} \\ (1 - \frac{cv}{\hat{\beta}'_{(2)}\hat{\beta}'_{(2)}})\hat{\beta}_{(2)} \end{Bmatrix} \qquad (3.4)$$

where $0 < c < \frac{2(q-2)}{(n-m+2)}$, has the property that

$$\gamma_1(\beta^*;\beta) < \gamma_1(\hat{\beta};\beta)$$

for all β. He points out that, when $\beta_{(2)} = 0$

$$F^* = \frac{\hat{\beta}'_{(2)}\hat{\beta}_{(2)}/q}{v/(n-m)} \sim F_{q,n-m}$$

and the estimate of $\beta_{(2)}$ is zero when $F^* < ((n-m)/q)c$. Use of β^* corresponds to making a preliminary test of hypothesis $\beta_{(2)} = 0$, at a level of significance $\alpha = P\{F_{q,n-m} > ((n-m)/q)c\}$. Hence, using β^* is very much like estimating β as

$$\begin{bmatrix} \hat{\beta}(1) \\ \hat{\beta}(2) \end{bmatrix}$$

when the hypothesis is rejected and as

$$\begin{bmatrix} \hat{\beta}(1) \\ 0 \end{bmatrix}$$

when the hypothesis is accepted. Some critical points are given in [19].

As an extension of the past results to nonorthogonal regression, he considers the model

$$y_j = X_j'\beta_1 + e_j , \quad j = 1, \ldots, n .$$

Letting

$$S = \sum_{j=1}^{n} X_j X_j' ,$$

suppose that $L'SL = I$ so that the transformation from the independent variables X_j to the orthogonal independent variables Z_j is $Z_j = L'X_j$ and $\beta_1 = L\beta$. The estimator (3.4) gives as an estimator for β the statistic

$$\beta^+ = \begin{cases} L_1 \begin{pmatrix} \hat{\beta}_0 \\ 0 \end{pmatrix} & \text{if } F^* \leq \frac{n-m}{q} c \\ L_1\hat{\beta}(1) + (1 - \frac{cv}{\hat{\beta}'(2)\hat{\beta}(2)})L_2\hat{\beta}(2) & \text{if } F^* > \frac{n-m}{q} c , \end{cases}$$

where $L = (L_1, L_2)$ and $\frac{1}{F^*} = \frac{qv}{(n-m)} \hat{\beta}'(1)\hat{\beta}(2)$.

The applicability of the results is limited because of the requirement of estimating at least three regression coefficients.

Another important work in the area of biased estimation in linear regression problems is the one written by Hoerl [9] who showed that in linear regression analysis with nonorthogonal data, it is helpful to augment the diagonal of the normal equation matrix by a small positive quantity in order to prevent "inflation" of the elements of the vector of regression coefficients; later Hoerl and Kennard [9] developed a comprehensive theory supporting Hoerl's procedure, showing that it is possible to improve linear estimation from nonorthogonal data by employing biased estimation, focusing on small MSE rather than least squares.

Hoerl and Kennard's formulation of the problem is as follows: Consider the model $(y, X\beta, I\sigma^2)$, and assume that the matrix $X'X$ has the form of a correlation matrix. If the eigenvalues of the $X'X$ matrix are denoted as in (2.5) by λ_i , $i = 1, \ldots, m$, then a seriously "ill-conditioned" problem is characterized by the fact that the smallest eigenvalue λ_m is very much smaller than 1 . They have summarized the dramatic inadequacy of LS for nonorthogonal problems in this situation by noting that σ^2/λ_m is a lower bound for the average squared distance between $\hat{\beta}$ and β .

Thus, for ill-conditioned data, the LS estimated regression coefficient vector $\hat{\beta}$ is expected to be far distant from the true vector β . Moreover, the LS coefficient vector is much too long, on the average, since $\lambda_m \ll 1$. The LS solution yields coefficients whose absolute values are too large and whose signs may actually reverse

with negligible changes in the data. The ridge regression estimator is obtained by solving

$$(X'X + dI)\hat{\beta}^{*} = X'y$$

yielding

$$\hat{\beta}^{*} = (X'X + dI)^{-1}X'y$$

for $d \geq 0$.

About the form of the estimator, Hoerl and Kennard says:

> "Estimation based on the matrix $[X'X + dI]$, $d \geq 0$ rather than $X'X$ has been found to be a procedure that can be used to help circumvent many of the difficulties associated with the usual LS estimates. In particular, the procedure can be used to portray the sensitivity of the estimates to the particular set of data being used, and it can be used to obtain a point estimate with smaller mean squared error."

There is an "optimum" value for d for any problem which cannot be determined in practice, so it is desirable to examine the ridge solution for a range of admissible values of d . They introduce the term "Ridge Trace" to describe the solution thought of, focused as a function of d . They also discuss the methods for choosing d . Some critical comments on the properties and goodness of the method had been given by Conniffee and Stone [4] and an extension and detailed study of the properties of the Ridge estimator has been done by Chapman [2].

Marquardt [13] proposed a class of regression estimators he called generalized inverse estimators. He argues that:

> "It is important to recognize that practical estimation problems give rise to matrices $X'X$ having eigenvalues that may be grouped qualitatively into three types - substantially greater than zero, slightly greater than zero, precisely zero (except for rounding error). In computations, it may be difficult to distinguish between adjacent types."

As an alternative to using precise methods to obtain "exact" solutions for this case, he suggests assuming a lower rank for $X'X$ and obtaining a solution under this assumption. For this he says:

> "Use of multiple precision arithmetic will not guarantee reliable results. Furthermore, inspection of the eigenvalues spectrum usually suggests that there is no "rank" clearly assignable to the matrix. Rather, there is a range of ranks that may be reasonable choices. One would like to be able to determine the generalized inverse for any assigned rank in this reasonable range."

The class of generalized inverse regression estimators is given by:

$$\hat{\beta}^{+} = A_{\gamma}^{+} X'y \ ,$$

where A_{γ}^{+} is the generalized inverse based upon

$$A_{\gamma}^{+} = \sum_{i=1}^{\gamma} \frac{1}{\lambda_i} V_i V_i' \qquad (3.5)$$

and the V_i is the eigenvector of $X'X$ corresponding to λ_i, for assigned rank γ. In general, there is an "optimum" value for γ for any problem, but it is desirable to examine the generalized inverse solution for a range of admissible values for γ.

Marquardt has shown that if $\hat{\beta}^{+}$ is the solution of the normal equations $X'X\hat{\beta} = X'y$ obtained by assinging rank γ to $X'X$, and

using the generalized inverse (3.5), then $\hat{\beta}^{+}$ minimizes the sum of squares of residuals

$$0(\hat{\beta}) = (y - X\hat{\beta})'(y - X\hat{\beta}) .$$

Marquardt makes a distinction between ridge and generalized inverse estimates. Precisely, he says:

> "Although the ridge and generalized inverse estimators share many desirable properties, the ridge estimator is not a generalized inverse estimator. For example,
>
> $$AA^{+}A = A .$$
>
> The ridge can be viewed as an approximate generalized inverse."

Rao [16] makes the remark that not much work has been done on BLE (minimum MSE estimators) compared to that on BLUE. Restricting attention to the model proposed in (2.1) and letting $\ell'y$ be an estimator of $p'\beta$, he points out that the MSE of $\ell'y$, namely,

$$E(\ell'y - p'\beta)^2 = \sigma^2\ell'\Sigma\ell + (X'\ell - p)'\beta\beta'(X'\ell - p) \qquad (3.6)$$

which involves both the unknown parameters σ^2 and β is not a suitable criterion for minimizing.

He proposes the following possibilities:

(1) Choose an a priori value of $\sigma^{-1}\beta$, say b, based on previous knowledge, and set up the criterion as $\sigma^2 S$, where

$$S_1 = \ell'\Sigma\ell + (X'\ell - p)'W(X'\ell - p)$$

and $W = bb'$.

(ii) If β is considered to have an a priori distribution with a dispersion matrix $\sigma^2 W$, where W is known, then the criterion is $\sigma^2 S_1$.

(iii) We observe that the expression (3.6) is the sum of two parts, one representing the variance and the second the bias. In such a case the choice of W in S_1 represents the relative weight we attach to bias compared to variance.

Taking S_1 as a criterion with an appropriate symmetric W, Rao gives the optimum choice of ℓ in the form of the following theorem.

Theorem 3.4

The BLE of any function $p'\beta$ is $p'\tilde{\beta}$, where

$$\tilde{\beta} = WX'(\Sigma + XWX')^{-1}y .$$

He later points out:

> "If we have some knowledge about the domain in which β is expected to lie, we may be able to choose W suitably to assure that BLE's have uniformly smaller mean dispersion error than BLUE's."

He also remarks that:

> "Further investigation in this direction such as comparison of the estimators given in Theorem 3.4 with the ridge estimator of Hoerl and Kennard will be useful."

Mayer and Willke [14] have studied the ridge estimators, and they consider them as a subclass of the class of linear transformation of

LS estimators. At the same time, they propose an alternative class of estimators that they call shrunken estimators. Mayer and Willke have shown that these estimators satisfy the admissibility condition proposed by Hoerl and Kennard and both are derived as minimum norm estimators in the class of linear transformations of LS estimators. In addition, they obtain a class of estimators that are minimum variance linear transformations of the LS estimators and that the members of this class are shown to be stochastically shrunken estimators.

The first class that they propose has as a typical member

$$C_\lambda = \lambda\hat{\beta} \ , \ \lambda \in [0,\infty)$$

where $\hat{\beta}$ is the OLS estimator and λ is labelled the shrinkage factor. If λ is a fixed scalar, then C_λ is called a deterministically shrunken; alternatively if $\lambda = F(\hat{\beta}'\hat{\beta})$ is a scalar function of $\hat{\beta}'\hat{\beta}$, then C_λ is called a stochastically shrunken estimator and is written $C(f)$.

Although the shrunken estimator, C_λ, may seem a rather simplistic alteration, Mayer and Willke has shown that it satisfies the following admissibility condition:

Proposition 3.1

For every β there exists a fixed λ in $[0,1]$ such that

$$E(C_\lambda - \beta)'(C_\lambda - \beta) < \mathrm{Var}(\hat{\beta})$$

and, thus, the subclass of deterministically shrunken estimators is admissible.

The second class of estimators that they propose are the ones that belongs to the following class:

Definition 3.2

Let C denote the class of linear transformations of $\hat{\beta}$. If $b \in C$, then $b = A\hat{\beta}$ for some $(m \times m)$ matrix A.

Definition 3.3

Let $C(\tau)$ denote the subclass of C such that $b(A_0)$ is in $C(\tau)$, if and only if,

$$\hat{\beta}'(A_0 - I)'S(A_0 - I)\hat{\beta} = \tau .$$

Thus, $C(\tau)$ is actually an equivalence class or orbit within the class C, the equivalence defined with respect to the sum of squares loss function.

Mayer and Willke show that both the ridge and the deterministically shrunken estimators can be characterized as minimum norm estimators in the class $C(\tau)$. In addition, they discuss some methods for choosing the proper shrinkage factor.

The third class of estimators that they proposed arises from the fact that different norms lead to different estimators and there is no obvious reason for preferring one norm over another. They consider estimators which have minimum total variance among all estimators in the given class. These estimators have the form

$$d_\delta = \delta\hat{\beta}'\hat{\beta}(I + \delta\hat{\beta}\hat{\beta}')^{-1}\beta , \quad \delta \in [0,\infty)$$

which are minimum variance within each equivalence class. Although d_δ looks quite complex, Mayer and Willke have shown that the estimators belongs to the class of shrunken estimators, in fact, d_δ is a stochastically shrunken estimator.

Finally, they propose that, if we use a shrinkage estimator with

$$[1 + \xi s^2(\hat{\beta}'\hat{\beta})^{-1}]$$

as the shrinkage factor, where $s^2 = y'y - \hat{\beta}'s\hat{\beta}$, then the class of estimators

$$e_\xi = [1 + \xi s^2(\hat{\beta}'\hat{\beta})^{-1}]\hat{\beta}$$

satisfies a stronger admissibility condition than the one presented above. In particular, if we let

$$W(B) = E(B - \beta)'S(B - \beta)$$

denote the weighted total MSE of the estimator B, then the following proposition is given by Sclove [19] and is based on results of James and Stein [10].

Proposition 3.2

If $m \geq 3$ and $0 < \xi < 2(m-2)(n-m+2)^{-1}$, then $W(e_\xi) < W(\hat{\beta})$ and if $\xi_0 = (m-2)(n-m+2)^{-1}$, then

$$W(e_{\xi_0}) = \min_{\xi} W(e_\xi) .$$

The class e_ξ is strongly admissible with respect to weighted MSE in the sense that is is known exactly which elements are better (in terms of MSE) than the LS estimator.

Goldstein and Smith [7] examine the mean square error properties of a class of shrinkage estimators for the normal regression model $(y, X\beta, I\sigma^2)$ which leads to a new derivation of the Hoerl and Kennard [9] ridge estimator and its generalizations; they also compare the proposed class with the James and Stein [10] estimator and the estimator proposed by Marquardt [13].

Goldstein and Smith decompose the model

$$y = U\Lambda^{1/2}V'\beta + e$$

where U, $\Lambda^{1/2}$ and V' are given in (2.4), as follows:

Let $P = \begin{bmatrix} U \\ \tilde{U} \end{bmatrix}$ where $\tilde{U}$ is a $(n \times n-m)$ matrix, the orthogonal complement of U, be such that

$$PXV = D = \begin{bmatrix} \Lambda^{1/2} \\ 0 \end{bmatrix}.$$

Writing

$$Z = Py, \quad \delta = V'\beta, \quad v = Pe,$$

they obtain

$$Z = PX\beta + Pe = D\delta + v = \begin{bmatrix} \Lambda^{1/2}V'\beta \\ 0 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

where v is normally distributed with mean 0 and covariance matrix $I\sigma^2$. Explicity

$$\left.\begin{aligned} Z_i &\sim N(\lambda_i^{1/2}\delta_i; \sigma^2) \quad (i = 1, 2, \ldots, m) \\ Z_i &\sim N(0, \sigma^2) \qquad\quad (i = m+1, \ldots, n) \end{aligned}\right\} . \tag{3.7}$$

They confine attention to estimators of the form

$$\delta_i^* = c_i Z_i \ ,$$

with

$$|c_i Z_i| \leq |Z_i/\lambda_i^{1/2}|$$

where $Z_i/\lambda_i^{1/2}$ are the OLS estimators.

The requirements that they impose on c_i are summarized as follows:

$$\left.\begin{array}{ll} \text{(i)} & c(\lambda_i,0) = \dfrac{1}{\lambda_i^{1/2}} \\ \text{(ii)} & c'(\lambda_i,d)/\lambda_i^{1/2} < 0 \quad \text{for all} \quad d \geq 0 \end{array}\right\} \ . \qquad (3.8)$$

Assuming from now on that (3.8) is satisfied, Goldstein and Smith have proven the following lemma:

Lemma 3.1

For any δ there exists $d > 0$ such that $\delta_i^* = c(\lambda_i,d)Z_i$ has smaller MSE than

$$\hat{\beta}_i = \frac{Z_i}{\lambda_i^{1/2}}$$

for all $i = 1, 2, \ldots, m$.

Later on they made some comments about the class of shrinkage estimators proposed by James and Stein [10]. These estimates are based on first line of (3.7) and take the form:

$$\tilde{\delta}[1 - ((m-2)/\sum_i^m Z_j^2)]Z_i \qquad i = 1, \ldots, m \ . \qquad (3.9)$$

They said about (3.9):

> "With respect to quadratic loss for $\tilde{\delta}_i$ it is well known that this has smaller average MSE than the LS estimator $\hat{\beta}_i = Z_i$. We note however that $\delta_i^* = \tilde{\delta}_i/\lambda_i^{1/2}$ the corresponding estimator for δ_i is now explicitly derived with respect to the loss function
>
> $$\sum_{i=1}^{m} \lambda_i(\delta^* - \delta_i)^2$$
>
> corresponding to
>
> $$\sum_{i=1}^{m} (\tilde{\delta}_i - \delta_i)^2$$
>
> rather than with respect to
>
> $$\sum_{i=1}^{m} (\delta^* - \delta_i)^2 .$$
>
> On an intuitive level, we see, therefore, that the James-Stein form is inappropriate in the sense that it implicity takes less account of the loss in precisely those directions where estimation is most inaccurate."

They have shown too that in the original parameters in the model $(y, X\beta, I\sigma^2)$, for any β there exists a $d > 0$ such that for $\beta^* = V\delta^*$, $\delta_i^* = c(\lambda_i, d)Z_i$. We will have that, β_i^* has smaller MSE than the corresponding LS estimator $\hat{\beta}_i$ for all $i = 1, \ldots, m$.

Hoerl and Kennard's ridge regression estimates and Marquardt's estimators belong to the class of estimators defined by β^* since by letting $c(\lambda_i, d) = \lambda_i^{1/2}/\lambda_i + d$ we get the ridge estimator and letting $\delta_i^* = Z_i/\lambda_i$ for $i = 1, \ldots, \gamma$, $\delta_i^* = 0$, $i = \gamma+1, \ldots, m$ we get Marquardt's estimator.

4. DIRECTIONALLY MINIMAX TRACE MEAN SQUARED ERROR ESTIMATOR

The goal in this chapter is to obtain an estimator of β which depends as little as possible on the parameters themselves and which, in some sense, minimizes the trace mean squared error, $T(Ay+b,\beta)$. Adopting a minimax philosophy leads to attempting to replace β by the value which maximizes $T(Ay+b,\beta)$. This, of course, is not helpful as T is unboundedly increasing with $||\beta||$. The modification adopted in this work is to express the parameters in the form $\beta = k\alpha$, where α are its direction cosines and k its length. Then, for fixed values of k, the expression, $T(Ay+b,k\alpha)$, can be maximized by choice of α. By Theorem 4.1 and the discussion which follows it, the ultimate choice of α is independent of the value of k. This fact indicates that, emanating from the origin, there is a direction corresponding to the worst choice of β with respect to mean squared error (or equivalently, squared bias). It is from along this ray that the value of β is chosen to maximize T. The exact location on the ray is set by choice of k.

The ideas expressed above are important enough to warrant the following definition.

Definition 4.1

A linear estimator $\tilde{\beta} = Ay + b$ is said to be the Directionally Minimax Trace Mean Squared Error (DMTMSE) Estimator of β if A and b are such that they minimize

$$S_T(A,b,k) = \sup_{\beta\in\mathcal{B}_k} T(Ay+b,\beta)$$

$$= \sigma^2 tr(A\Sigma A') + \sup_{\beta\in\mathcal{B}_k} [b + (AX-I)\beta]'[b + (AX-I)\beta]$$

where

$$\mathcal{B}_k = \{\beta \mid \beta = k\alpha \ , \ \|\alpha\| = 1\} \ .$$

In order to simplify our work, so that we can accomplish our goal, we will first prove the following theorem.

Theorem 4.1

In the linear model $(y,X\beta,\Sigma\sigma^2)$, $S_T(A,b,k) \geq S_T(A,0,k)$ for any A , and in particular, the DMTMSE estimator is of the form Ay .

Proof: Either $\|F\beta + b\|^2 \leq \|F\beta\|^2$ or $\|F\beta + b\|^2 \geq \|F\beta\|^2$. Suppose the former holds. Then it follows that $\|b\|^2 \leq -2b'F\beta$ and hence,

$$\|-F\beta + b\|^2 = \|F\beta\|^2 + \|b\|^2 - 2b'F\beta \geq \|F\beta\|^2$$

$$+ 2\|b\|^2 \geq \|F\beta\|^2 \ .$$

Thus,

$$\max(\|F\beta + b\|^2 \ , \ \|-F\beta + b\|^2) \geq \|F\beta\|^2 \ .$$

Now,

$$\sup_{\beta\in\mathcal{B}_k} \|F\beta + b\|^2 = \sup_{\beta\in\mathcal{B}_k} \{\max[\|F\beta + b\|^2 \ , \ \|-F\beta + b\|^2]\}$$

$$\geq \sup_{\beta\in\mathcal{B}_k} \|F\beta\|^2 \ ,$$

with equality when $b = 0$. Therefore, $S_T(A,b,k) \geq S_T(A,0,k)$, for all A and k . □

In view of Theorem 4.1, the criterion can be reduced to:

$$S_T(A,k) = \sigma^2 tr(A\Sigma A') + \sup_{\beta \in \mathcal{Q}_k} \beta'F'F\beta$$

$$= \sigma^2 tr(A\Sigma A') + k^2 Ch_M(F'F) \quad , \qquad (4.1)$$

where $F = AX-I$. It can be seen from (4.1) that the characteristic vector of $F'F$ associated with its largest characteristic root gives the direction cosines of the least favorable direction for estimation. This is the value of α which leads to maximum bias. Fortunately, this direction is independent of k so it is meaningful to hold k constant.

4.1 DMTMSE Estimation In The Class of OLS Estimators Computed Subject To False Restrictions

One class of biased linear estimators of regression coefficients that has been proposed is that obtained by computing least squares estimates under sets of false restrictions. These estimators were studied by Toro and Wallace [18]. Their work is primarily concerned with testing whether or not a particular set of false restrictions will lead to estimators with smaller mean squared error. They leave the choice of restrictions up to the experimenter.

For this case, we shall restrict our attention to the linear model $(y,X\beta,I\sigma^2)$ and assume that X is of full rank $(rank(X) = m)$.

To obtain the class of estimators, we shall impose u false restrictions on the parameters:

$$R\beta = h \ ,$$

where R is a $(u \times m)$ matrix of rank $u(\leq m)$ and h is a $(u \times 1)$ vector. So that the equations are consistent, we shall assume that $h \in \mathcal{C}(R)$.

It is well-known (see, for example, Pringle and Rayner [15]) that under this setup the estimator is of the form $\tilde{\beta} = \tilde{\beta}(R,h)$

$$\tilde{\beta} = \tilde{\beta}(R,h) = [I - S^{-1}R'(RS^{-1}R')^{-1}R]S^{-1}X'y$$

$$+ S^{-1}R'(RS^{-1}R')^{-1}h \ ,$$

where S is given by (2.2).

In the following series of theorems, we will see that it is possible to limit the class of estimators of interest.

Theorem 4.1.1

In the class of least squares estimators subject to false restrictions for the setup $(y, X\beta, I\sigma^2)$, restrictions $R\beta = 0$ are preferred over $R\beta = h$ with respect to DMTMSE estimation.

Proof: The theorem follows from Theorem 4.1 since, for $R\beta = h$, the estimator will be of the form $Ay + b$, where A is independent of h and b equals 0 whenever $h = 0$. □

In view of Theorem 4.1.1, attention will be limited to the class of least squares estimators obtained subject to $R\beta = 0$.

Theorem 4.1.2

The class of estimators, $\tilde{\beta}(R)$, obtained by least squares subject to constraints $R\beta = 0$, where $R \in \{R \mid R \text{ is } (u \times m) \text{ of rank } u\}$ is equivalent to that where $R \in \{R \mid R \text{ is } (u \times m) \text{ of rank } u \text{ and } RS^{-1}R' = I\}$, where S is given by (2.2).

Proof: Since Λ and V given in (2.4) are positive definite, the rows of $\Lambda^{1/2}V$ form a basis for Euclidean m-space and, hence, R can be written as $R = B\Lambda^{1/2}V'$, for some $(u \times m)$ matrix B. Since $u = \text{rank}(R) \leq \text{rank}(B) \leq u$, B must have full row rank. Now,

$$RS^{-1}R' = B\Lambda^{1/2}V'V\Lambda^{-1}V'V\Lambda^{1/2}B' = BB'$$

and BB' is positive definite. Thus, if we let $RS^{-1}R' = GG'$, $|G| \neq 0$, it is easily seen that $\tilde{\beta}(R) = \tilde{\beta}(G^{-1}R)$. Also, if we let $G^{-1}R = \tilde{R}$, we have

$$\tilde{R}S^{-1}\tilde{R}' = G^{-1}GG'G'^{-1} = I .$$

Thus, for every R there exists a corresponding $\tilde{R}$ such that

$$\tilde{\beta}(R) = \tilde{\beta}(\tilde{R}) \quad \text{and} \quad \tilde{R}S^{-1}\tilde{R}' = I . \quad \square$$

Using Theorem 4.1.2, we can respecify the class of estimators of interest as

$$\tilde{\beta} = \tilde{\beta}(R) = (I - S^{-1}R'R)S^{-1}X'y$$

for all R such that $RS^{-1}R' = I$. Within this class we wish to find the optimum with respect to the DMTMSE criterion set out in Chapter 4.

It should be noted that the value of u is assumed to be given and fixed. Also, because of the assumption that R is of full row rank, the ordinary least squares (OLS) estimator is not in the class. However, there exist members of the class that are arbitrarily "close" to the OLS estimator.

Subsections 4.1.1 and 4.1.2 below deal with deriving the optimum estimator for $u = 1$ and for general u, respectively, and subsection 4.1.3 examines some properties of the resulting estimators.

4.1.1 The Single Constraint Case

The first case we shall consider is when $u = 1$. In this case the restrictions being imposed are of the form $r'\beta = 0$, where r is an $(m \times 1)$ vector satisfying $r'S^{-1}r = 1$.

By putting $A = (I - S^{-1}rr')S^{-1}X'$ in (4.1), and since

$$Ch_M(F'F) = Ch_M(rr'S^{-2}rr') = r'rr'S^{-2}r ,$$

we get

$$S_T(r,k) = \sigma^2 tr[(I - S^{-1}rr')S^{-1}X'XS^{-1}(I - rr'S^{-1})]$$

$$+ k^2 r'rr'S^{-2}r$$

$$= \sigma^2 tr[S^{-1} - S^{-1}rr'S^{-1} - S^{-1}rr'S^{-1} + S^{-1}rr'S^{-1}rr'S^{-1}]$$

$$+ k^2 r'rr'S^{-2}r$$

$$= \sigma^2 \mathrm{tr}[S^{-1}] + (k^2 r'r - \sigma^2) r'S^{-2}r$$

$$= \sigma^2 \sum_{i=1}^{m} \lambda_i^{-1} + (k^2 r'r - \sigma^2) r'S^{-2}r \ ,$$

where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$ and Λ is defined in (2.4) and (2.5).

The following theorem gives the vector r which minimizes $S_T(r,k)$.

Theorem 4.1.1.1

For the class of least squares estimates computed subject to one incorrect restriction,

$$S_T(r,k) \geq \sigma^2 \sum_{i=1}^{m-1} \lambda_i^{-1} + k^2$$

with equality when $r = \lambda_m^{1/2} v_m$. Here $V = (v_1, v_2, \ldots, v_m)$ and V is defined in (2.4). That is, the DMTMSE estimator for this class of estimators is

$$\tilde{\beta} = (I - v_m v_m')S^{-1}X'y \ .$$

Proof: Let $w = \Lambda^{-1/2}V'r$ so that $r = V\Lambda^{1/2}w$. Also, $r'S^{-1}r = 1$ implies $w'w = 1$. Now

$$S_T(r,k) = S_T(V\Lambda^{1/2}w,k) = \sigma^2 \sum_{i=1}^{m} (w_i^2\lambda_i - \sigma^2) \sum_{i=1}^{m} w_i^2\lambda_i^{-1} + \sigma^2 \sum_{i=1}^{m} \lambda_i^{-1}$$

$$= \sigma^2 \sum_{i=1}^{m} \lambda_i^{-1} + k^2 \Big(\sum_{i=1}^{m} w_i^2\lambda_i \Big) \Big(\sum_{i=1}^{m} w_i^2\lambda_i^{-1} \Big)$$

$$- \sigma^2 \sum_{i=1}^{m} w_i^2\lambda_i^{-1} \ . \qquad (4.1.1.1)$$

Since

$$\sum_{i=1}^{m} w_i^2 = 1$$

and since the harmonic mean is always less than or equal to the arithmetic mean,

$$\left(\sum_{i=1}^{m} w_i^2\lambda_i\right)\left(\sum_{i=1}^{m} w_i^2\lambda_i^{-1}\right) \geq 1 . \qquad (4.1.1.2)$$

Furthermore, since the value of a weighted arithmetic mean is always less than or equal to the largest value being averaged,

$$\sum_{i=1}^{m} w_i^2\lambda_i^{-1} \leq \lambda_m^{-1} . \qquad (4.1.1.3)$$

Using (4.1.1.2) and (4.1.1.3) in (4.1.1.1), we get

$$S_T(r,k) \geq \sigma^2 \sum_{i=1}^{m} \lambda_i^{-1} + k^2 - \sigma^2 \sum_{i=1}^{m} w_i^2\lambda_i^{-1}$$

$$\geq \sigma^2 \sum_{i=1}^{m} \lambda_i^{-1} + k^2 - \sigma^2\lambda_m^{-1}$$

$$= \sigma^2 \sum_{i=1}^{m-1} \lambda_i^{-1} + k^2 .$$

Noting that equality holds in (4.1.1.2) whenever $\lambda_i = \lambda$ for all values of i for which $w_i > 0$ and in (4.1.1.3) whenever $w_1 = w_2 = \ldots = w_{m-1} = 0$, $w_m = 1$, we see that both equalities hold when $w' = (0, 0, \ldots, 0, 1)$, in which case $r = V\Lambda^{1/2}w = \lambda_m^{1/2}v_m$.

Thus,

$$S_T(\lambda_m^{1/2}v_m,k) = \sigma^2 \sum_{i=1}^{m-1} \lambda_i^{-1} + k^2$$

and the resulting estimator becomes

$$\tilde{\beta} = (I - v_m v_m')S^{-1}X'y \ . \ \square$$

The estimator thus derived can be seen to be of the form $\tilde{\beta} = P\hat{\beta}$, where P is an orthogonal projection matrix and $\hat{\beta}$ is the OLS estimator. The estimator can be obtained by projecting $\hat{\beta}$ into the space orthogonal to the characteristic vector corresponding to the smallest characteristic root of $X'X$.

Another interesting feature of this result is the fact that the estimator does not depend on the value of k and, hence, need not be specified beforehand. Of course, the resulting mean squared errors will involve k .

4.1.2 The General Case

For the general case we shall assume that u is a fixed number $(1 \leq u \leq m)$. The solution has the appearance of that for $u = 1$ but the proof of the theorem is more involved.

Theorem 4.1.2.1

For the class of least squares estimates computed subject to a set of u independently incorrect restrictions,

$$S_T(R,k) \geq \sigma^2 \sum_{i=1}^{m-u} \lambda_i^{-1} + k^2$$

with equality holding whenever $R = \Lambda_2^{1/2} V_2'$, where Λ_2 and V_2 are given in (2.7) and (2.8). That is, for this class of estimators, the DMTMSE estimator is

$$\tilde{\beta} = (I - V_2 V_2')S^{-1}X'y = V_1 V_1' S^{-1} X'y \ .$$

Proof: We can write

$$R = (BB')^{-1/2} B \Lambda^{1/2} V'$$

where B is any $(u \times m)$ matrix of rank u. From this we have that

$$\begin{aligned} & tr[S^{-1} - S^{-1}R'RS^{-1}] \\ & = tr[V\Lambda^{-1}V' - V\Lambda^{-1}V'V\Lambda^{1/2}B'(BB')^{-1/2}(BB')^{-1/2}B\Lambda^{1/2}V'V\Lambda^{-1}V'] \\ & = tr[\Lambda^{-1}] - tr[\Lambda^{-1/2}P_B\Lambda^{-1/2}] \\ & = \sum_{i=1}^{m} \lambda_i^{-1} - \sum_{i=1}^{m} p_{ii}\lambda_i^{-1} \ , \end{aligned}$$

where

$$P_B = B'(BB')^{-1}B = (p_{ij}) \ .$$

Since P_B is a symmetric idempotent matrix, we have that

$$\sum_{j=1}^{m} p_{ij}^2 = p_{ii}$$

which implies that $p_{ii} \geq 0$. Further, if $p_{ii} > 0$, we have that

$$0 = \sum_{j \neq i} p_{ij}^2 + p_{ii}^2 - p_{ii} \geq p_{ii}^2 - p_{ii}$$

implies $p_{ii} \leq 1$. It is well-known that

$$\sum_{i=1}^{m} p_{ii} = u \ .$$

Therefore,

$$\sum_{i=1}^{m} p_{ii}\lambda_i^{-1} \ ,$$

is maximized subject to $0 \leq p_{ii} \leq 1$ and

$$\sum_{i=1}^{m} p_{ii} = u$$

by choosing

$$p_{11} = \cdots = p_{(m-u)(m-u)} = 0$$

and

$$p_{(m=u+1)(m-u+1)} = \cdots = p_m = 1$$

giving

$$\sum_{i=1}^{m} p_{ii}\lambda_i^{-1} = \sum_{i=m-u+1}^{m} \lambda_i^{-1} \ .$$

This corresponds to taking $B = [0 \vdots I_u]$ and

$$P_B = \begin{bmatrix} 0 & \vdots & 0 \\ \cdots & \vdots & \cdots \\ 0 & \vdots & I_u \end{bmatrix} \ .$$

Using this we have

$$\mathrm{tr}[S^{-1} - S^{-1}R'RS^{-1}] \geq \sum_{i=1}^{m-u} \lambda_i^{-1} \ . \qquad (4.1.2.1)$$

Next, we have that

$$Ch_M(R'RS^{-1}R'R) = Ch_M(V\Lambda^{1/2}P_B\Lambda^{-1/2}\Lambda^{-1/2}P_B\Lambda^{1/2}V')$$

$$= Ch_M(\Lambda^{1/2}P_B\Lambda^{-1/2}\Lambda^{-1/2}P_B\Lambda^{1/2})$$

$$= Ch_M(Q'Q) ,$$

where

$$Q = \Lambda^{1/2}P_B\Lambda^{-1/2} .$$

Note that $Q^2 = Q$ but is not symmetric. Now suppose that $\ell \in \mathcal{C}(Q)$ then $\ell = Qz$ for some z and, hence, $Q\ell = \ell$. Furthermore, if $\ell^* = \ell/(\ell'\ell)^{1/2}$ then $Q\ell^* = \ell^*$ and $\|Q\ell^*\|^2 = \|\ell^*\|^2 = 1$. Therefore,

$$Ch_M(Q'Q) = \sup_{x(x'x=1)} \|Qx\|^2 \geq 1$$

and

$$Ch_M(R'RS^{-1}R'R) \geq 1 . \qquad (4.1.2.2)$$

Using (4.1.2.1) and (4.1.2.2) we may conclude that

$$\sigma^2 tr(S^{-1} - S^{-1}R'RS^{-1}) + k^2 Ch_M(R'RS^{-2}R'R)$$

$$\geq \sigma^2 \sum_{i=1}^{m-u} \lambda_i^{-1} + k^2 . \qquad (4.1.2.3)$$

It remains to demonstrate that the lower bound is attainable. Putting $B = (0 \vdots I_u)$ we get

$$R = (BB')^{-1/2}B\Lambda^{1/2}V' = \Lambda_2^{1/2}V_2'$$

and substituting this into the left hand side of (4.1.2.3) gives

$$\sigma^2 tr(V_1\Lambda_1^{-1}V_1') + k^2 Ch_M(V_2V_2')$$

$$= \sigma^2 \sum_{i=1}^{m-u} \lambda_i^{-1} + k^2$$

which is the right hand side of (4.1.2.3). Finally, the estimator can take various forms:

$$\tilde{\beta} = (I - V_2V_2')S^{-1}X'y = (I - V_2V_2')\hat{\beta} = V_1V_1'\hat{\beta} ,$$

where $\hat{\beta}$ is the OLS estimator. □

4.1.3 Some Properties of The Estimator

Marquardt [13] proposed a class of regression estimators he called generalized inverse estimators. He argued that, as an alternative to using precise computing methods for obtaining "exact" solutions in situations where the $X'X$ matrix has some small but positive characteristic roots, it might be preferable to assign a lower rank to $X'X$ and obtain a solution under this assumption. In our notation, his estimator can be written as

$$\hat{\beta}^+ = A_v^+X'y ,$$

where v is the assigned rank of $X'X$, $A_v^+ = V_1\Lambda_1^{-1}V_1'$, and V_1 and Λ_1 are defined in (2.7) and (2.8), with the exception that the dimensions of V_1 is $(m \times v)$. In this setting the first of a

series of properties that the estimator possesses is introduced in the following theorem.

Theorem 4.1.3.1

The DMTMSE estimator for the class of least squares estimators computed subject to false restrictions is equivalent to Marquardt's generalized inverse estimators when the assigned rank of $X'X$ equals $m - u$.

Proof: Since we are assuming that $X'X$ is of full rank,

$$\hat{\beta}^{+} = A_v^{+}X'y = V_1\Lambda_1^{-1}V_1'(V_1\Lambda_1V_1' + V_2\Lambda_2V_2')S^{-1}X'y$$

$$= V_1V_1'S^{-1}X'y = \tilde{\beta} \ . \ \square$$

Some further results which follow easily from previous results are summarized without proof in the following theorem.

Theorem 4.1.3.2

For any $(m \times 1)$ vector ℓ

(i) $\mathrm{Var}(\ell'\tilde{\beta}) \leq \mathrm{Var}(\ell'\hat{\beta})$ with equality if, and only if, $\ell \in \mathcal{C}(V_1)$ in which case $\ell'\tilde{\beta} = \ell'\hat{\beta}$.

(ii) $E(\ell'\tilde{\beta}) = \ell'\beta + \ell'V_2V_2'\beta$ and the second term (the bias) vanishes if, and only if, $\ell \in \mathcal{C}(V_1)$ or $\beta \in \mathcal{C}(V_1)$.

(iii) The estimator $\tilde{\beta}$ is shorter than $\hat{\beta}$, i.e., $\tilde{\beta}'\tilde{\beta} < \hat{\beta}'\hat{\beta}$.

(iv) For the more general model $(y, X\beta, \Sigma\sigma^2)$, where $|\Sigma| \neq 0$, the corresponding estimator is of the form

$$\tilde{\beta} = \tilde{V}_1\tilde{V}_1'(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y ,$$

where $X'\Sigma^{-1}X = \tilde{V}\tilde{\Lambda}\tilde{V}$, $\tilde{V}\tilde{V}' = \tilde{V}'\tilde{V} = I$, $\tilde{\Lambda}$ is diagonal, $\tilde{V} = (\tilde{V}_1 : \tilde{V}_2)$, and $\tilde{V}_1$ is a $(m \times (m-u))$ matrix.

Toro and Wallace [18] give a statistical test for determining whether imposing a given set of false restrictions, $R\beta = 0$, will result in a reduced mean square error. This test is appropriate for use in this setting provided we make the assumption that the errors in the model are normally distributed. In this case, the form of the test is

$$\text{Reject } H_0 \text{ if } W \geq W_\alpha ,$$

where

$$W = [y'U_2U_2'y/u]/[y'(I - UU')y/(n-m)]$$

and U and U_2 are defined in (2.4) and (2.6). Note that the hypothesis H_0 states that $\tilde{\beta}$ is superior to the OLS estimator, $\hat{\beta}$, with respect to mean squared error and accepting H_0 suggests that $\tilde{\beta}$ is preferred over $\hat{\beta}$. It can be seen that W is a noncentral F random variable with u and $(n - m)$ degrees of freedom and noncentrality $\delta = \beta'V_2\Lambda_2V_2'\beta/2\sigma^2$, where V_2 and Λ_2 are defined in (2.7) and (2.8). It can be seen also that

$$\delta \leq \beta' V_2 V_2' \beta \lambda_{m-u+1}/2\sigma^2 ,$$

where λ_{m-u+1} is defined in (2.4) and (2.4). Therefore, δ is small whenever the projection of β on $\mathcal{C}(V_2)$ is small or λ_{m-u+1} is small. The latter condition is precisely that for which the generalized inverse estimators are intended. Some critical points and power computations are given in [18].

4.2 The DMTMSE Estimator For The Class of Shrinkage Estimators

Many regression estimators have been proposed that have the form

$$\tilde{\beta} = Ay = V\Gamma U'y ,$$

where U and V are defined by (2.4) and

$$\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_m) .$$

Included in this class are:

OLS: $\gamma_j = \lambda_j^{-1/2}$, $j = 1, \ldots, m$

Generalized Inverse Regression: $\gamma_j = \lambda_j^{-1/2}$, $j = 1, \ldots, m-u$

$\gamma_j = 0$, $j = m-u+1, \ldots, m$

Ridge Regression: $\gamma_j = \lambda_j^{1/2}/(\lambda_j + k^2)$, $j = 1, \ldots, m$

as well as others. This class is discussed by Goldstein and Smith [7].

In the following theorem, the DMTMSE estimator for this class is derived.

Theorem 4.2.1

In the linear model $(y, X\beta, I\sigma^2)$, the DMTMSE estimator from the class of estimators of the form $\tilde{\beta} = V\Gamma U'y$, where U and V are defined by (2.4) and $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_m)$, is $\tilde{\beta} = t\hat{\beta}$, where $\hat{\beta}$ is the OLS estimator,

$$t = c^2/(\sum_{j=1}^{m} \lambda_j^{-1} + c^2) ,$$

and $c^2 = k^2/\sigma^2$.

Proof: First,

$$S_T(A,k) = \sigma^2 tr(AA') + k^2 Ch_M[(I-AX)(I-X'A')]$$

$$= \sigma^2 \sum_{j=1}^{m} \gamma_j^2 + k^2 Ch_M(V(I - \Lambda^{1/2}\Gamma)^2 V')$$

$$= \sigma^2 \sum_{j=1}^{m} \gamma_j^2 + k^2 \max[(1 - \lambda_j^{1/2}\gamma_j)^2, \; j = 1, 2, \ldots, m] .$$

Suppose that the value of the second term is p^2. Then any set of γ_j's that minimizes S_T must satisfy

$$(1 - \lambda_j^{1/2}\gamma_j)^2 \leq p^2$$

for $j = 1, \ldots, m$ and, hence,

$$\gamma_j \geq (1-p)\lambda_j^{-1/2}, \quad j = 1, \ldots, m .$$

But S_T is a minimum, therefore, γ_j must equal $(1-p)\lambda_j^{-1/2}$, its smallest possible value, so as to make

$$\sum_{j=1}^{m} \gamma_j^2$$

as small as possible. This implies that

$$S_T/\sigma^2 = (1-p)^2 \sum_{j=1}^{m} \lambda_j^{-1} + p^2 c^2 .$$

Since this is concave in p, the minimum can be found by setting first derivatives equal to zero resulting in

$$p = \sum_{j=1}^{m} \lambda_j^{-1} / \left(\sum_{j=1}^{m} \lambda_j^{-1} + c^2 \right) .$$

Thus,

$$\gamma_j = \left[c^2 / \left(\sum_{i=1}^{m} \lambda_i^{-1} + c^2 \right) \right] \lambda_j^{-1/2}$$

where $c^2 = k^2/\sigma^2$. □

The estimator obtained in Theorem 4.2.1 is a deterministically shrunken estimator as defined and studied in Mayer and Willke [14]. Its form is simply a scalar times the OLS estimator. The scalar factor is between 0 and 1 and, as such, has the effect of shortening the length of the vector $\hat{\beta}$.

4.3 The DMTMSE Estimator For The Class of General Linear Functions

In this section we will be concerned with the search for the DMTMSE estimator for the class of general linear functions of the form Ay, where A is any $(m \times n)$ matrix.

The form of the estimators that we have obtained in previous sections and standard results from linear model theory, suggest an admissibility condition which is elaborated in the following theorem.

Theorem 4.3.1

With respect to DMTMSE estimation, all estimators not of the form Ay, where $A = CU'$ for some C, are inadmissible.

Proof: Suppose that Ay is any linear estimation of β, then it can be seen that, with respect to DMTMSE estimation, $AUU'y$, where U is defined by (2.4), is superior to it. This is because

$$k^2 Ch_M[(I - AUU'X)'(I - AUU'X)]$$

$$= k^2 Ch_M[(I - AX)'(I - AX)]$$

and, since $I - UU'$ is n.n.d., we can write $I - UU' = DD'$, say, so that

$$0 \leq tr[(AD)(AD)'] = tr(A(I - UU')A')$$

$$= tr(AA') - tr(AUU'A') .$$

Thus,

$$S_T(A,k) \geq S_T(AUU',k)$$

for any A. □

This theorem states that the original model

$$y = X\beta + e = U\Lambda^{1/2}V'\beta + e$$

can be partitioned into

$$y_1 = \Lambda^{1/2}V'\beta + U'e$$

and

$$y_2 = \tilde{U}'e$$

where $\tilde{U}$ is the orthogonal complement of U, $y_1 = U'y$, and $y_2 = \tilde{U}'y$, and that all estimation should be based solely on y_1.

Thus, without loss of generality we may restrict our attention to the problem of estimation of β in the model:

$$y = X\beta + e \qquad (4.3.1)$$

where y is $(m \times 1)$, $X = \Lambda^{1/2}V'$ is $(m \times m)$ of rank m, and e is $(m \times 1)$ with $E(e) = 0$ and $Var(e) = I\sigma^2$.

Restricting our attention to the model defined in (4.3.1), an admissibility condition for estimators of the form Ay is introduced in the next theorem.

Theorem 4.3.2

In the model $y = X\beta + e$, estimators Ay of β are inadmissible with respect to DMTMSE estimation if $AX \neq (AX)'$.

Proof: For the model defined in (4.3.1), we have

$$S_T(A,k) = \sigma^2 tr[AA'] + k^2 Ch_M[(I-AX)(I-AX)'] . \qquad (4.3.2)$$

Let

$$A_0 = AX \qquad (4.3.3)$$

so that

$$A = A_0 X^{-1} . \qquad (4.3.4)$$

By substituting (4.3.3) and (4.3.4) in (4.3.2), we have

$$S_T(A_0,k) = \sigma^2 tr[A_0(X'X)^{-1}A_0'] + k^2 Ch_M[(I-A_0)(I-A_0)'] .$$

Consider now the symmetric matrix

$$A_1 = I - [(I-A_0)(I-A_0)']^{1/2} ,$$

where

$$[(I-A_0)(I-A_0)']^{1/2}$$

is a positive symmetric square root.

Then it follows easily that

$$Ch_M[(I-A_0)(I-A_0)'] = Ch_M[(I-A_1)(I-A_1)']$$

so it remains to prove that

$$\mathrm{tr}[A_0(X'X)^{-1}A_0'] \geq \mathrm{tr}[A_1(X'X)^{-1}A_1]$$

or

$$\mathrm{tr}[A_0V\Lambda^{-1}V'A_0'] \geq \mathrm{tr}[A_1V\Lambda^{-1}V'A_1'] \ , \qquad (4.3.5)$$

since $X = \Lambda^{1/2}V'$.

Before proving (4.3.5), observe the following:

$$\mathrm{tr}[(I-A_1)'(X'X)^{-1}(I-A_1)]$$

$$= \mathrm{tr}\left\{[(I-A_0)(I-A_0)']^{1/2}(X'X)^{-1}[(I-A_0)(I-A_0)']^{1/2}\right\}$$

$$= \mathrm{tr}[(I-A_0)(I-A_0)'(X'X)^{-1}]$$

$$= \mathrm{tr}[(I-A_0)'(X'X)^{-1}(I-A_0)] \ ,$$

therefore,

$$-2\mathrm{tr}[A_1'(X'X)^{-1}] + \mathrm{tr}[A_1'(X'X)^{-1}A_1]$$

$$= -2\mathrm{tr}[A_0'(X'X)^{-1}] + \mathrm{tr}[A_0'(X'X)^{-1}A_0]$$

or

$$\mathrm{tr}[A_0'(X'X)^{-1}A_0] - \mathrm{tr}[A_1'(X'X)^{-1}A_1]$$

$$= 2\left\{\mathrm{tr}[A_0'(X'X)^{-1}] - \mathrm{tr}[A_1'(X'X)^{-1}]\right\} \ .$$

Hence,

$$\mathrm{tr}[A_0'(X'X)^{-1}] \geq \mathrm{tr}[A_1'(X'X)^{-1}] \qquad (4.3.5)$$

will imply

$$\mathrm{tr}[A_0'(X'X)^{-1}A_0] \geq \mathrm{tr}[A_1'(X'X)^{-1}A_1] .$$

Therefore, we have only to prove (4.3.5). This will be done as follows:

$$\mathrm{tr}[V\Lambda^{-1}V'] - \mathrm{tr}[A_1V\Lambda^{-1}V'] = \mathrm{tr}[(I-A_1)V\Lambda^{-1}V']$$

$$= \mathrm{tr}\{[(I-A_0)(I-A_0)']^{1/2}V\Lambda^{-1}V'\}$$

$$= \mathrm{tr}\{[V\Lambda^{-1}V'(I-A_0)(I-A_0)'V\Lambda^{-1}V']^{1/2}\}$$

$$\geq \mathrm{tr}[(I-A_0)'V\Lambda^{-1}V']$$

$$= \mathrm{tr}[V\Lambda^{-1}V'] - \mathrm{tr}[A_0'V\Lambda^{-1}V'] , \qquad (4.3.6)$$

where the inequality holds, since for any real matrix W ,

$$\mathrm{tr}[(W'W)^{1/2}] \geq \mathrm{tr}[W] ,$$

(see for example, Marcus and Minc [12], Section 4.2). From (4.3.6) we may conclude

$$\mathrm{tr}[A_0'V\Lambda^{-1}V'] \geq \mathrm{tr}[A_1'V\Lambda^{-1}V']$$

as required. Therefore,

$$S_T(A_0,k) \geq S_T(A_1,k) . \quad \square$$

Our main purpose now is to prove the following theorem:

Theorem 4.3.3

In the model $(y, X\beta, I\sigma^2)$, where y is $(m \times 1)$, $X = \Lambda^{1/2}V'$ is $(m \times m)$ or rank m , the DMTMSE estimator of β is given by $\tilde{\beta} = t\hat{\beta}$, where

$$t = c^2/(\sum_{i=1}^{m} \lambda_i^{-1} + c^2) ,$$

and $c^2 = k^2/\sigma^2$.

Proof: By employing the same transformation used in the previous theorem, we have

$$S_T(A_0,k) = \sigma^2 tr[A_0(X'X)^{-1}A_0'] + k^2 Ch_M[(I-A_0)(I-A_0)']$$

$$= \sigma^2 tr[G\Gamma^2 G'V\Lambda^{-1}V'] + k^2 Ch_M[G(I-\Gamma)^2 G']$$

where by Theorem 4.3.3, A_0 is a symmetric matrix, that can be written as

$$A_0 = G\Gamma G' .$$

From the above we have,

$$S_T(A_0,k) = \sigma^2 tr[\Gamma G'V\Lambda^{-1}V'G\Gamma] + k^2 \max_i[(1-\gamma_i)^2]$$

$$= \sigma^2 \sum_i \gamma_i^2 \sum_k (g_i'v_k)^2\lambda_k^{-1} + k^2 \max_i[(1-\gamma_i)^2]$$

$$= \sigma^2 \sum_i \gamma_i^2 z_i^2 + k^2 \max_i [(1-\gamma_i)^2]$$

$$= \sigma^2 \sum_i \delta_i^2 + k^2 \max_i [(1-\delta_i z_i^{-1})^2]$$

where

$$\delta_i^2 = \gamma_i^2 z_i^2 \ ;$$

$$\gamma_i = \delta_i z_i^{-1} \ ,$$

since $z_i \neq 0$, g_i is the $i^{\underline{th}}$ column of G and v_k the $k^{\underline{th}}$ column of V :

Using the identical argument of Theorem 4.2.1, the minimizing values are

$$\delta_i = c^2/(\sum_{j=1}^{m} z_i^2 + c) z_i \ , \quad i = 1, \ldots, m \ ,$$

and, therefore,

$$t = \gamma_i = c^2/(\sum_{j=1}^{m} \lambda_i^{-1} + c^2) \ , \quad i = 1, \ldots, m$$

since

$$\sum_{i=1}^{m} z_i^2 = tr[G'V\Lambda^{-1}V'G] = \sum_{i=1}^{m} \lambda_i^{-1} \ .$$

Furthermore, it is easily seen that the choice of G is arbitrary. So we have

$$A_0 = AX = tI$$

which implies that

$$A = tX^{-1}$$

and, therefore, from Theorem 4.3.1 the resulting DMTMSE estimator of β is of the form:

$$AU'y = tX^{-1}U'y = tV\Lambda^{-1/2}U'y = t\hat{\beta} \ .\square$$

From Theorem 4.3.3, we may conclude that the DMTMSE estimator for the class of general linear functions coincides with that of the class of shrinkage estimators.

Unfortunately, for the general linear model $(y,X\beta,I\sigma^2)$, where y is $(n \times 1)$, and X is $(n \times m)$ of rank $= m \leq n$, the DMTMSE $[P'\beta]$ is not, in general, equal to $P'[\text{DMTMSE}(\beta)]$, where P is any $(t \times m)$ matrix. The following counterexample proves the assertion.

Suppose we want to estimate $p'\beta$ with $\ell'y$, where p' and ℓ' are, respectively, $(1 \times n)$ vectors.

$$S_T[\ell', p'\beta] = \sigma^2\ell'\ell + k^2(X'\ell - p)'(X'\ell - p) \ .$$

Taking the derivative of S_T with respect to ℓ and equating it to zero, we have

$$[\sigma^2 I + k^2XX']\ell = k^2Xp$$

so that

$$\ell' = k^2[\sigma^2 I + k^2XX']^{-1}Xp$$

giving the estimator

$$\text{DMTMSE}(p'\beta) = \ell'y = p'X[c^{-2}I + XX']^{-1}y$$

$$= p'V\Gamma U'y \ , \qquad (4.3.7)$$

where Γ is a diagonal matrix with elements

$$\gamma_i = \frac{\lambda_i^{1/2}}{(\lambda_i + c^{-2})} \ , \quad i = 1, \ldots, m \ .$$

Thus, DMTMSE $(p'\beta)$ is clearly different from

$$p'\text{DMTMSE}(\beta) = p't'\hat{\beta} = p't'V\Lambda^{-1/2}U'y \ ,$$

where

$$t = c^2/(\sum_{i=1}^{m} \lambda_i^{-1} + c^2) \ .$$

5. DIRECTIONALLY MINIMAX MEAN SQUARED ERROR ESTIMATOR

In Chapter 4, we were concentrating on the search for an estimator of β which depends as little as possible on the parameters themselves and which, in some sense, minimizes the trace of the MSE, $T(Ay,\beta)$. This chapter will be devoted to the search for an estimator of β which we will expect to be free, as much as possible, of the parameters themselves and which minimizes, in some sense, the mean squared error matrix. The strategy that we will adopt, similar to the last chapter, is to express the parameters in the form $\beta = k\alpha$, where the vector α contains its direction cosines and k its length. Then, for a fixed value of k, the MSE matrix, $M(Ay,k\alpha)$, can be "maximized", in a sense that we will define later, by choice of α. Then we will find the matrix A that will "minimize" the resulting criterion.

5.1 A Matrix Ordering

Let H be any square matrix. As usual, H will be called positive definite (p.d.) if $x'Hx > 0$ for all $x \neq 0$; positive semi-definite (p.s.d.) if $x'Hx \geq 0$ for all x and $= 0$ for some $x \neq 0$; nonnegative definite (n.n.d.) if it is either p.d. or p.s.d.; zero definite (z.d.) if $x'Hx = 0$ for all x. For these four we write $H > 0$, $H \gtrsim 0$, $H \geq 0$ and $H = 0$, respectively, where 0 is the null matrix. Finally, we define $H \geq B$ to mean $H - B \geq 0$; this also can be written $B \leq H$.

Chipman [3] has proved the following important lemma regarding this ordering among square matrices.

Lemma 5.1.1

The relation $\geq$ among square matrices is transitive and, among symmetric matrices, it is also anti-symmetric.

In view of this lemma, we shall speak of minimizing a symmetric, nonnegative definite matrix over a set $\mathcal{H}$, that is finding a matrix $H \in \mathcal{H}$, where $\mathcal{H}$ is a certain class of matrices, such that $B \geq H$ for all $B \in \mathcal{H}$.

Owing to the anti-symmetry of the relation $\geq$, if a set of symmetric matrices has a minimum, the minimum matrix is a fortori unique.

5.2 Geometric Interpretation of The Matrix Ordering

In order to develop more insight about the matrix ordering that we have defined, we shall state a theorem which will permit us to visualize what is meant by the relation $H \leq B$. We will be considering only the case where $B \geq 0$ and $H \geq 0$.

Observe first that according to our definition, $H \leq B$, if and only if, $x'Hx \leq x'Bx$ for all x, therefore $Bx = 0$ will imply $Hx = 0$, and, hence, $\{x \mid Bx = 0\} \subseteq \{x \mid Hx = 0\}$ so that if $x \notin \{x \mid Hx = 0\}$ then clearly $x \notin \{x \mid Bx = 0\}$.

Theorem 5.2.1

Let H and B be any two n.n.d. symmetric matrices, and define $E_H = \{x \mid x'Hx = 1\}$ and $E_B = \{x \mid x'Bx = 1\}$, then $x \in E_H$ implies that

there exists a number $m \leq 1$ and $m \neq 0$ such that $mx \in E_B$, if, and only if, $H \leq B$.

Proof: Assume first $H \leq B$ and let $x \in E_H$, therefore

$$1 = x'Hx \leq x'Bx .$$

Now it is clear from the note above that $x'Bx > 0$, so that if we choose

$$m = \frac{1}{\sqrt{x'Bx}} \leq 1$$

we will obtain

$$mx'Bxm = \frac{1}{x'Bx} x'Bx = 1$$

that is, $mx \in E_B$.

Now let x be an arbitrary vector. Then either $x'H = 0$ or $m_1 x \in E_H$ for some $m_1 \neq 0$. If the former holds, then

$$x'Hx = 0 \leq x'Bx .$$

If the second assertion holds, we have

$$m_1^2 x'Hx = 1$$

and by hypothesis there exist $m_2 \neq 0$ and $m_2 \leq 1$ such that

$$m_2 m_1 x'Bx m_1 m_2 = m_1^2 m_2^2 x'Bx = 1 .$$

Hence,

$$m_1^2 x'Hx = 1 = m_1^2 m_2^2 x'Bx .$$

Dividing this expression by m_1^2 we finally get

$$x'Hx = m_2^2 x'Bx \leq x'Bx \ .$$

Since x was chosen arbitrarily,

$$H \leq B \ . \ \square$$

By Theorem 5.2.1, we can intuitively conclude that $H \leq B$ means that the graph defined by the point set E_B is entirely contained in that for E_H, in the sense that, if we choose a point in E_H and travel towards the origin, we will cross E_B before reaching the origin. The following example in a two-dimensional space will help to illustrate the idea.

Example 5.2.1

Let $E_B = \{x \mid x'(Ia'a)x = 1\}$ and $E_H = \{x \mid x'aa'x = 1\} = \{x \mid x'a = 1 \text{ or } x'a = -1\}$, where $x' = (x_1, x_2)$ and $a' = (a_1, a_2)$ are (1×2) vectors, $B = Ia'a$ and $H = aa'$ are two (2×2) symmetric n.n.d. matrices.

It follows from the Cauchy-Schwarz inequality that

$$x'Hx = x'aa'x \leq a'ax'x = x'(Ia'a)x = x'Bx$$

for all x and, hence,

$$H = aa' \leq Ia'a = B \ .$$

It is clear from Figure 5.2.1 that if we choose a point that lies either on the lines $x'a = 1$ or $x'a = -1$ and we travel towards the origin, we will cross the circle defined by $x'(Ia'a)x = 1$, before reaching the origin.

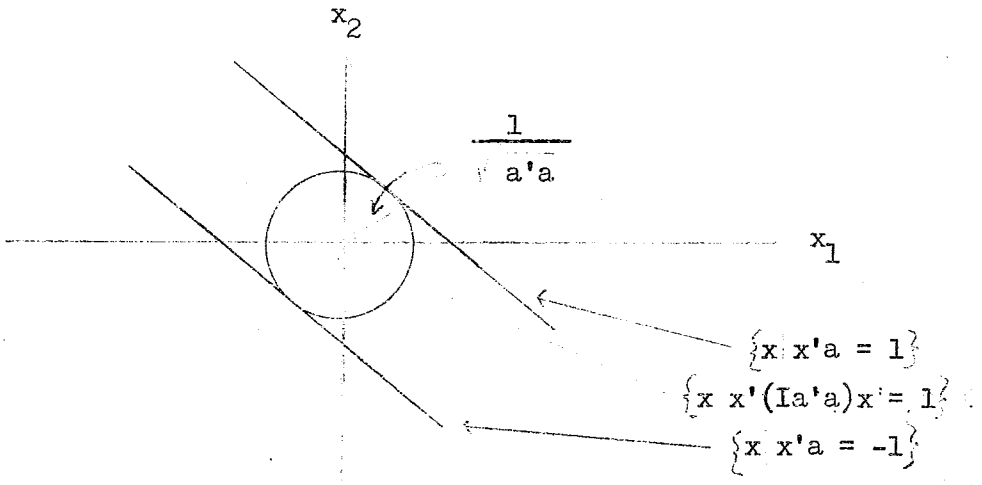

Figure 5.2.1 Graphs for the point sets $E_H = \{x \mid x'aa'x = 1\}$ and $E_B = \{x \mid x'(Ia'a)x = 1\}$

5.3 Definition of a Supremum in The Matrix Ordering

In Section 5.1, we have defined the meaning of the relation $H \leq B$ for two n.n.d., symmetric matrices. In this section, we will be concerned with formalizing the concept of a supremum in the context of the ordering that we are considering. This concept will permit us to define a criterion that can be used to obtain biased estimators which will be superior to OLS with respect to MSE.

Before we formalize the definition of the supremum, let us consider the following sets of matrices:

$$V = \{B \mid B \text{ is a n.n.d., symmetric matrix}\} ,$$

$$Z \subseteq V, \text{ where } Z \text{ is a non-empty set,}$$

$$W = \{B \mid B \in V \text{ and } (H \in Z \Rightarrow H \leq B)\} ,$$

$$= \text{set of upper bounds for the set } Z \text{ in } V .$$

Now the precise meaning of a supremum is contained in the following definition.

Definition 5.3.1

A square matrix $B \in V$ is the least upper bound or supremum for the set Z in V if, and only if,

(1) B is an upper bound for each $H \in Z$, that is,
$H \leq B$ for all $H \in Z$,

and

(2) if $B_1 \in V$ and $B_1 \geq H$ for all $H \in Z$ then
$B_1 \geq B$, that is B is a minimum in W.

It is important for later work to point out that the supremum or least upper bound as has been defined does not always exist, even in the case when the set Z under consideration is bounded above. The following counter-example will illustrate this.

Example 5.3.1

Consider the set $Z = \{H_1, H_2\}$, where

$$H_1 = \begin{bmatrix} 3 & 0 \\ 0 & 2.5 \end{bmatrix}, \quad H_2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

It is clear that H_1 is a p.d., symmetric matrix, and since the characteristic roots of H_2 are 3 and 1, it is also a p.d., symmetric matrix.

Let

$$B_1 = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

and

$$B_2 = \begin{bmatrix} 4 & 0 \\ 0 & 2.5 \end{bmatrix}.$$

These are clearly two p.d., symmetric matrices. Moreover, it follows immediately that the following relations hold: $B_1 \geq H_1$, $B_2 \geq H_1$. Since the characteristic roots of the matrix $B_1 - H_2$ are 0 and 2, and the ones corresponding to the matrix $B_2 - H_2$ are 0 and 3/2, the following relations also hold: $B_1 \geq H_2$, $B_2 \geq H_2$. Hence, we may conclude that the set $Z = \{H_1, H_2\}$ is bounded above by the matrices B_1 and B_2; i.e., B_1 and B_2 are elements of W. However, there does not exist a n.n.d., symmetric matrix,

$$C = \begin{bmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{bmatrix},$$

such that the following relations hold:

(i) $B_1 \geq C$,

(ii) $B_2 \geq C$,

(iii) $C \geq H_1$,

and

(iv) $C \geq H_2$.

In order to confirm the above assertion, suppose (i), (ii) and (iii) hold. Then it follows immediately that $c_{11} = 3$ and $c_{22} = 2.5$. Using this and (iii), we have that

$$\begin{bmatrix} 3 & c_{12} \\ c_{12} & 2.5 \end{bmatrix} \geq \begin{bmatrix} 3 & 0 \\ 0 & 2.5 \end{bmatrix}$$

which implies that for any $x' = (x_1, x_1)$

$$(x_1, x_2)\begin{bmatrix} 3 & c_{12} \\ c_{12} & 2.5 \end{bmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \geq 3x_1^2 + 2.5x_2^2$$

$\forall$, x_1 and x_2 .

Thus, letting $x_1 = 1$ and $x_2 = 1$, we have

$$5.5 + 2c_{12} \geq 5.5$$

and, hence,

$$c_{12} \geq 0 \, . \tag{5.3.1}$$

Similarly, from (ii) and (iii), we get the following inequalities:

$$3x_1^2 + 3x_2^2 \geq 3x_1^2 + 2.5x_2^2 + 2x_1x_2c_{12}$$

and

$$4x_1^2 + 2.5x_2^2 \geq 3x_1^2 + 2.5x_2^2 + 2x_1x_2c_{12}$$

and, letting $x_1 = x_2 = x_0$, we have

$$\frac{1}{2}\, x_0^2 \geq 2x_0^2c_{12}$$

and

$$x_0^2 \geq 2x_0^2c_{12}$$

which, in turn, implies

$$0 \geq x_0^2c_{12} .$$

Hence,

$$0 \geq c_{12} , \qquad (5.3.2)$$

and from (5.3.1) and (5.3.2) we may conclude

$$c_{12} = 0 .$$

So, if we assume (i), (ii), and (iii) true we must conclude that

$$C = H_1 .$$

But, since the characteristic roots of $H_1 - H_2$ are 1.78 and -0.28, we can see that H_1 and H_2 are not comparable in the sense that we have defined. Therefore, we may conclude that there does not exist a n.n.d., symmetric matrix C, such that (i), (ii), (iii), and (iv)

hold and, therefore, that the supremum for our set $Z = \{H_1, H_2\}$ does not exist, even though the set is bounded above.

Before we set out some sufficient conditions for the existence of a supremum, let us first develop some useful notations.

Suppose that the n.n.d., symmetric matrix H is a function of a. We shall denote it by $H(a)$, where a belongs to a certain non-empty set $\mathcal{A}$, so that our set Z can be described as $\{H \mid H = H(a), a \in \mathcal{A}\}$. Under this set up we can obtain (if it exists) the supremum B of the set of matrices $H(a)$ in V over all $a \in \mathcal{A}$ and we shall write this assertion as

$$B = \sup_{a \in \mathcal{A}} H(a) .$$

Theorem 5.3.1

Let $B \in V$ and $H(a) \in Z \subseteq V$ for all $a \in \mathcal{A}$, moreover, let $E_B = \{x \mid x'Bx = 1\}$ and suppose the following conditions hold:

(i) $H(a) \leq B$ for all $a \in \mathcal{A}$

(ii) For every $x \in E_B$ there exists an $a \in \mathcal{A}$ such that

$$x'Bx = x'H(a)x ,$$

then

$$\sup_{a \in \mathcal{A}} H(a) = B .$$

Proof: By assumption (i) B is an upper bound for Z in V. So it remains to be proven that if $B_1 \in W$ then $B \subseteq B_1$, that is

$$x'Bx \leq x'B_1x \quad \text{for all } x .$$

To prove this, let x be an arbitrary vector. Then one of the two following cases must occur:

(1) $Bx = 0$, in which case $x'Bx = 0 \leq x'B_1x$, since $B_1 \in V$.

(2) There exists $c_1 \neq 0$ such that $c_1x \in E_B$; that is $c_1^2x'Bx = 1$. But by (ii) there exists an $a \in \mathcal{A}$ such that

$$1 = c_1^2x'Bx = c_1^2x'H(a)x \leq c_1^2x'B_1x$$

where again the inequality hold since $B_1 \in W$.

By (1) and (2) we may conclude that $B \leq B_1$ and hence

$$\sup_{a\in\mathcal{A}} H(a) = B \ . \ \square$$

5.4 Directionally Minimax Mean Squared Error Estimation

The goal of this section is to develop a meaningful criterion to obtain an estimator of β in the general linear model $(y, X\beta, \Sigma\sigma^2)$, $|\Sigma| \neq 0$, which depends, as little as possible, on the parameters themselves and which, in some sense, minimizes the mean squared error matrix. We shall be concerned with the joint estimation of β using linear functions of the observations of the form Ay , where A is some $(t \times n)$ matrix, $t \leq n$.

The strategy adopted, as was done in Chapter 4, is to express the parameters in the form $\beta = k\alpha$, where α is the vector of its direction cosines and k its length. Then, for fixed values of k ,

the matrix $M(Ay,k\alpha)$, can be maximized, in the sense defined in Section 5.3, by choice of α . The precise meaning of what we will call a Directionally Minimax Mean Squared Error Estimator is contained in the following definition:

Definition 5.4.1

A linear estimator $\widetilde{P'\beta} = Ay$ is said to be Directionally Minimax Mean Squared Error (DMMSE) estimator of $P'\beta$ if A is such that it minimizes

$$S_M(Ay,P'\beta;k,\Sigma\sigma^2) = \sup_{\beta\in\Theta_k} M(Ay,P'\beta;\beta,\Sigma\sigma^2) =$$

$$= \sigma^2 A\Sigma A' + \sup_{\beta\in\Theta_k} (P' - AX)\beta\beta'(P' - AX)$$

where

$$\Theta_k = \{\beta \mid \beta = k\alpha \ , \ \|\alpha\| = 1\} \ .$$

In view of Definition 5.4.1, we will be concerned first with the problem of finding a n.n.d., symmetric matrix B such that

$$B = \sup_{\beta\in\Theta_k} (P' - AX)\beta\beta'(P' - AX)' \ .$$

This problem will be confronted in the next two theorems, in which we will assume, without loss of generality that $k^2 = 1$.

Theorem 5.4.1

Let $T = DD'$, where D is a $(t \times r)$ matrix of rank r. Then $T = DD' \geq F\beta\beta'F'$ for all $\beta \in \mathcal{Q}_1$ if, and only if, $F = DC$ for some $(r \times m)$ matrix C and $Ch_M(C'C) \leq 1$.

Proof: Assume first that

$$DD' \geq F\beta\beta'F' \quad \text{for all} \quad \beta \in \mathcal{Q}_1 .$$

This inequality holds, if and only if,

$$x'F\beta\beta'F'x \leq x'DD'x \quad \text{for all} \quad x \quad \text{and for all} \quad \beta \in \mathcal{Q}_1 . \qquad (5.4.1)$$

Now if $D'x = 0$ we will have from (5.4.1) that

$$\beta'F'x = 0 \quad \text{for all} \quad \beta \in \mathcal{Q}_1$$

and, hence, $F'x = 0$. Thus,

$$\mathcal{Q}^{\perp}(D') \subseteq \mathcal{Q}^{\perp}(F')$$

therefore,

$$\mathcal{Q}(F') \subseteq \mathcal{Q}(D')$$

and from this we have that

$$F = DC \quad \text{for some} \quad C . \qquad (5.4.2)$$

If, as we have assumed,

$$F\beta\beta'F' \leq DD' \quad \text{for all} \quad \beta \in \mathcal{Q}_1 ,$$

then by (5.4.2) we have

$$DC\beta\beta'C'D' \leq DD' \quad \text{for all} \quad \beta \in \mathfrak{B}_1 .$$

This last inequality holds if, and only if,

$$x'DC\beta\beta'C'D'x \leq x'DD'x \quad \text{for all} \quad \beta \in \mathfrak{B}_1 \quad \text{and all} \quad x .$$

Hence, in particular,

$$m'D'DC\beta\beta'C'D'Dm \leq m'D'DD'Dm \quad \text{for all} \quad \beta \in \mathfrak{B}_1 \quad \text{and all} \quad m .$$

This happens if, and only if,

$$D'DC\beta\beta'C'D'D \leq D'DD'D \quad \text{for all} \quad \beta \in \mathfrak{B}_1 ,$$

where $|D'D| \neq 0$ and, therefore, the relation holds if, and only if,

$$C\beta\beta'C' \leq I \quad \text{for all} \quad \beta \in \mathfrak{B}_1 . \qquad (5.4.3)$$

By (5.4.3) we have that

$$Ch_M(C\beta\beta'C') \leq 1 \quad \text{for all} \quad \beta \in \mathfrak{B}_1$$

which is true if, and only if,

$$\beta'C'C\beta \leq 1 \quad \text{for all} \quad \beta \in \mathfrak{B}_1$$

which in turn is true if, and only if,

$$Ch_M(C'C) \leq 1 .$$

Let us now assume that $F = DC$ for some C and $Ch_M(C'C) \leq 1$. Then,

$$F\beta\beta'F' = DC\beta\beta'C'D'$$

and, by the Cauchy-Schwarz inequality, we will obtain

$$x'DC\beta\beta'C'D'x \leq x'DD'x\beta'C'C\beta \leq x'DD'x \quad \text{for all} \quad \beta \in \mathcal{B}_1 ,$$

and all x .

This implies that

$$DC\beta\beta'C'D \leq DD'$$

and, hence, by our assumption we will conclude

$$F\beta\beta'F' \leq DD' = T . \quad \square$$

Lemma 5.4.1

$F'F$ is an upper bound for $F\beta\beta'F'$, that is,

$$F\beta\beta'F' \leq FF' \quad \text{for all} \quad \beta \in \mathcal{B}_1 .$$

Proof: The proof follows immediately from the Cauchy-Schwarz inequality. $\square$

Theorem 5.4.2

If $T = DD' \geq F\beta\beta'F'$ for all $\beta \in \mathcal{B}_1$ then the n.n.d., symmetric matrix FF' is such that

$$FF' \leq DD' = T .$$

Proof: By Theorem 5.4.1 we can write $F = DC$ for some C. Let $CC' = G\Delta G'$, where $GG' = G'G = I$ and Δ is a diagonal matrix, that is, $\Delta = \text{diag}(\delta_i)$.

It follows immediately from Theorem 5.4.1 and Lemma 5.4.1 that $\delta_i \leq 1$ for all i. Therefore,

$$\Delta - I \leq 0$$

if, and only if,

$$x'(\Delta - I)x \leq 0 \quad \text{for all} \quad x .$$

In particular,

$$m'DG(\Delta - I)G'D'm \leq 0 \quad \text{for all} \quad m .$$

Therefore,

$$m'(DCC'D' - DD')m \leq 0 \quad \text{for all } m .$$

Thus, we may conclude by Theorem 5.4.1 that

$$FF' \leq DD' . \quad \square$$

Theorem 5.4.2 states that if $T \in V$ is an upper bound for $F\beta\beta'F'$, then it satisfies $T \geq FF'$. Then, in view of Lemma 5.4.1, which states that FF' is also an upper bound for $F\beta\beta'F'$,

$$FF' = \sup_{\beta \in \mathcal{B}_1} F\beta\beta'F' .$$

Therefore, our criterion can be reduced to

$$S_M(Ay, P'\beta; k, \Sigma\sigma^2) = \sigma^2 A\Sigma A' + \sup_{\beta \in \mathcal{B}_k} (P'-AX)\beta\beta'(P'-AX)'$$

$$= \sigma^2 A\Sigma A' + k^2(P' - AX)(P' - AX)' \quad . \quad (5.4.4)$$

In Theorem 5.3.1 we have developed a sufficient condition for the existence of the supremum that will be used in the following example to illustrate another route for obtaining the

$$\sup_{\beta \in \mathcal{B}_k} F\beta\beta'F' \quad .$$

Example 5.4.1

It is clear that

$$F\beta\beta'F' \leq k^2 FF' \quad \text{for all} \quad \beta \in \mathcal{B}_k \quad .$$

Now let $x \in E_{k^2 FF;}$, that is,

$$x'FF'k^2x = 1 \quad ,$$

We must be able to find a $\beta \in \mathcal{B}_k$, such that

$$x'F\beta\beta'F'x = 1 \quad .$$

To do so, let $\beta_0 = cF'x$. Since $\beta_0'\beta_0 = k^2$ that implies that

$$k^2 = c^2 x'FF'x$$

and, hence, that

$$c^2 = k^2/x'FF'x \quad .$$

Thus,

$$x'F\beta_0\beta_0'F'x = (x'FF'x)^2c^2 = x'FF'k^2x = 1$$

and, hence, by virtue of Theorem 5.3.1 we may conclude

$$\sup_{\beta \in \Theta_k} F\beta\beta'F' = FF'k^2 \ . \ \square$$

The problem on which we shall concentrate our attention now is the search for an $(m \times n)$ matrix A, which minimizes our criterion given in 5.4.4 and which will yield a possibly biased estimator of the form $\tilde{\beta} = Ay$ which will be superior to OLS with respect to MSE.

In the process of solving a different problem, Foster [6] and Chipman [3] obtained the matrix A which minimizes the matrix $AVA' + (I-AX)U(I-AX)'$, where U and V are aribtrary p.d. matrices. The form of their solution is $A_0 = UX'(XUX' + V)^{-1}$. This result can be applied to the criterion (5.4.4).

Prior to discovering their work, the following theorem was proved. It solves the same problem for the case when $V = \Sigma\sigma^2$ and $U = k^2I$. Because of its simplicity, the theorem along with its proof will be presented.

Theorem 5.4.3

In the linear model $(y, X\beta, \Sigma\sigma^2)$, $|\Sigma| \neq 0$,

$$\min_A Q(A) = \min_A \left\{k^{-2}S_M(Ay, \beta; k, \Sigma\sigma^2)\right\}$$

$$= \min_A \left\{c^{-2}A\Sigma A' + (I-AX)(I-AX)'\right\} = (I + X'\Sigma^{-1}Xc^2)^{-1}$$

and the equality holds when $A = X'(\Sigma c^{-2} + XX')^{-1}$, where

$$c^2 = \frac{k^2}{\sigma^2} .$$

That is, the DMMSE estimator is

$$\tilde{\beta} = X'(\Sigma c^{-2} + XX')^{-1}y .$$

Proof: Let $E = (\Sigma c^{-2} + XX')$. From this we have

$$Q(A) = c^{-2}A\Sigma A' + (I-AX)(I-AX)'$$

$$= c^2 A\Sigma A' + I - AX - X'A' + AXX'A'$$

$$= A(\Sigma c^{-2} + XX')A + I - AX - X'A'$$

$$= AEA' - AEE^{-1}X - X'E^{-1}EA' + X'E^{-1}EE^{-1}X + I - X'E^{-1}X$$

$$= (A - X'E^{-1})E(A - X'E^{-1})' + (I - X'E^{-1}X) .$$

Since

$$I - X'(\Sigma c^{-2} + XX')^{-1}X = (I + X'\Sigma^{-1}Xc^2)^{-1}$$

we have

$$Q(A) = (A - X'E^{-1})E(A - X'E^{-1})' + (I - X'\Sigma^{-1}Xc^2)^{-1} . \quad (5.4.5)$$

Observe that only the first term in (5.4.5) depends on A and is equal to the zero matrix whenever $A = X'E^{-1}$. From this, we may conclude

$$\min_A Q(A) = (I + X'\Sigma^{-1}Xc^2)$$

and the equality holds when $A = X'(\Sigma c^{-2} + XX')^{-1}$. Finally, our estimator becomes

$$\tilde{\beta} = Ay = X'(\Sigma c^{-2} + XX')^{-1}y \ . \ \square$$

The following theorem states that, holding the model $(y, X\beta, \Sigma\sigma^2)$, $|\Sigma| \neq 0$, fixed, the DMMSE estimator $(P'\beta)$, of the linear function $P'\beta$ can be obtained by forming the corresponding linear functions, $P'\tilde{\beta}$, of the DMMSE estimator, $\tilde{\beta}$, of β .

Theorem 5.4.4

In the linear model $(y, X\beta, \Sigma\sigma^2)$, $|\Sigma| \neq 0$, and for any $(t \times m)$ matrix P'

$$DMMSE(P'\beta) = P'[DMMSE(\beta)]$$

and hence,

$$MSE[DMMSE(P'\beta)] = P'\{MSE[DMMSE(\beta)]\}P \ .$$

Proof: Let $E = (\Sigma c^{-2} + XX')$. From this we have

$$Q(A) = c^{-2}A\Sigma A' + (P'-AX)(P'-AX)'$$

$$= c^{-2}A\Sigma A' + P'P - P'X'A' - AXP + AXX'A'$$

$$= A(c^{-2}\Sigma + XX')A' + PP' - P'X'A' - AXP$$

$$= AEA - P'X'E^{-1}EA' - AEE^{-1}XP + P'P$$

$$+ P'X'E^{-1}EE^{-1}XP - P'X'E^{-1}XP$$

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

$$= (A - P'X'E^{-1})E(A' - E^{-1}XP) + P'(I + X'\Sigma^{-1}Xc^{2})^{-1} P \,, \tag{5.4.6}$$

Observe that only the first term in (5.4.6) depends on A and is equal to the zero matrix whenever $A = P'X'E^{-1}$. From this, we may conclude

$$\min_{A} Q(A) = P'(I + X'\Sigma^{-1}Xc^{2})^{-1}P$$

and the equality holds when

$$A = P'X'(\Sigma c^{-2} + XX')^{-1} \,,$$

Finally, our estimator $(\widetilde{P'\beta})$ becomes

$$(\widetilde{P'\beta}) = Ay = P'X'(\Sigma c^{-2} + XX')^{-1}y = P'\tilde{\beta} \,. \quad \square$$

5.5 An Alternative Procedure To Obtain The DMMSE Estimator

In view of Theorem 5.4.4, an alternative procedure for obtaining the DMMSE estimator, is to apply the minimax argument that has been developed in the last sections, to the MSE of an arbitrary linear combination of the observations, $\ell'y$, used to estimate an arbitrary linear parametric function of the parameters, $p'\beta$, in the linear model $(y, X\beta, \Sigma\sigma^2)$, $|\Sigma| \neq 0$. That is, we will apply the minimax argument to:

$$M(\ell'y,p'\beta) = \sigma^2\ell'\Sigma\ell + \beta'(X'\ell - p)(X'\ell - p)'\beta \quad . \qquad (5.5.1)$$

Then, expressing the parameters in the form $\beta = k\alpha$, where α are the direction cosines and k its length, then for fixed k , we can maximize $M(\ell'y,p'\beta)$ by choice of α , as follows:

$$S_M(\ell'y,p'\beta) = \sup_{\beta\in\Omega_k} M(\ell'y,p'\beta) = \sigma^2\ell'\Sigma\ell + k^2(X'\ell-p)'(X'\ell-p) \quad , \qquad (5.5.2)$$

where $\Omega_k = \{\beta;\beta = k\alpha \ , \ \|\alpha\| = 1\}$.

For S_M as in (5.5.2), Theorem 5.5.1 gives the optimum choice of ℓ.

Theorem 5.5.1

The directionally minimax MSE estimator of any function $p'\beta$ is $p'\tilde{\beta}$ where

$$\tilde{\beta} = k^2x'[\sigma^2\Sigma + k^2XX']^{-1}y$$

$$= X'[c^{-2}\Sigma + XX']^{-1}y \quad ,$$

where

$$c^{-2} = \frac{\sigma^2}{k^2} \quad .$$

Proof: Taking the derivative of S_M with respect to ℓ and equating it to zero, we have

$$\sigma^2 2\Sigma\ell + 2k^2XX'\ell - 2k^2Xp = 0 \quad . \qquad (5.5.3)$$

So the value of ℓ that minimizes S_M satisfies the equation

$$[\sigma^2\Sigma + k^2XX']\ell = k^2Xp$$

so that

$$\ell = k^2[\sigma^2\Sigma + k^2XX']^{-1}Xp$$

giving the estimator

$$\ell'y = p'k^2X'[\sigma^2\Sigma + k^2XX']^{-1}y$$

$$= p'X'[c^{-2}\Sigma + XX']^{-1}y$$

$$= p'\tilde{\beta} \qquad (5.5.4)$$

where

$$c^{-2} = \frac{\sigma^2}{k^2} . \quad \square$$

The estimator $\tilde{\beta}$ of β given in (5.5.4) is the DMMSE estimator of β.

The results which appear in Theorem 5.5.1 are certainly included in Theorem 5.4.4. The former were obtained early in the investigation and suggested that the DMMSE estimator of β might be

$$\tilde{\beta} = X'[c^{-2}\Sigma + XX']^{-1}y .$$

Without this hint, the proof of Theorem 5.4.3 would have been impossible. The results are included to indicate the steps which provided the motivation for the main results of this chapter.

5.6 Relation Between The DMMSE Estimator And The Ridge Regression Estimator

In the next theorem we will show the relation of the DMMSE estimator and the ridge regression estimator. For this we will restrict our attention to the model $(y, X\beta, I\sigma^2)$.

Theorem 5.6.1

The DMMSE estimator is equivalent to the ridge regression estimator $\tilde{\beta}^* = [I + c^{-2}(X'X)^{-1}]^{-1}(X'X)^{-1}X'y$ of Hoerl and Kennard [9].

Proof: The result follows immediately if we apply a formula given by Rao [17] to the form $X'[c^{-2}I + XX']^{-1}$. Indeed

$$X'[c^{-2}I + XX']^{-1} = X'[\frac{1}{c^{-2}}I - \frac{1}{c^{-2}} X(X'X)^{-1}X'$$

$$+ X(X'X)^{-1}[I + c^{-2}(X'X)^{-1}]^{-1}(X'X)^{-1}X']$$

$$= [I + c^{-2}(X'X)^{-1}]^{-1}(X'X)^{-1}X' .$$

Thus,

$$\tilde{\beta}^* = [I + c^{-2}(X'X)^{-1}]^{-1}\hat{\beta} = \tilde{\beta} ,$$

where $\hat{\beta}$ is the OLS estimators. □

5.7 Minimization of The Tr[S(Ay,k)]

For The Restriction Case

In Chapter 4 we studied the class of biased estimators of regression coefficients, by computing LS estimates under sets of false restrictions. For that case we were restricting our attention to the linear model $(y, X\beta, I\sigma^2)$ in which we assume that $\text{rank}(X) = m$. To obtain the class of estimators we impose u independent false restriction on the parameter space and by Theorem 4.1 we observed that restrictions of the form $R\beta = 0$ were preferred over $R\beta = h$. Moreover, in view of Theorem 4.2, we saw that we can specify the class of estimators of interest as

$$\tilde{\beta} = \tilde{\beta}(R) = (I - S^{-1}R'R)S^{-1}X'y \qquad (5.7.1)$$

for all R such that $RS^{-1}R' = I$, where S has been defined in 2.5. Within this class we found that the DMTMSE estimator, given by Theorem 4.4, has the form

$$\tilde{\beta} = (I - V_2V_2')S^{-1}X'y = V_1V_1'S^{-1}X'y .$$

By putting $A = (I - S^{-1}R'R)S^{-1}X'$, and $\Sigma = I$, $p = I$, in (5.4.4), we will propose an alternative procedure for obtaining the DMTMSE estimator. This procedure consists of finding, for fixed k, the matrix R that minimizes

$$T_S(R,k) = \text{tr}[S_M(R,k)] = \text{tr}[S^{-1} - S^{-1}R'RS^{-1}]$$

$$+ k^2\text{tr}[R'RS^{-2}R'R] . \qquad (5.7.2)$$

The optimum value of R is given in the following theorem.

Theorem 5.7.1

For the class of least squares estimates computed subject to a set of u independently false restrictions

$$T_S(R,k) = tr[S_M(R,k)] \geq \sigma^2 \sum_{i=1}^{m-u} \lambda_i^{-1} + k^2 u ,$$

where the equality holds whenever $R = \Lambda_2^{1/2} V_2'$, where λ_i, Λ_2, and V_2 are defined in (2.4), (2.7) and (2.8), respectively. That is, for this class of estimators, our estimator is

$$\tilde{\beta} = V_1 V_1' S^{-1} X'y = (I - V_2 V_2') S^{-1} X'y .$$

Proof: We can write $R = (BB')^{-1/2} B \Lambda^{1/2} y'$, where B is any $(u \times m)$ matrix of rank u. From this we have that

$$tr[S_M(R,k)] = \sigma^2 tr[S^{-1} - S^{-1}R'RS^{-1}] + k^2 tr[R'RS^{-1}R'R]$$

$$= \sigma^2 tr[\Lambda^{-1} - \Lambda^{-1/2} P_B \Lambda^{-1/2}]$$

$$+ k^2 tr[\Lambda^{-1/2} P_B \Lambda^{1/2} \Lambda^{1/2} P_B \Lambda^{-1/2}]$$

where $P_B = B'(BB')^{-1}B$, and Λ has been defined in (2.4).

In Theorem 4.4 we have proved that

$$tr[\Lambda^{-1} - \Lambda^{-1/2} P_B \Lambda^{-1/2}] \geq \sum_{i=1}^{m-u} \lambda_i^{-1} . \qquad (5.7.2)$$

Now let $Q = \Lambda^{-1/2} P_B \Lambda^{1/2}$ and observe that $Q^2 = Q$ and $\text{rank}(Q) = \text{rank}(P_B) = u$. Let v_i, $i = 1, \ldots, m$ be a

orthonormal basis for E^m such that the first u of them form an orthonormal basis for $\mathcal{R}(Q)$, we have

$$\begin{aligned} \operatorname{tr}(Q'Q) &= \sum_{i=1}^{m} v_i'Q'Qv_i \\ &= \sum_{i=1}^{u} v_i'Q'Qv_i + \sum_{i=u+1}^{m} v_i'Q'Qv_i \\ &= \sum_{i=1}^{u} v_i'v_i = u \ . \end{aligned} \tag{5.7.3}$$

By using (5.7.2) and (5.7.3) we may conclude

$$\begin{aligned} &\sigma^2\operatorname{tr}[S^{-1}-S^{-1}R'RS^{-1}] + k^2\operatorname{tr}[R'RS^{-2}R'R] \\ &\qquad \geq \sigma^2 \sum_{i=1}^{m-u} \lambda_i^{-1} + k^2u \ . \end{aligned} \tag{5.7.4}$$

It remains to demonstrate that the lower bound in (5.7.4) is attainable. Putting $B = (0 \vdots I_u)$ we get

$$R = (BB')^{-1/2}B\Lambda^{1/2}V' = \Lambda_2^{1/2}V_2' \tag{5.7.5}$$

and substituting (5.7.5) into the left hand side of (5.7.4) gives

$$\sigma^2\operatorname{tr}[V_1\Lambda_1^{-1}V_1'] + k^2\operatorname{tr}[V_2V_2'] = \sigma^2 \sum_{i=1}^{m-u} \lambda_i^{-1} + k^2u$$

which is the right hand side of (5.7.4).

Finally, our estimator takes the form

$$\tilde{\beta} = Ay = (I - V_2V_2')S^{-1}X'y = V_1V_1'S^{-1}X'y = V_1V_1'\hat{\beta} \ ,$$

where $\hat{\beta}$ is the OLS estimator. □

In view of Theorem 5.7.1, we may conclude the following important fact.

FACT: For the class of biased estimation computed subject to possibly false restrictions, the DMTMSE estimator can be obtained by minimizing with respect to R either of the following two expressions:

$$\min_R S_T(R,k) = \min_R \Big\{\sigma^2 \mathrm{tr}[S^{-1} - S^{-1}R'RS^{-1}] + \sup_{\beta \in \mathcal{Q}_k} \mathrm{tr}[F\beta\beta'F']\Big\},$$

or

$$\min_R \Big\{\sigma^2 \mathrm{tr}[S^{-1} - S^{-1}R'RS^{-1}] + \mathrm{tr}[\sup_{\beta \in \mathcal{Q}_k} F\beta\beta'F']\Big\} \qquad (5.7.6)$$

where

$$F = S^{-1}R'R \ .$$

6. JOINT DIRECTIONALLY MINIMAX MEAN SQUARED ERROR ESTIMATION

This chapter is a continuation of Chapter 5 in that the form of the estimator obtained there is utilized here. The main subject matter is the problem of estimating regression coefficients for a set of dependent regression equations describing different variables but sharing the same design matrix, when the $X'X$ matrix has some small but positive eigenvalues.

DMMSE estimation applied equation-by-equation yields efficient coefficient estimators under special conditions. For conditions generally encountered we propose an estimator in which all parameters are estimated simultaneously.

In Section 6.2, the proposed estimator is presented and its asymptotic distribution is studied. In Section 6.3 a comparison between the variance-covariance matrices of the joint LS estimator, the joint ridge regression and the joint DMMSE estimation is presented.

6.1 Some Basic Definitions and Notation

In this section we will define the Kronecker product of matrices, order of magnitude of a sequence, and some notation concerning asymptotically multinormal random vectors.

Definition 6.1.1 (Rao [17], p. 29)

Let $A = (a_{ij})$ and $B = (b_{ij})$ be $(m \times n)$ and $(p \times q)$ matrices, respectively. Then the Kronecker product

$$A \otimes B = (a_{ij}B)$$

is an $(mp \times nq)$ matrix expressible as a partitioned matrix with $a_{ij}B$ as the $(i,j)^{\underline{th}}$ partition, $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

Definition 6.1.2

The sequence a_n is said to be of smaller order than n^k, indicated by $o(n^k)$, if the sequence $n^{-k}a_n$ converges to zero, and moreover, we say that the sequence a_n is at most of order n^k, written $O(n^k)$, when the sequence $n^{-k}a_n$ is bounded.

Definition 6.1.3

If a random vector W_n converges in distribution to a multinormal distribution with certain mean and variance-covariance, we will write this either:

$$W_n \xrightarrow{d} N(.,.)$$

or

$$W_n \curvearrowright AN(.,.) .$$

6.2 The Joint Generalized DMMSE Estimator

In Chapter 5 we have been considering a linear model of the form $(y, X\beta, \Sigma\sigma^2)$, $|\Sigma| \neq 0$, and we have derived what we call the DMMSE estimator for this model, namely:

$$\tilde{\beta} = k^2X'\left[\sigma^2\Sigma + k^2XX'\right]^{-1}y .$$

In this section we will suppose that we have p general linear models of the form that we have defined above, that is, we will have a set of models of the form:

$$y_j = X\beta_j + \epsilon_j \ , \quad j = 1, \ldots, p \ , \tag{6.2.1}$$

where we will assume

$$\mathrm{Var}[\epsilon_j] = I\sigma^2_{jj}; \mathrm{Cov}[\epsilon_i, \epsilon_j] = I\sigma^2_{ij} \ .$$

The system described by (6.2.1) may be written as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} X & 0 & \ldots & 0 \\ 0 & X & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & X \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{bmatrix} \ . \tag{6.2.2}$$

Using the notation that we defined in Section 6.1, (6.2.2) may be written as

$$\underline{Y} = \underline{\underline{X}}\underline{\beta} + \underline{\epsilon} \tag{6.2.3}$$

where

$$\underline{Y} = [y_1', \ldots, y_p'] \ ; \quad \underline{\underline{X}} = I \otimes X \ ;$$

$$\underline{\epsilon}' = [\epsilon_1', \ldots, \epsilon_p'] \ ; \quad \underline{\beta}' = [\beta_1', \ldots, \beta_p']$$

and

$$\epsilon_i' = [\epsilon_{i1}, \ldots, \epsilon_{in}] \ .$$

Furthermore, we will assume that

$$\underline{V} = \mathrm{Var}[\underline{Y}] = \Sigma \otimes I \; ; \quad E[\underline{\epsilon}_i] = 0$$

so that if we define the vector

$$\tilde{\epsilon}_j' = [\epsilon_{1j}, \ldots, \epsilon_{pj}]$$

we will have

$$E[\tilde{\epsilon}_j'] = 0 \quad \text{and} \quad \mathrm{Var}[\tilde{\epsilon}_j] = \Sigma \; ,$$

and $\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n$ is a set of independent random vectors.

The setting that we have defined above suggests the form for what we shall call the joint DMMSE estimator; namely,

$$\tilde{\beta} = k^2 \underline{X}' [\underline{V} + k^2 \underline{X}\underline{X}']^{-1} \underline{Y}$$

$$= k^2 (I \otimes X')[(\Sigma \otimes I) + k^2 (I \otimes X)(I \otimes X')]^{-1} \underline{Y}$$

where

$$k^2 = \underline{\beta}' \underline{\beta} \; .$$

In the following Theorem an alternative form for our estimator $\tilde{\beta}$ is proposed which will be very useful in future studies.

Theorem 6.2.1

The joint DMMSE estimator is a linear transformation of the joint least squares estimator. That is,

$$\tilde{\underline{\beta}} = k^2[\Sigma \otimes (X'X)^{-1} + k^2 I]^{-1}\hat{\underline{\beta}} \,, \qquad (6.2.4)$$

where

$$\hat{\underline{\beta}} = (I \otimes (X'X)^{-1}X')\underline{Y}$$

is the joint LS estimator.

Proof: The proof follows immediately if we apply a formula proposed in Problem 2.9 by Rao [17, p. 33] to our estimator

$$\hat{\underline{\beta}} = k^2(I \otimes X')[(\Sigma \otimes I) + k^2(I \otimes X)(I \otimes X')]^{-1}\underline{Y}$$

$$= k^2(I \otimes X')[(\Sigma^{-1} \otimes I)$$

$$- (\Sigma^{-1} \otimes I)(I \otimes X)D(I \otimes X')(\Sigma^{-1} \otimes I)$$

$$+ (\Sigma^{-1} \otimes I)(I \otimes X)D(D + C)^{-1}D(I \otimes X')(\Sigma^{-1} \otimes I)]\underline{Y} \,,$$

where

$$D = [(I \otimes X')(\Sigma^{-1} \otimes I)(I \otimes X)]^{-1}$$

$$= [\Sigma \otimes (X'X)^{-1}] \,; \quad C = k^2 I \,.$$

By doing some algebra in the expression above, we get

$$\tilde{\underline{\beta}} = k^2(I \otimes X')[(\Sigma^{-1} \otimes I) - (\Sigma^{-1} \otimes X(X'X)^{-1}X')$$

$$+ (I \otimes X(X'X)^{-1})[\Sigma \otimes (X'X)^{-1} + k^2 I]^{-1}(I \otimes (X'X)^{-1}X')]\underline{Y}$$

$$= k^2[\Sigma \otimes (X'X)^{-1} + k^2 I]^{-1}[I \otimes (X'X)^{-1}X']\underline{Y}$$

$$= k^2[\Sigma \otimes (X'X)^{-1} + k^2 I]^{-1}\hat{\underline{\beta}} \ . \ \square$$

In most of the cases we are faced with Σ unknown. Following Zellner [21], a two-step procedure is proposed which starts the estimation of Σ using the matrix of mean squares and products of the LS residuals:

$$\hat{\Sigma} = \frac{1}{n}\begin{bmatrix} e_1' \\ \vdots \\ e_p' \end{bmatrix}[e_1, \ldots, e_p] = (\hat{\sigma}_{ij}) \ .$$

where

$$e_j = y_j - X\hat{\beta}_j$$

and

$$\hat{\beta}_j = (X'X)^{-1}X'y_j$$

so that

$$\hat{\sigma}_{ij} = (y_i - X\hat{\beta}_i)'(y_j - X\hat{\beta}_j)(1/n)$$

is a consistent estimator of σ_{ij}, the $(i,j)^{\underline{th}}$ element of Σ .

Although the LS estimator $\hat{\beta}_j = (X'X)^{-1}X'y_j$ may be highly variable when $X'X$ is "close" to colinearity; that is, when some of the eigenvalues are "very small relative to the others", the product $X\hat{\beta}_j$ does not depend on the eigenvalues of $X'X$ and, therefore, their

use in this circumstance need not be avoided. Indeed, by using the decomposition of X defined in (2.4) we have that

$$\hat{\beta}_j = V\Lambda^{-1/2}U'y_j \ .$$

However,

$$X\hat{\beta}_j = U\Lambda^{1/2}V'V\Lambda^{-1/2}U'y_j = UU'y_j \ ;$$

which is clearly independent of the eigenvalues of $X'X$, and hence,

$$\hat{\sigma}_{ij} = y_i'[I - UU']y_j$$

is independent of the eigenvalues of $X'X$ itself.

Next, we apply (6.2.4) with Σ replaced by $\hat{\Sigma}$, so that the joint DMMSE estimator becomes:

$$\tilde{\underline{\beta}}^* = [I + \frac{1}{k^2}(\hat{\Sigma} \otimes (X'X)^{-1})]^{-1}\hat{\underline{\beta}} \ . \qquad (6.2.5)$$

6.2.1 Asymptotic Distribution Results For The Joint DMMSE Estimator And The Joint Ridge Regression Estimator

In this section we will obtain some asymptotic results for the joint DMMSE and joint ridge regression estimators.

The following theorem gives the asymptotic distribution of

$$\frac{1}{\sqrt{n}}(\Sigma^{-1} \otimes X')\underline{\epsilon} \ .$$

This will be useful in later work.

Theorem 6.2.1.1

For the model $\underline{Y} = (I \otimes X)\underline{\beta} + \underline{\epsilon}$, assume that

(i) $\frac{1}{n} X'X$ converges, as $n \to \infty$, to a positive definite matrix Q .

(ii) $\|x_i\| \leq t$ for all i , where t is a given number and x_i is the i^{th} column of the X' matrix.

Then,

$$\frac{1}{\sqrt{n}} (\Sigma^{-1} \otimes X')\underline{\epsilon}$$

has a limiting multinormal distribution with mean zero and variance-covariance $(\Sigma^{-1} \otimes Q)$, that is,

$$\frac{1}{\sqrt{n}}(\Sigma^{-1} \otimes X')\underline{\epsilon} \frown AN(0, \Sigma^{-1} \otimes Q) .$$

Proof: Following the lines of the proof of Theorem 8.2 of Theil [20, p. 330] , observe first that

$$\frac{1}{\sqrt{n}} (\Sigma^{-1} \otimes X')\underline{\epsilon} = \frac{1}{\sqrt{n}} (\Sigma^{-1} \otimes I)(I \otimes X')\underline{\epsilon}$$

and moreover,

$$\frac{1}{\sqrt{n}} (I \otimes X')\underline{\epsilon} = \frac{1}{\sqrt{n}} (I \otimes x_1)\tilde{\epsilon}_1 + \ldots + \frac{1}{\sqrt{n}} (I \otimes x_n)\tilde{\epsilon}_n ,$$

where the sum is over a set of independent random vectors. Now let

$$S_n = \frac{1}{n} \sum_{i=1}^{n} x_i x_i' = \frac{1}{n} X'X .$$

Therefore by assumption (ii),

$$\left\| S_n \right\| \leq \frac{1}{n} \sum_{i=1}^{n} \left\| k_i x_i' \right\| \leq t^2$$

and

$$\frac{1}{\sqrt{n}} x_i = o(\frac{1}{\sqrt{n}}) \ .$$

By assumption

$$\mathrm{Var}[(\Sigma^{-1} \otimes I)(I \otimes x_j)\tilde{\epsilon}_j] = (\Sigma^{-1} \otimes x_j x_j')$$

and, furthermore, the characteristic function of

$$(\Sigma^{-1} \otimes I)(I \otimes x_j)\tilde{\epsilon}_j$$

can be expressed (see Cramer [5], p. 21.3) as

$$\phi_{w_j}(t) = 1 - \frac{1}{2} t'(\Sigma^{-1} \otimes x_j x_j')t + o(t't) \ ,$$

where

$$w_j = (\Sigma^{-1} \otimes I)(I \otimes x_j)\tilde{\epsilon}_j \ .$$

Thus,

$$\phi_{\frac{1}{\sqrt{n}} w_j}(t) = 1 - \frac{1}{2} t'(\Sigma^{-1} \otimes \frac{1}{n} x_j x_j')t + o(t't/n) \qquad (6.2.1.1)$$

and from (6.2.1.1) we may conclude

$$\phi_w(t) = \prod_{j=1}^{n} \phi_{\frac{1}{\sqrt{n}} w_j}(t) = \prod_{j=1}^{n} [1 - \frac{1}{2} t'(\Sigma^{-1} \otimes \frac{1}{n} x_j x_j')t$$

$$+ o(t't/n)] \ , \qquad (6.2.1.2)$$

where

$$w = \sum_{j=1}^{n} \frac{1}{\sqrt{n}} w_j .$$

Taking the log of (6.2.1.2), we have

$$\log \phi_w(t) = \sum_{j=1}^{n} \left[1 - \frac{1}{2} t'(\Sigma^{-1} \otimes \frac{1}{n} x_j x_j')t + R_{j,n}\right] ,$$

where for each t , $R_{j,n} = o(\frac{1}{n})$. Therefore, if we take n sufficiently large to make

$$\left| - \frac{1}{2} t'(\Sigma^{-1} \otimes \frac{1}{n} x_j x_j')t + R_{j,n} \right| < 1$$

we will obtain

$$\log \phi_w(t) = - \frac{1}{2} t' \sum_{j=1}^{n} (\Sigma^{-1} \otimes \frac{1}{n} x_j x_j')t + o(1)$$

since for sufficiently large n ,

$$\left| \sum_{j=1}^{n} R_{j,n} \right| \leq \sum_{j=1}^{n} |R_{j,n}| \leq \sum_{j=1}^{n} \frac{\delta}{n} = \delta .$$

So we will get

$$\log \phi_w(t) = - \frac{1}{2} t'(\Sigma^{-1} \otimes \frac{1}{n} \sum_{j=1}^{n} x_j x_j')t + o(1) . \qquad (6.2.1.3)$$

Finally, we will obtain from assumption (i) that (6.2.1.3) converges to

$$- \frac{1}{2} t'(\Sigma^{-1} \otimes Q)t$$

and this corresponds to the multinormal distribution with mean zero and covariance matrix $(\Sigma^{-1} \otimes Q)$. Therefore, by the Continuity Theorem, we may conclude that

$$\frac{1}{\sqrt{n}} (\Sigma^{-1} \otimes X')\underline{\epsilon}$$

is distributed asymptotically as a $N[0, \Sigma^{-1} \otimes Q]$. □

Corollary 6.2.1

Under the assumptions of Theorem 6.2.1.1,

$$\frac{1}{\sqrt{n}} X'\epsilon_j$$

has a limiting multinomial distribution with mean zero and variance-covariance matrix $\sigma_{jj} Q$.

Proof: Premultiply

$$\frac{1}{\sqrt{n}} (I \otimes X')\underline{\epsilon}$$

by $[0, \ldots, 0, I, 0, \ldots, 0]$, where I is located at the j^{th} position and obtain from Theorem 6.2.1.1, by letting

$$w_j = (I \otimes x_j)\tilde{\epsilon}_j ,$$

that

$$\frac{1}{\sqrt{n}} X'\epsilon_j$$

has limiting multinormal distribution with mean zero and variance-covariance $\sigma_{jj} Q$. □

One of the asymptotic distribution results about the joint DMMSE estimator is contained in the following theorem.

Theorem 6.2.1.2

For the model $(\underline{Y}, X\underline{\beta}, \Sigma \otimes I)$, $|\Sigma| \neq 0$, assume that

(i) $\frac{1}{n} X'X$ converges, as $n \to \infty$, to a p.d. matrix Q .

(ii) $\|x_i\| \leq m$, for all i , where m is a given but fixed number and x_i is the $i^{\underline{th}}$ column of X' .

Then, $\sqrt{n}(\tilde{\underline{\beta}}^* - \underline{\beta})$ has limiting multinormal distribution with mean zero and variance-covariance $(\Sigma \otimes Q^{-1})$.

Proof: From the definition of $\tilde{\underline{\beta}}^*$, given by (6.2.5), we obtain

$$\tilde{\underline{\beta}}^* - \underline{\beta} = \left[I + \frac{1}{k^2}(\hat{\Sigma} \otimes (X'X)^{-1})\right]^{-1}(\hat{\Sigma}^{-1} \otimes X'X)^{-1}$$

$$\times (\hat{\Sigma}^{-1} \otimes X'X)\left[I \otimes (X'X)^{-1}X'\right]\underline{Y} - \underline{\beta}$$

$$= \left[(\hat{\Sigma}^{-1} \otimes X'X) + \frac{1}{k^2} I\right]^{-1}(\hat{\Sigma}^{-1} \otimes X')\underline{Y} - \underline{\beta}$$

$$= \left[(\hat{\Sigma}^{-1} \otimes X'X) + \frac{1}{k^2} I\right]^{-1}(\Sigma^{-1} \otimes X')\underline{\beta}$$

$$+ \left[\hat{\Sigma}^{-1} \otimes X'X) + \frac{1}{k^2} I\right]^{-1}(\hat{\Sigma}^{-1} \otimes X')\underline{\epsilon}$$

$$- \left[(\hat{\Sigma}^{-1} \otimes X'X) + \frac{1}{k^2} I\right]^{-1}\left[(\hat{\Sigma}^{-1} \otimes X'X) + \frac{1}{k^2} I\right]\underline{\beta}$$

$$= -\frac{1}{k^2}\left[(\hat{\Sigma}^{-1} \otimes X'X) + \frac{1}{k^2} I\right]^{-1}\underline{\beta}$$

$$+ \left[(\hat{\Sigma}^{-1} \otimes X'X) + \frac{1}{k^2} I\right]^{-1}(\hat{\Sigma}^{-1} \otimes X')\underline{\epsilon} .$$

Hence,

$$\sqrt{n}(\tilde{\underline{\beta}}^{*} - \underline{\beta}) = -\frac{1}{k^2\sqrt{n}} A_n\underline{\beta} + A_n B_n \qquad (6.2.1.4)$$

where

$$A_n = \left[\frac{1}{k^2 n} I + (\hat{\Sigma}^{-1} \otimes \frac{1}{n} X'X)\right]^{-1} ,$$

and

$$B_n = \frac{1}{\sqrt{n}} (\hat{\Sigma}^{-1} \otimes X')\underline{\epsilon} . \qquad (6.2.1.5)$$

Observe that, by defining

$$K_n = \frac{1}{\sqrt{n}} (\Sigma^{-1} \otimes X')\underline{\epsilon} ,$$

we obtain

$$B_n = \frac{1}{\sqrt{n}} (\hat{\Sigma}^{-1} \otimes X')\underline{\epsilon} = \frac{1}{\sqrt{n}} (\Sigma^{-1} \otimes X')\underline{\epsilon}$$

$$+ \frac{1}{\sqrt{n}} [(\hat{\Sigma}^{-1} - \Sigma^{-1}) \otimes X']\underline{\epsilon}$$

$$= K_n + \frac{1}{\sqrt{n}} [(\hat{\Sigma}^{-1} - \Sigma^{-1}) \otimes X']\underline{\epsilon} . \qquad (6.2.1.6)$$

The second vector on the right hand side of Equation (6.2.1.6) has as its $j^{\underline{th}}$ element

$$(\hat{\sigma}^{j1} - \sigma^{j1}) \frac{1}{\sqrt{n}} X'\epsilon_1 + \ldots + (\hat{\sigma}^{jp} - \sigma^{jp}) \frac{1}{\sqrt{n}} X'\epsilon_p . \qquad (6.2.1.7)$$

It has been shown in Corollary 6.2.1.1 that

$$\frac{1}{\sqrt{n}} X'\epsilon_j \cap AN(0, \sigma^2_{jj} Q)$$

and, by the consistency of $\hat{\Sigma}$, which implies that

$$\hat{\sigma}^{j\alpha} - \sigma^{j\alpha} \xrightarrow{P} 0 .$$

We may conclude (by the continuity of the inverse of nonsingular matrices) that

$$(\hat{\sigma}^{j\alpha} - \sigma^{j\alpha}) \frac{1}{\sqrt{n}} X'\epsilon_{\alpha} \xrightarrow{P} 0$$

and, therefore, (6.2.1.7) converges in probability to zero. Hence,

$$\sqrt{n}(\tilde{\beta}^{*} - \underline{\beta}) = \frac{1}{k^2\sqrt{n}} A_n \underline{\beta} + A_n K_n$$

$$+ A_n \frac{1}{\sqrt{n}} [(\hat{\Sigma}^{-1} - \Sigma^{-1}) \otimes X'] \epsilon .$$

Now,

$$A_n \xrightarrow{P} \Sigma \otimes Q^{-1} .$$

Therefore,

$$\frac{1}{k^2\sqrt{n}} A_n \xrightarrow{P} 0 ,$$

and since it has been shown in Theorem 6.2.1.1 that

$$K_n \frown AN(0, \Sigma^{-1} \otimes Q)$$

we may conclude that

$$\sqrt{n}(\tilde{\beta}^{*} - \underline{\beta}) \frown AN(0, \Sigma \otimes Q^{-1}) . \quad \square$$

In Section 5.6, we have found that for the case when we restrict ourselves to the model $(y, X\beta, I\sigma^2)$, the DMMSE estimator is

equivalent to the ridge regression estimator of Hoerl and Kennard; namely,

$$\tilde{\beta}_{RR} = [I + c^{-2}(X'X)^{-1}]^{-1}\hat{\beta}$$

where $\hat{\beta}$ is the OLS estimator.

An obvious extension of the ridge regression estimator to a joint regression estimator is:

$$\tilde{\underline{\beta}}_{RR} = [I + c^{-2}(I \otimes (X'X)^{-1})]^{-1}[I \otimes (X'X)^{-1}X']\underline{Y}$$

$$= [I + c^{-2}(I \otimes (X'X)^{-1})]^{-1}\hat{\underline{\beta}} \; . \qquad (6.2.1.8)$$

In the following theorem the asymptotic distribution of the estimator defined in (6.2.1.8) is presented.

Theorem 6.2.1.3

Under the assumptions of Theorem 6.2.1.2, we have that $\sqrt{n}(\tilde{\underline{\beta}}_{RR} - \underline{\beta})$ has a limiting multinormal distribution with mean zero and variance-covariance $\Sigma \otimes Q^{-1}$; that is,

$$\sqrt{n}(\tilde{\underline{\beta}}_{RR} - \underline{\beta}) \; /) \; AN(0, \Sigma \otimes Q^{-1}) \; .$$

Proof: Following the lines of the proof of Theorem 6.2.1.1, and because of the form of the joint ridge regression estimator, we have

$$\tilde{\underline{\beta}}_{RR} - \underline{\beta} = [(I \otimes X'X) + c^{-2}I]^{-1}(I \otimes X'X)(I \otimes (X'X)^{-1}X')\underline{Y}$$

$$- \underline{\beta} = -c^{2}[(I \otimes X'X) + c^{-2}I]^{-1}\underline{\beta}$$

$$+ [(I \otimes X'X) + c^{-2}I]^{-1}(I \times X')\underline{\epsilon} .$$

Hence,

$$\sqrt{n}(\hat{\underline{\beta}}_{RR} - \underline{\beta}) = - \frac{c^{-2}}{\sqrt{n}} A_{1n}B_{n} + A_{1n}B_{n}$$

where

$$A_{1n} = [(I \otimes \frac{1}{n} X'X) + \frac{c^{-2}}{n} I]^{-1}$$

and

$$B_{1n} = \frac{1}{\sqrt{n}} (I \otimes X')\underline{\epsilon} .$$

Now,

$$A_{1n} \xrightarrow{p} (I \otimes Q^{-1})$$

and

$$\frac{c^{-2}}{\sqrt{n}} A_{1n} \xrightarrow{p} 0 ,$$

and

$$\frac{1}{\sqrt{n}} (I \otimes X')\underline{\epsilon} \bigcap AN(0,\Sigma \otimes Q) ,$$

therefore,

$$\sqrt{n}\ (\hat{\underline{\beta}}_{RR} - \underline{\beta}) \bigcap AN(0,\Sigma \otimes Q^{-1}) . \quad \square$$

From Theorems 6.2.1.2, 6.2.1.3, and a result given by Theil [20, p. 400], we may conclude that the joint LS, joint DMMSE, and the joint ridge regression estimators have the same limiting distribution.

6.3 Comparison Between The Variances of The Joint LS, Joint DMMSE and Joint Ridge Regression Estimators

We have concluded in the last section that the estimators under consideration have the same limiting distribution. The obvious next step is to compare, for small samples, its mean square errors. It was found that the expressions involved were too complicated and that nothing useful could be learned. It is for this reason that in this section we shall only be concerned with the comparison of their variances.

Note first that both the joint DMMSE and joint ridge regression estimators are transformations of the joint LS estimators; indeed,

$$\tilde{\underline{\beta}} = k^2[\Sigma \otimes (X'X)^{-1} + k^2 I]^{-1}\hat{\underline{\beta}}$$

$$= [\frac{1}{k^2}(\Sigma \otimes (X'X)^{-1}) + I]^{-1}\hat{\underline{\beta}}$$

and

$$\tilde{\underline{\beta}}_{RR} = [c^{-2}(I \otimes (X'X)^{-1}) + I]^{-1}\hat{\underline{\beta}} ,$$

where

$$c^{-2} = \frac{\ddot{\gamma}}{k^2} ,$$

and $\bar{\gamma}$ is any generalized variance derived from Σ such that $\gamma_p < \bar{\gamma} < \gamma_1$, (e.g., $\bar{\gamma} = tr[\Sigma]/p$ or $\bar{\gamma} = |\Sigma|^{1/p}$).

Now,

$$Var(\hat{\underline{\beta}}) = \Sigma \otimes (X'X)^{-1} = D \tag{6.3.1}$$

therefore,

$$Var(\tilde{\underline{\beta}}) = ADA \tag{6.3.2}$$

and

$$Var(\tilde{\underline{\beta}}_{RR}) = BDB \ , \tag{6.3.3}$$

where

$$A = \left[I + \frac{1}{k^2}(\Sigma \otimes (X'X)^{-1})\right]^{-1} \ , \tag{6.3.4}$$

and

$$B = \left[I + \frac{\bar{\gamma}}{k^2}(I \otimes (X'X)^{-1}\right]^{-1} \ . \tag{6.3.5}$$

6.3.1 Comparison Between $Var(\hat{\underline{\beta}})$ and $Var(\tilde{\underline{\beta}}_{RR})$

The criterion that we will be using to compare variance-covariance matrices is the one we have defined in Chapter 5; that is, we shall say

$$Var[\hat{\underline{\beta}}] \geq Var[\tilde{\underline{\beta}}_{RR}]$$

if, and only if, $Var[\hat{\underline{\beta}}] - Var[\tilde{\underline{\beta}}]$ is n.n.d. In the next theorem we will present such a comparison.

Theorem 6.3.1.1

For the variance-covariance matrices of the joint LS and joint ridge regression estimators, the following inequality holds:

$$\mathrm{Var}[\hat{\beta}] \geq \mathrm{Var}[\tilde{\beta}_{RR}] \ .$$

Proof: By definition, and from (6.3.1) and (6.3.3),

$$\mathrm{Var}[\hat{\beta}] \geq \mathrm{Var}[\tilde{\beta}_{RR}]$$

if, and only if,

$$D \geq BDB \ ,$$

This occurs if, and only if,

$$\ell' D \ell \geq \ell' DBD \ \ell, \quad \text{for all} \quad \ell,$$

since BDB is p.d.; this happens if, and only if,

$$\frac{\ell' D \ell}{\ell' BDB \ell} \geq 1 \quad \text{for all} \quad \ell \ ,$$

or equivalently,

$$\inf_{\ell} \frac{\ell' D \ell}{\ell' BDB \ell} \geq 1 \ ,$$

using the result in Problem 22 in Rao [17, p. 74].

This last inequality holds if, and only if,

$$Ch_m[DB^{-1}D^{-1}B^{-1}] \geq 1 \ . \qquad (6.3.1.1)$$

Now let $P = I \otimes (X'X)^{-1}$ and observe that the following relations holds:

$$DPD^{-1} = P;\ PD^{-1} = D^{-1}P\ ;\ DP = PD\ . \qquad (6.3.1.2)$$

Using the above results, we find that (6.3.1.1) holds if, and only if,

$$Ch_m\left[D(I + \frac{\bar{\gamma}}{k^2} P)D^{-1}(I + \frac{\bar{\gamma}}{k^2}P)\right] = Ch_m\left[(I + \frac{\bar{\gamma}}{k^2} P)^2\right] \geq 1\ . \qquad (6.3.1.3)$$

By writing

$$P = C\underline{\Lambda}^{-1}C' = (\underline{X}'\underline{X})^{-1}\ ,$$

where

$$CC' = C'C = I\ ,$$

we see that (6.3.1.3) holds, if, and only if,

$$Ch_m\left[(I + \frac{\bar{\gamma}}{k^2}\underline{\Lambda}^{-1})^2\right] \geq 1$$

and this holds, if, and only if,

$$1 + \frac{\bar{\gamma}}{k^2}\frac{1}{\underline{\lambda}_{pm}} \geq 0$$

where $\underline{\lambda}_{pm}$ is the smallest eigenvalue of P. Hence, the inequality holds if, and only if,

$$\frac{\bar{\gamma}}{k^2\underline{\lambda}_{pm}} \geq 0$$

which, by assumption, is true. □

6.3.2 Comparison Between $\mathrm{Var}(\hat{\underline{\beta}})$ and $\mathrm{Var}(\tilde{\underline{\beta}})$

The comparison between $\mathrm{Var}(\hat{\underline{\beta}})$ and $\mathrm{Var}(\tilde{\underline{\beta}})$ will be presented in the following theorem.

Theorem 6.3.2.1

For the variance-covariance matrix of the joint LS and joint DMMSE estimators, the following inequality holds:

$$\mathrm{Var}(\hat{\underline{\beta}}) \geq \mathrm{Var}(\tilde{\underline{\beta}}) \ .$$

Proof: Using agruments similar to those in Theorems 6.3.1.1, 6.3.2, and 6.3.3, we have that

$$\mathrm{Var}(\hat{\underline{\beta}}) - \mathrm{Var}(\tilde{\underline{\beta}}) = D - ADA$$

is n.n.d. if, and only if,

$$\mathrm{Ch}_m(DA^{-1}D^{-1}A^{-1}) \geq 1 \ ,$$

where A and D have been defined in (6.3.4) and (6.3.1), respectively. The above inequality holds if, and only if,

$$\mathrm{Ch}_m\left[\left(\frac{1}{k^2} D + I\right)^2\right] \geq 1$$

and this happens if, and only if,

$$\frac{\delta_{pm}}{k^2} \geq 0 \ , \qquad (6.3.2.1)$$

where δ_{pm} is the smallest eigenvalue of the D matrix. By assumption (6.3.2.1) is true. □

6.3.3 Comparison Between $\mathrm{Var}(\tilde{\beta})$ and $\mathrm{Var}(\tilde{\beta}_{RR})$

In this section we will be concerned with the comparison between the variances of the joint DMMSE and ridge regression estimators. This will be elaborated in the following theorem.

Theorem 6.3.3.1

The variance-covariance matrices of the joint DMMSE and ridge regression estimators are not comparable in the defined sense.

Proof: The proof of this theorem will be done by contradiction. We will assume first that

$$\mathrm{Var}[\tilde{\beta}_{RR}] - \mathrm{Var}[\tilde{\beta}] = BDB - ADA \qquad (6.3.3.1)$$

is a n.n.d. matrix, where D, A and B have been defined in (6.3.1) and (6.3.4), respectively. Now, by similar argument to that in Theorem 6.3.1.1, we have that (6.3.3.1) is true if, and only if,

$$Ch_m[BDB(ADA)^{-1}] \geq 1$$

and this is true if, and only if,

$$Ch_m[(D^{-1} + \frac{\bar{\gamma}}{k^2} D^{-1}P)(I + \frac{\bar{\gamma}}{k^2} P)^{-1}(I + \frac{1}{k^2} D)(D^{-1} + \frac{1}{k^2} I)] \geq 1 . \qquad (6.3.3.2)$$

By using the relations between P and D given in (6.3.1.2), we have that (6.3.3.2) holds if, and only if,

$$Ch_m[(I + \frac{1}{k^2} D)(I + \frac{\bar{\gamma}}{k^2} P)^{-2}(I + \frac{1}{k^2} D)] \geq 1 .$$

Equivalently,

$$Ch_m[A^{-2}B^2] \geq 1 \ . \qquad (6.3.3.2)$$

Now, using the definitions given in (2.5), and letting $\Sigma = E\Lambda E'$, where $E'E = EE' = I$, we have

$$P = I \times V\Lambda^{-1}V' = (I \otimes V)(I \otimes \Lambda^{-1})(I \otimes V') \ ;$$

from this

$$I + \frac{\bar{\gamma}}{k^2} P = (I \otimes V)(I \otimes V')$$

$$+ (I \otimes V)(I \otimes \frac{\bar{\gamma}}{k^2}\Lambda^{-1})(I \otimes V')$$

$$= (I \otimes V)[(I \otimes I) + (I \otimes \frac{\bar{\gamma}}{k^2}\Lambda^{-1})](I \otimes V')$$

$$= (I \otimes V)(I \otimes M)(I \otimes V')$$

where

$$M = \text{diag}(1 + \frac{\bar{\gamma}}{k^2\lambda_i}) = \text{diag}(\frac{\bar{\gamma} + k^2\lambda_i}{k^2\lambda_i}) \ .$$

Hence,

$$B^2 = (I + \frac{\bar{\gamma}}{k^2} P)^{-2} = (I \otimes V)(I \otimes R)(I \otimes V') \ ,$$

where

$$R = \text{diag}(\frac{k^2\lambda_i}{\bar{\gamma} + k^2\lambda_i})^2 \ .$$

Now,

$$A^{-1} = \left[I + \frac{1}{k^2} D\right] = I + \frac{1}{k^2} (E\Gamma E' \otimes V\Lambda^{-1}V')$$

$$= (E \otimes V)(E' \otimes V')$$

$$+ (E \otimes V)(\Gamma \otimes \frac{1}{k^2} \Lambda^{-1})(E' \otimes V')$$

$$= (E \otimes V)(I + \Gamma \otimes \frac{1}{k^2} \Lambda^{-1})(E' \otimes V') \ .$$

Therefore,

$$A^{-2} = (E \otimes V)(I + \Gamma \otimes \frac{1}{k^2} \Lambda^{-1})^2(E' \otimes V') \ .$$

Thus,

$$Ch_m[A^{-2}B^2] = Ch_m[(E \otimes V)(I + \Gamma \otimes \frac{1}{k^2} \Lambda^{-1})^2$$

$$x\ (E' \otimes V')(I \otimes V)(I \otimes R)(I \otimes V')]$$

$$= Ch_m[(I \otimes R)(I + \Gamma \otimes \frac{1}{k^2} \Lambda^{-1})^2]$$

But,

$$(I \otimes R)(I + \Gamma \otimes \frac{1}{k^2} \Lambda^{-1})^2 = \text{diag}(\frac{k^2\lambda_i + \gamma_k}{\bar{\gamma} + k^2\lambda_i})^2 \ .$$

Therefore, (6.3.3.2) holds if, and only if,

$$\gamma_m \geq \bar{\gamma}$$

and we arrive at a contradiction.

Assume now that

$$\text{Var}[\tilde{\beta}] - \text{Var}[\tilde{\beta}_{RR}] = ADA - BDB \qquad (6.3.3.3)$$

is a n.n.d. matrix. Using the relationships between P and D given in (6.3.1.2), we may obtain that (6.3.3.3) is true if, and only if,

$$Ch_m[ADAB^{-1}D^{-1}B^{-1}] = Ch_m[ADAB^{-2}D^{-1}]$$

$$= Ch_m[D^{-1}ADAB^{-2}] \geq 1 .$$

But,

$$D^{-1}ADAB^{-2} = D^{-1}[I + \frac{1}{k^2} D]^{-1}D[I + \frac{1}{k^2} D]^{-1}B^{-2}$$

$$= [I + \frac{1}{k^2} D]^{-1}[I + \frac{1}{k^2} D]^{-1}B^{-2} = A^2B^{-2} .$$

Furthermore,

$$B^{-1} = (I \otimes V)(I + I \otimes \frac{\bar{\gamma}}{k^2} \Lambda^{-1})(I \otimes V') .$$

Thus,

$$B^{-2} = (I \otimes V)(I \otimes W)(I \otimes V')$$

where

$$W = \text{diag}(\frac{k^2\lambda_i + \bar{\gamma}}{k^2\lambda_i})^2 .$$

Now

$$A = [I + E\Gamma E' \otimes V \frac{1}{k^2} \Lambda^{-1}V']^{-1}$$

$$= [(E \otimes V)(I + \Gamma \otimes \frac{1}{k^2} \Lambda^{-1})(E' \otimes V')]^{-1} .$$

Therefore,

$$A^2 = (E \otimes V)(I + \Gamma \otimes \frac{1}{k^2} \Lambda^{-1})^{-2}(E' \otimes V') \ .$$

Hence,

$$Ch_m[ADAB^{-1}D^{-1}B^{-1}] = Ch_m[D^{-1}ADAB^{-2}] = Ch_m[A^2B^{-2}]$$

$$= Ch_m[(E \otimes V)(I + \Gamma \otimes \frac{1}{k^2} \Lambda^{-1})^{-2}$$

$$x \ (E' \otimes V')(I \otimes V)(I \otimes W)(I \otimes V')]$$

$$= Ch_m[(E \otimes I)(I + \Gamma \otimes \frac{1}{k^2} \Lambda^{-1})^{-2}$$

$$x \ (E' \otimes W)]$$

$$= Ch_m[(I \otimes W)(I + \Gamma \otimes \frac{1}{k^2} \Lambda^{-1})^{-2}] \ .$$

However, since

$$(I \otimes W)(I + \Gamma \otimes \frac{1}{k^2} \Lambda^{-1})^{-2} = \text{diag}(\frac{k^2\lambda_i + \bar{\gamma}}{k^2\lambda_i + \gamma_k})^2$$

we have that

$$Ch_m[ADAB^{-1}D^{-1}B^{-1}] = Ch_m[(I \otimes W)(I + \Gamma \otimes \frac{1}{k^2} \Lambda^{-1})^{-2}] \geq 1$$

if, and only if,

$$\frac{k^2\lambda_m + \bar{\gamma}}{k^2\lambda_m + \gamma_1} \geq 1 \ .$$

This last inequality holds if, and only if,

$$\bar{\gamma} \geq \gamma_1 \qquad (6.3.3.4)$$

where γ_1 is the largest eigenvalue of $\Sigma = E\Gamma\Gamma'$.

Clearly (6.3.3.4) is a contradiction.

In review, what we have is the following:

(i) If we assume $\mathrm{Var}(\tilde{\underline{\beta}}_{RR}) - \mathrm{Var}(\tilde{\underline{\beta}})$ is n.n.d., we arrive at a contradiction.

(ii) If we assume $\mathrm{Var}(\tilde{\underline{\beta}}) - \mathrm{Var}(\tilde{\underline{\beta}}_{RR})$ is n.n.d., we also arrive at a contradiction.

Therefore, from (i) and (ii) we may conclude that $\mathrm{Var}(\tilde{\underline{\beta}})$ and $\mathrm{Var}(\tilde{\underline{\beta}}_{RR})$ are not comparable in the sense that we have defined. □

From Theorem 6.3.1.1 and Theorem 6.3.2.1, we may conclude that for small samples, both the joint DMMSE and ridge regression estimators have "smaller" variance-covariance matrices than the joint LS estimators. Unfortunately, from Theorem 6.3.3.1 we shall conclude that the variance-covariance matrices of the joint DMMSE and joint ridge regression estimators are not comparable in the sense we have defined before.

7. PROBLEMS FOR FURTHER RESEARCH

In this section we will present some ideas and problems that arose in the course of this research, and which warrant further research.

Problem 7.1

Considering the setting in Chapter 4, Obenchain[1] has proposed that instead of using the Euclidean length of the parameter vector β, we could define

$$\mathcal{B}_{k,Q} = \{\beta \mid \beta'Q\beta = k^2\}.$$

This is, consider β's whose squared length in the norm of a p.d. Q is k^2, and proceed along the same lines of Chapter 4 with this modification.

The same modification could be introduced in Chapter 5.

Problem 7.2

Under the setting of Chapter 5, try to find (if it exists)

$$\sup_{\beta \in \mathcal{B}_k} M(Ay + b, P'\beta; \beta, \Sigma\sigma^2) \tag{7.1}$$

or prove that

$$\sup_{\beta \in \mathcal{B}_k} M(Ay+b, P'\beta; \beta, \Sigma\sigma^2) \geq \sup_{\beta \in \mathcal{B}_k} M(Ay, P'\beta; \beta, \Sigma\sigma^2) \tag{7.2}$$

with equality holding for $b = 0$.

---

[1] Obenchain, R. L., September 13, 1976. Personal communication. Professor. Department of Applied Statistics, Bell Laboratories, Holmdel, N.J. 07733.

If (7.2) is not true, and if (7.1) exists, use it to obtain the DMMSE estimator.

Problem 7.3

Compare the MSE matrices for the joint DMMSE and joint ridge regression estimators.

Problem 7.4

Let $\hat{\theta}$ be a biased estimator of θ. Using $\hat{\theta}$, construct an at least $(1 - \alpha)$ confidence interval for θ.

Problem 7.5

For the matrix form (Chapter 5) of the minimax MSE criterion, the optimum estimator for the class of OLS estimators calculated subject to false restrictions has not been obtained.

8. LIST OF REFERENCES

1. Baranchik, A. J. 1964. Multiple regression and estimation of the mean of a multivariate normal distribution. Tech. Report No. 51, Dept. of Statistics, Stanford Univ., Stanford, Calif.

2. Chapman, D. 1974. An extension and investigation of the properties of the ridge regression. Unpubl. Ph.D. Thesis, Dept. of Statistics, N.C. State Univ., Raleigh, N.C. Univ. Microfilms, Ann Arbor, Mich.

3. Chipman, J. S. 1964. On least squares with insufficient observations. J. Amer. Stat. Assoc. 59:1078-1111.

4. Conniffe, D. and J. Stone. 1973. A critical view of ridge regression. The Statistician 22:181-187.

5. Cramer, H. 1966. Mathematical Methods of Statistics. Princeton Univ. Press, Princeton, N.J.

6. Foster, M. 1961. An application of the Wiener Kolomogorov smoothing theory of matrix inversion. J. Soc. Ind. and App. Math. 9:387-392.

7. Goldstein, M. and A. F. M. Smith. 1974. Ridge type estimators for regression analysis. J. Royal Stat. Soc., Ser. B, 36:284-291.

8. Hoerl, A. E. 1962. Application of ridge analysis to regression problems. Chem. Eng. Progr. 58:54-59.

9. Hoerl, A. E. and R. W. Kennard. 1970. Ridge regressions: Applications to nonorthogonal problems. Technometrics 12:55-67, 69-82.

10. James, W. and C. Stein. 1960. Estimation with Quadratic Loss. Proc. of Fourth Berkeley Symp. on Math. Stat. and Prob. Univ. of Calif., Berkeley, Calif.

11. Lancaster, P. 1969. Theory of Matrices. Academic Press, Inc., New York City, N. Y.

12. Marcus, M. and H. Minc. 1969. Survey of Matrix Theory and Matrix Inequalities. Allyn and Bacon, Boston, Mass.

13. Marquardt, D. W. 1970. Generalized inverse, ridge regression, biased linear estimation and nonlinear estimation. Technometrics 72:591-612.

14. Mayer, L. S. and T. A. Willke. 1973. On biased estimation in linear models. Technometrics 15:497-508.

15. Pringle, R. M. and A. A. Rayner. 1971. Generalized Inverse Matrices With Applications to Statistics. Hafner Publ. Co., New York City, N.Y.

16. Rao, C. R. 1971. Unified theory of linear estimation. Sankhya, Ser. A, 33:371-394.

17. Rao, C. R. 1973. Linear Statistical Inference and Its Applications. (Second Edition). John Wiley and Sons, Inc., New York City, N.Y.

18. Toro-Viscarrondo, C. and T. D. Wallace. 1968. A test of the mean square error criterion for restrictions in linear regression. J. Amer. Stat. Assoc. 63:558-572.

19. Sclove, S. L. 1968. Improved estimators for coefficients in linear regression. J. Amer. Stat. Assoc. 63:596-606.

20. Theil, H. 1971. Principles of Econometrics. John Wiley and Sons, Inc., New York City, N. Y.

21. Zellner, A. 1962. An efficient method of estimating seemingly unrelated regressions and test for aggregation bias. J. Amer. Stat. Assoc. 57:348-368.