Ridge And Related Biased

Estimators In Linear Regression Models

Thomas F. Willis

A Thesis

in

The Department

of

Mathematics

Presented in Partial Fulfillment of the Requirements

for the degree of Master of Science

Concordia University

Montreal, Quebec, Canada

April 1979

© Thomas F. Willis, 1979

ABSTRACT

Ridge And Related Biased Estimators In Linear Regression Models

Thomas F. Willis

This thesis presents a survey of the ridge regression literature. The ridge estimators were originally put forward by Hoerl (1962), (1964) to cope with the effects of severe multicollinearity between the dependent variables of the general linear model. A number of properties and criticisms of the estimators are detailed in this thesis. In addition to their classical formulation, the ridge estimators are considered within a Baysian framework. The ridge estimators are compared with other biased estimators which have been proposed to counter the effects of an ill-conditioned $X'X$ matrix.

In addition to providing a review of the existing ridge estimators, a combined estimator of the form:

$$\hat{\beta}_2^*(k,r) = P_r(\Lambda_r + kI_r)^{-1}P_r'X'Y$$

is studied in this thesis. $\hat{\beta}_2^*(k,r)$ is a combination of the ordinary ridge estimator proposed by Hoerl (1962), (1964) and Marquardt's (1970) generalized least squares estimator. It is proposed that this combined estimator be employed in situations where some of the eigenvalues for the $X'X$ matrix are assumed to be equal to zero and others close to zero. A number of properties of the combined estimator $\hat{\beta}_2^*(k,r)$ are developed. The results of a series of simulation experiments are also presented to illustrate the potential usefulness of the estimator.

ACKNOWLEDGEMENTS

The author would like to thank Dr. T.D. Dwivedi for his encouragement and suggestions in the preparation of this thesis.

TABLE OF CONTENTS

Chapter

1 Introduction 1

2 The Ridge Estimator 15

3 The Relationships Between The Ridge Estimator And Other Biased Estimators 55

4 Criticisms Of The Ridge Estimator 98

5 Generalizations Of The Ridge Estimator 109

6 Tests Of Hypotheses And Confidence Intervals For Ridge Regression 141

7 The Bayesian Interpretation Of Ridge Regression .. 158

8 A Combined Estimator 166

Appendix - A Convergence Theorem 186

References 189

Chapter 1

Introduction

Consider the general linear model:

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

where: Y is a n x 1 vector of n observations of the variable to be explained or predicted; X is a n x p matrix of n observations on p explanatory or control variables; β is a p x 1 vector of p coefficients; and ε is a n x 1 vector of unobservable disturbances. The unobservable disturbances are assumed to satisfy:

$$E(\varepsilon) = 0 \qquad \text{Var}(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I_n \tag{1.2}$$

In some situations, it will be assumed that the unobservable disturbances are normally distributed. Further, assume that the unobservable disturbances are independent of the values of X. Unless specified otherwise, it will be assumed that the design matrix will be of full rank.

The ordinary least squares or Gauss-Markov solution is the most widely applied procedure for estimating the unknown vector of parameters β in model (1.1). This estimator is obtained by choosing that estimator b of β which minimizes the resulting residual sum of squares. Suppose that X is a design matrix and Y the corresponding vector of observations. The sum of squares of the residual function is given by:

$$\begin{aligned}\phi(b) &= (Y - Xb)'(Y - Xb) \\ &= Y'Y - Y'Xb - b'X'Y + b'X'Xb \\ &= Y'Y - 2b'X'Y + b'X'Xb \end{aligned} \tag{1.3}$$

The critical value of $\phi(b)$ satisfies:

$$\frac{d}{db}\phi(b) = -2X'Y + 2X'Xb = 0 \tag{1.4}$$

or:

$$X'Xb = X'Y \tag{1.5}$$

If X is of full rank so that X'X is invertible, then

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{1.6}$$

is the critical value of $\Phi(b)$. When X is of full rank, the Hessian matrix for $\Phi(b)$:

$$\frac{d^2}{db^2}\ \Phi(b) = 2X'Y \tag{1.7}$$

is a positive definite symmetric matrix so that $\hat{\beta}$ must minimize the sum of squares of the residuals function. Throughout the remainder of this thesis, the symbol $\hat{\beta}$ will be reserved for the ordinary least squares estimator of β in model (1.1).

Suppose that in addition to the assumptions (1.2), the unobservable disturbances are assumed to be normally distributed. The joint probability density function for the unobservable disturbances would be:

$$f(\varepsilon|\sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\ \varepsilon'\varepsilon\right\} \tag{1.8}$$

If the transformation:

$$Y = X\beta + \varepsilon \tag{1.9}$$

is made, the resulting joint probability density function becomes:

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

$$= L(\beta,\sigma^2|Y,X) \tag{1.10}$$

Maximizing the likelihood function $L(\beta,\sigma^2|Y,X)$ produces the maximum likelihood estimators:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{1.11}$$

and:

$$\hat{\sigma}^2 = \frac{(Y - X\beta)'(Y - X\beta)}{n} \tag{1.12}$$

for β and σ^2. Therefore, the least squares estimator $\hat{\beta}$ is also the maximum likelihood estimator when the unobservable disturbances are assumed to follow a normal distribution.

In many situations, it is convenient to standardize the observations before estimating the parameters. Assume that the model defined by (1.1) includes a constant term. Suppose that the sample mean and standard deviation are calculated for the dependent variable and each of the explanatory variables. The observations corresponding to each variable can be standardized by subtracting the variable's sample mean from each observation and dividing by the standard deviation. This transformation of the observations results in a $X'X$ matrix equal to the sample correlation matrix for the explanatory variables and a $X'Y$ vector containing the correlations between the dependent variable and each of the explanatory variables. Let $\bar{y}$, s_y, $\bar{x}_i$ and s_i represent the sample means and standard deviations for the dependent variable and i'th explanatory variable respectively. Suppose that $\tilde{\beta}$ denotes the estimated coefficient vector obtained using the standardized model. Estimates of the parameters for the non-standardized model may be calculated according to:

$$\tilde{\beta}_i^* = \begin{cases} \bar{y} - s_y \sum_{i=1}^{p} \frac{\tilde{\beta}_i}{s_i} \bar{x}_i & \text{for } i = 0 \\ \frac{s_y}{s_i} \tilde{\beta}_i, & \text{for } i = 1,2,3, \ldots p \end{cases} \qquad (1.13)$$

If the general model does not contain a constant term, the observations for each variable can be standardized by dividing each observation by the corresponding sample standard deviation.

Standardizing the data as described above leads to a form of

the model whose estimated parameters lend themselves to more straight-forward interpretations in terms of the correlations between the different variables in the model. In addition, Marquardt and Snee (1975) emphasized that the variable transformations remove the effects of correlations between the constant term and the explanatory variables from the estimation procedure. They pointed out that the presence of significant correlations between these variables should be particularly distressing. The resultant nonessential ill-conditioning would not be due to any real deficiency in the data set but rather the result of an arbitrary assignment of origins to the scales employed to express the explanatory variables. McCabe (1978) pointed out that the transformations defined by (1.13) impose a constraint upon the predicted values of Y. He noted that the estimated model will predict $\bar{y}$ for the vector $(\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_p)$ irregardless of the estimator employed to estimate β. McCabe (1978) suggested that it might be more meaningful to constrain the prediction of Y at some other point in some situations.

A number of properties for $\hat{\beta}$ follow directly from the assumptions for the unobservable disturbances. $\hat{\beta}$ is an unbiased estimator of β since:

$$\begin{aligned} E(\hat{\beta}) &= (X'X)^{-1}X'E(Y) \\ &= (X'X)^{-1}X'E(X\beta + \varepsilon) \\ &= \beta + (X'X)^{-1}X'E(\varepsilon) \\ &= \beta \end{aligned} \tag{1.14}$$

The variance-covariance matrix for $\hat{\beta}$ is given by:

$$\begin{aligned} Var(\hat{\beta}) &= E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \\ &= E((X'X)^{-1}X'\varepsilon)((X'X)^{-1}X'\varepsilon)' \\ &= (X'X)^{-1}X'E(\varepsilon\varepsilon')X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned} \tag{1.15}$$

Suppose that there exists another unbiased estimator of β with a smaller variance than $\hat{\beta}$. The estimator must be of the form:

$$\hat{\beta}^* = \hat{\beta} + B \tag{1.16}$$

with a variance-covariance matrix:

$$\begin{aligned} Var(\hat{\beta}^*) &= E(\hat{\beta}^* - \beta)(\hat{\beta}^* - \beta)' \\ &= E(\hat{\beta} + B - \beta)(\hat{\beta} + B - \beta)' \\ &= E((X'X)^{-1}X'\varepsilon + B)((X'X)^{-1}X'\varepsilon + B)' \\ &= E((X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}) + E((X'X)^{-1}X'\varepsilon B') \\ &\quad + E(B\varepsilon'X(X'X)^{-1}) + E(BB') \\ &= Var(\hat{\beta}) + (X'X)^{-1}X'E(\varepsilon B') + E(B\varepsilon')X(X'X)^{-1} \\ &\quad + E(BB') \end{aligned} \tag{1.17}$$

Since the unobservable disturbances are assumed to be independent of the values of X, the estimators $\hat{\beta}$ and $\hat{\beta}^*$ are independent of ε. It therefore follows that:

$$\begin{aligned} E(B\varepsilon') &= E((\hat{\beta}^* - \hat{\beta})\varepsilon') \\ &= E(\hat{\beta}^* - \hat{\beta})E(\varepsilon') \\ &= 0 \end{aligned} \tag{1.18}$$

The expression for the variance-covariance matrix of $\hat{\beta}^*$ reduces to:

$$\begin{aligned} Var(\hat{\beta}^*) &= Var(\hat{\beta}) + E(BB') \\ &\geq Var(\hat{\beta}) \end{aligned} \tag{1.19}$$

with equality if and only if B is a constant vector of zeros. In other words, any unbiased estimator of β different from the ordinary least squares estimator has a larger variance. Therefore, $\hat{\beta}$ is the minimum variance unbiased estimator (MVU) of β in model (1.1).

The efficiency of any estimator b of β may be quantified by considering the squared distance between b and β defined by:

$$L^2(b) = (b - \beta)'(b - \beta) \tag{1.20}$$

The expected value of $L^2(b)$:

$$\begin{aligned} EL^2(b) &= E(b - \beta)'(b - \beta) \\ &= E(b - Eb)'(b - Eb) + (Eb - \beta)'(Eb - \beta) \\ &= \sum_{i=1}^{p} Var(b_i) + \sum_{i=1}^{p} (Eb_i - \beta_i)^2 \end{aligned} \tag{1.21}$$

is usually denoted as the mean squared error (MSE) for the estimator. $EL^2(b)$ measures not only the dispersions of the individual parameters but also their squared biases.

In the case of the least squares estimator:

$$
\begin{aligned}
EL^2(\hat{\beta}) &= E(\hat{\beta} - \beta)'(\hat{\beta} - \beta) \\
&= E((X'X)^{-1}X'\varepsilon)'((X'X)^{-1}X'\varepsilon) \\
&= E(\varepsilon'X(X'X)^{-2}X'\varepsilon) \\
&= tr(X(X'X)^{-2}X\sigma^2 I_n) \\
&= \sigma^2 tr(X'X)^{-1} \qquad (1.22)
\end{aligned}
$$

since the unobservable disturbances are assumed to satisfy conditions (1.2). If X is of full rank, the $X'X$ matrix will be a positive definite symmetric matrix with p eigenvalues, $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p$, greater than zero. Further, there exists an orthogonal matrix P such that:

$$P'(X'X)P = \Lambda \qquad (1.23)$$

where Λ is a diagonal matrix with the eigenvalues λ_i as it's diagonal elements. Since multiplication is commutative within the trace operator:

$$
\begin{aligned}
EL^2(\hat{\beta}) &= \sigma^2 tr((X'X)^{-1}P'P) \\
&= \sigma^2 tr(P'(X'X)^{-1}P) \\
&= \sigma^2 tr(\Lambda^{-1}) \\
&= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \qquad (1.24)
\end{aligned}
$$

In the same manner,

$$
\begin{aligned}
Var(L^2(\hat{\beta})) &= Var((\hat{\beta} - \beta)'(\hat{\beta} - \beta)) \\
&= Var(\varepsilon'X(X'X)^{-2}X'\varepsilon) \\
&= 2tr(X(X'X)^{-2}X'\sigma^2 I_n)^2
\end{aligned}
$$

(1.) Searle, S.R.: Linear Models. New York: John Wiley & Sons, Inc., 1971, p. 55.

(2.) Ibid., p. 57.

$$= 2\sigma^4 tr((X'X)^{-2})$$

$$= 2\sigma^4 tr(\Lambda^{-2})$$

$$= 2\sigma^4 \sum_{i=1}^{p} \frac{1}{\lambda_i^2} \qquad (1.25)$$

If the explanatory or control variables exhibit small interdependencies so that the design matrix is nearly orthogonal, the least squares estimates may not be too bad. However, in reality, it is often found that there are large degrees of multicollinearity between the explanatory or control variables. For example, Newhouse and Oman (1971) described problems in estimating the budgetary costs for various sizes of volunteer armies. They suggested that it is necessary to know the effects of relative military-civilian wages for a given constant unemployment rate. Data at one point in time from different regions has been employed in the estimation of these effects. However, Newhouse and Oman (1971) pointed out that if high unemployment occurs in the same areas as high ratios of military to civilian pay, it is hard to estimate the effects of adjusting only the military to civilian pay ratios.

If the non-orthogonality is at all severe, the $X'X$ matrix will be ill-conditioned and have some eigenvalues close to zero. Since both $EL^2(\hat{\beta})$ and $Var(L^2(\hat{\beta}))$ are dependent upon the values of λ_i^{-1}, it can be seen that the estimates $\hat{\beta}_i$ of the individual parameters may fluctuate wildly. It is even possible that some parameters will have the wrong signs. Using a Euclidean norm, the length of the ordinary least squares estimator is defined to be:

$$\|\hat{\beta}\| = (\hat{\beta}'\hat{\beta})^{1/2}, \qquad (1.26)$$

so that the expected squared length of the least squares estimator becomes:

$$E(\|\hat{\beta}\|^2) = E(\hat{\beta}'\hat{\beta})$$

$$= \beta'\beta + E(\hat{\beta} - \beta)'(\hat{\beta} - \beta)$$

$$= \beta'\beta + \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \quad . \qquad (1.27)$$

Therefore, it is apparent that as the $X'X$ matrix becomes more ill-conditioned, the ordinary least squares estimates will tend to become too large in absolute value. Vinod (1976b) examined two economy of scale functions which exhibited serious multicollinearity in their exogenous variables to show that ordinary least squares solutions may be misleading.

Swindel (1974) constructed a series of examples to illustrate the kind of instability that can result when the design matrix is ill-conditioned. Based upon his examples, a series of simulations of the model:

$$Y = X\begin{pmatrix}1\\1\end{pmatrix} + \varepsilon \qquad (1.28)$$

$$\varepsilon \sim N(0.0, \sigma^2 I_3)$$

were constructed. The four design matrices:

$$X_1 = \begin{pmatrix}1.0000 & 0.0000\\0.0000 & 1.0000\\0.0000 & 0.0000\end{pmatrix} \quad X_2 = \begin{pmatrix}0.9578 & 0.2873\\0.2873 & 0.9578\\0.0000 & 0.0000\end{pmatrix} \qquad (1.29)$$

$$X_3 = \begin{pmatrix}0.8944 & 0.4472\\0.4472 & 0.8944\\0.0000 & 0.0000\end{pmatrix} \quad X_4 = \begin{pmatrix}0.8000 & 0.6000\\0.6000 & 0.8000\\0.0000 & 0.0000\end{pmatrix}$$

were employed in the simulations. Sequences of one thousand observations were generated for each design matrix. Tables 1 and 2 summarize the simulation results when σ^2 is taken to be 0.25 and 0.36 respectively. The theoretical means for the sum of squares of the ordinary least squares estimates and variances for the sum of squares of the deviations are compared with the simulation results. A breakdown of the distribution of errors in the signs of the simulated ordinary least squares estimates is provided.

Table 1 - A Summary Of N = 1,000 Simulations Of The Least Squares Estimates For The Model:

$$Y = X_i \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \varepsilon, \qquad \varepsilon \sim N(0.0, 0.25 I_3)$$

		Design Matrix			
		X_1	X_2	X_3	X_4
	Theoretical Values				
1.	Eigenvalues For $X_i'X_i$: a) λ_1	1.00000	1.55027	1.79989	1.96000
	b) λ_2	1.00000	0.44957	0.19999	0.04000
2.	Sum Of Squares Of The $\hat{\beta}_i$	2.00000	2.00000	2.00000	2.00000
3.	Theoretical Sum Of Squares Of The $\hat{\beta}_i$	2.50000	2.71735	3.38897	8.37756
4.	Theoretical Variance For $L^2(\hat{\beta})$	0.25000	0.67048	3.16396	78.15773
	Simulation Results				
1.	Average Value Of: a) $\hat{\beta}_1$	0.97868	0.96843	0.95284	0.89491
	b) $\hat{\beta}_2$	1.02061	1.03096	1.04659	1.10454
2.	Average Sum Of Squares Of The $\hat{\beta}_i$	2.48585	2.68854	3.32607	8.06728
3.	Variance For $L^2(\hat{\beta})$	0.23085	0.58169	2.70966	66.68130
4.	Percentages Of Correct Signs:				
	a) Neither $\hat{\beta}_1$ Nor $\hat{\beta}_2$	0.00%	0.00%	0.00%	0.00%
	b) Only $\hat{\beta}_1$	1.80%	3.80%	9.90%	26.20%
	c) Only $\hat{\beta}_2$	2.70%	4.60%	12.30%	30.20%
	d) Both $\hat{\beta}_1$ And $\hat{\beta}_2$	95.50%	91.60%	77.80%	43.60%

Table 2 - A Summary Of N = 1,000 Simulations Of The Least Squares Estimates For The Model:

$$Y = X_i \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \varepsilon, \qquad \varepsilon \sim N(0.0, 0.36 I_3)$$

		Design Matrix			
		X_1	X_2	X_3	X_4
Theoretical Values					
1. Eigenvalues For $X_i'X_i$:	a) λ_1	1.00000	1.55027	1.79989	1.96000
	b) λ_2	1.00000	0.44957	0.19999	0.04000
2. Sum Of Squares Of The $\hat{\beta}_i$		2.00000	2.00000	2.00000	2.00000
3. Theoretical Sum Of Squares Of The $\hat{\beta}_i$		2.72000	3.03298	4.00012	11.18369
4. Theoretical Variance For $L^2(\hat{\beta})$		0.51840	1.39030	6.56079	162.06784
Simulation Results					
1. Average Value Of:	a) $\hat{\beta}_1$	0.97443	0.96213	0.94342	0.87391
	b) $\hat{\beta}_2$	1.02475	1.03717	1.05593	1.12547
2. Average Sum Of Squares Of The $\hat{\beta}_i$		2.69992	2.99175	3.90977	10.73727
3. Variance For $L^2(\hat{\beta})$		0.47868	1.20620	5.61791	138.27827
4. Percentages Of Correct Signs					
a) Neither $\hat{\beta}_1$ Nor $\hat{\beta}_2$		0.30%	0.00%	0.00%	0.00%
b) Only $\hat{\beta}_1$		3.90%	7.00%	14.90%	29.00%
c) Only $\hat{\beta}_2$		4.80%	8.20%	17.40%	34.60%
d) Both $\hat{\beta}_1$ And $\hat{\beta}_2$		91.00%	84.80%	67.70%	36.40%

An indication of the increasing instability of the least squares estimator as the smallest eigenvalue tends to zero or as σ^2 increases is provided by considering the breakdown of correct signs for the estimated parameters. Using X_1, 95.5% of the estimated vectors $\hat{\beta}$ had both signs correct when σ^2 was set to 0.25 and 91.00% when σ^2 equalled 0.36. On the other hand, in the case of the most non-orthogonal design matrix, only 43.60% of the estimated parameters had both signs correct when σ^2 equalled 0.25. The corresponding figure was 36.40% when σ^2 was set to 0.36.

Smith and Goldstein (1975) noted that the occurrence of small eigenvalues and thus ill-conditioning probably is the result of one of two situations. First, the data points may lie close to a hyperplane in the parameter space. The design is not adequate to estimate all the parameters of the regression model. One remedy would be to collect more data and respecify the model. Secondly, the variables in the model might be highly interdependent so that at least one variable might be considered redundant. A possible solution to this problem would be to drop factors in order to destroy the correlation bonds among the explanatory variables. In this case, the user may be left with 'dangling' controllables. Alternatively, if it is not desirable to drop explanatory or control variables, the parameters for the complete model may be estimated. However, the prediction function must be treated as a 'black box' and the function's derivatives should not be used.

In order to reduce the problems of parameter inflation and instability, Hoerl (1962) introduced the method of ridge regression and defined a class of estimators of the form:

$$\hat{\beta}_k^* = (X'X + kI_p)^{-1}X'Y \qquad (1.30)$$

where:

$$k \geq 0 \qquad (1.31)$$

The ridge estimator defined by (1.30) is a biased estimator for which it is hoped that a major reduction in the variances of the individual parameter estimates can be achieved by introducing a small amount of bias into the estimation process. The remainder of this thesis consists of a comprehensive survey of the various properties of the ridge estimator. The ridge estimator is contrasted with other biased estimators which have been proposed to deal with the effects of severe multicollinearity in the explanatory variables. Several generalizations of the estimator defined by (1.30) are considered. In addition to the classical formulation of the estimator, the ridge estimator is considered within a Bayesian framework. Finally, several criticisms of the ridge regression procedures are outlined.

Chapter 2

The Ridge Estimator

In a discussion of an application of regression analysis in the estimation of control equations for some chemical processes, A.E. Hoerl noted that:

> " The exactness of mathematics can in some cases defeat its own utility as a tool for solving industrial problems. Too often the mathematically exact solution is taken as that whereas subsequent data belies that solution and thus discredits this approach. [1] "

He pointed out that the use of pre-designed experiments in the development of control equations for industrial processes is rarely feasible. Rather the data that is gathered is usually 'poorly conditioned' and as a result unstable coefficients are produced. Simulation results were presented in the previous section to demonstrate that misleading results can often be produced when the interrelationships between the independent variables are strong enough. In order to derive more credible control equations, Hoerl (1962), (1964) introduced the ridge estimator:

$$\hat{\beta}_k^* = (X'X + kI_p)^{-1}X'Y \qquad (2.1)$$

where:

$$k \geq 0 \qquad (2.2)$$

and defined ridge analysis in terms of quadratic response functions. A.E. Hoerl suggested that:

> " Given the analytic solution, the ridge analysis determines the unique combinations of solutions which minimize the lack of fit of the data while decreasing the size of the coefficients. In addition, a solution is determined ... which not only 'fits' the data but is simultaneously a stable solution - the

---

(1.) Hoerl, A.E. : Applications Of Ridge Analysis To Regression Problems . Chemical Engineering Progress 58, 1962, p.56 .

real crux of the matter.[1]"

Hoerl and Kennard (1970a) compiled a comprehensive summary of the properties of the ridge estimator. Adopting the criterion:

> Mean Squared Error Admissibility Criterion: A class of estimators E will be called admissible if for every X and Y, there exists an estimator $e \varepsilon E$ such that:

$$MSE(e) < MSE(\hat{\beta})$$

$$= \sum_{i=1}^{p} Var(\hat{\beta}_i) \quad , \qquad (2.3)$$

they demonstrated the existence of a k such that $\hat{\beta}_k^*$ is mean squared error admissible. Newhouse and Oman (1971) detailed the geometrical implications of the ridge solution. A summary of these results is presented in this section.

The ridge estimator $\hat{\beta}_k^*$ reduces to the least squares estimator when k is set to zero. $\hat{\beta}_k^*$ may be rewritten as:

$$\hat{\beta}_k^* = (X'X + kI_p)^{-1}X'Y$$

$$= (I_p + k(X'X)^{-1})^{-1}(X'X)^{-1}X'Y$$

$$= (I_p + k(X'X)^{-1})^{-1}\hat{\beta}$$

$$= Z_k\hat{\beta} \qquad (2.4)$$

where:

$$Z_k = (I_p + k(X'X)^{-1})^{-1} \quad . \qquad (2.5)$$

Therefore, the ridge estimator $\hat{\beta}_k^*$ is a linear transformation

---

(1.) Ibid., p. 58.

of the least squares estimator $\hat{\beta}$ with the transformation dependent only on k through Z_k. If λ_i, $i=1,2,\ldots,p$, are the eigenvalues of the $X'X$ matrix, then Z_k has p eigenvalues equal to:

$$\lambda_i' = \frac{\lambda_i}{\lambda_i + k} \qquad i=1,2,\ldots,p \tag{2.6}$$

which tend to zero as k increases. It is a common result from matrix algebra that if the eigenvalues for a matrix have limiting values equal to zero then the matrix will tend to a matrix of zeros. Therefore:

$$\begin{aligned} \lim_{k\to+\infty} \hat{\beta}_k^* &= \lim_{k\to+\infty} Z_k \hat{\beta} \\ &= 0\, I_p \hat{\beta} \\ &= 0 \quad , \end{aligned} \tag{2.7}$$

so that $\hat{\beta}_k^*$ tends to a vector of zeros as $k\to+\infty$.

Invoking the results from Chapter 1, the expected value of $\hat{\beta}_k^*$ becomes:

$$\begin{aligned} E(\hat{\beta}_k^*) &= E(Z_k\hat{\beta}) \\ &= Z_k E(\hat{\beta}) \\ &= Z_k \beta \\ &= (I_p + k(X'X)^{-1})\beta \quad . \end{aligned} \tag{2.8}$$

Applying the binomial inverse theorem[1]:

$$(I_p + k(X'X)^{-1})^{-1} = I_p - k(X'X + kI_p)^{-1} \tag{2.9}$$

and so the expected value of the ridge regression estimator becomes:

(1.) Press, S.J.: Applied Multivariate Analysis. New York: Holt, Rinehart And Winston, Inc., 1972, p. 23.

$$E(\hat{\beta}_k^*) = \beta - k(X'X + kI_p)^{-1}\beta \quad . \tag{2.10}$$

It follows from (2.10) that the ridge regression estimator has a bias equal to $-k(X'X + kI_p)^{-1}\beta$. Furthermore, the variance-covariance matrix for $\hat{\beta}_k^*$ is:

$$\begin{aligned}
Var(\hat{\beta}_k^*) &= E(\hat{\beta}_k^* - E(\hat{\beta}_k^*))(\hat{\beta}_k^* - E(\hat{\beta}_k^*))' \\
&= E(Z_k\hat{\beta} - E(Z_k\hat{\beta}))(Z_k\hat{\beta} - E(Z_k\hat{\beta}))' \\
&= Z_k E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' Z_k' \\
&= \sigma^2 Z_k (X'X)^{-1} Z_k' \\
&= \sigma^2 (X'X + kI_p)^{-2}(X'X) \quad .
\end{aligned} \tag{2.11}$$

By definition, the total residual sum of squares for any estimator b of β equals:

$$\begin{aligned}
\Phi(b) &= (Y - Xb)'(Y - Xb) \\
&= (Y - X\hat{\beta} + X\hat{\beta} - Xb)'(Y - X\hat{\beta} + X\hat{\beta} - Xb) \\
&= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (X\hat{\beta} - Xb)'(X\hat{\beta} - Xb) \\
&= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (\hat{\beta} - b)'X'X(\hat{\beta} - b) \\
&= \Phi_{min} + (\hat{\beta} - b)'X'X(\hat{\beta} - b)
\end{aligned} \tag{2.12}$$

where Φ_{min} is the global minimum of the residual sum of squares function. It can be seen from (2.12) that the residual sum of squares function is a quadratic function in $(\hat{\beta} - b)$. There are a continuum of values b which satisfy:

$$\Phi(b) = \Phi_{min} + \Phi_0 \tag{2.13}$$

for any fixed $\Phi_0 > 0.0$. These values form contours of constant residual sum of squares which are hyperellipsoids centered at the least squares solution.

Hoerl and Kennard (1970a) noted that, on the average, the distance between $\hat{\beta}$ and the true parameters β will be large if $X'X$ is ill-conditioned. The worse ill-conditioned $X'X$; the more $\hat{\beta}$ can be expected to differ from β. However, the smaller the eigenvalues λ_i, the further one may move away from the least squares estimator without incurring an appreciably large increase in the residual sum of squares. In light of:

$$EL^2(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \tag{2.14}$$

and:

$$Var(L^2(\hat{\beta})) = 2\sigma^4 \sum_{i=1}^{p} \frac{1}{\lambda_i^2} \tag{2.15}$$

Hoerl and Kennard suggested that "it seems reasonable that if one moves away from the minimum sum of squares of the residuals point, the movement should be in a direction which will shorten the length of the regression parameters.[1]". Applying this criterion, an estimator b should be choosen which will minimize the sum of squares of the residuals subject to the constraint that the squared length of b equals c^2. The corresponding Lagrangian equation:

$$F(b) = (Y - Xb)'(Y - Xb) + k(b'b - c^2) \tag{2.16}$$

is minimized when:

$$\begin{aligned} \frac{d}{db} F(b) &= -2X'Y + 2X'Xb + 2kb \\ &= 2(X'X + kI_p)b - 2X'Y \\ &= 0 \end{aligned} \tag{2.17}$$

or:

$$b = (X'X + kI_p)^{-1}X'Y$$

(1.) Hoerl, A.E. and Kennard, R.W.: Ridge Regression: Biased Estimation For Nonorthogonal Problems. Technometrics 12, 1970, p. 58.

$$= \hat{\beta}_k^* \quad . \tag{2.18}$$

Therefore the ridge regression estimator minimizes the sum of squares of the residuals for a fixed squared parameter length.

Meeter (1966) relaxed the constraint on $\hat{\beta}_k^*$ by showing that $\hat{\beta}_k^*$ minimizes the residual sum of squares on and within a sphere of radius $r = \|\hat{\beta}_k^*\|$. Later, Newhouse and Oman (1971) proved the following theorem:

<u>Theorem 2.1</u>: Let $\|\hat{\beta}_k^*\| = r$ where $k>0$. Then $\hat{\beta}_k^*$ is a unique vector which minimizes $\phi(b)$ subject to $\|b\| \leq r$.

If r is taken large enough, the global minimum sum of squares will be achieved and $\hat{\beta}_k^*$ will in fact be $\hat{\beta}$. The proof of Theorem 2.1 requires the following sequence of lemmas:

<u>Lemma 2.1</u>: Suppose that $0 \leq s < t$. Then

$$\|\hat{\beta}_t^*\| < \|\hat{\beta}_s^*\| \quad . \tag{2.19}$$

<u>Proof</u>: Consider the orthogonal matrix P defined in (1.27) which diagonalizes the $X'X$ matrix. Letting $X^* = XP$ and $\alpha = P'\beta$, the general linear model may be rewritten as:

$$\begin{aligned} Y &= X\beta + \varepsilon \\ &= XPP'\beta + \varepsilon \\ &= X^*\alpha + \varepsilon \quad . \end{aligned} \tag{2.20}$$

The ridge estimator for α is by definition:

$$\begin{aligned} \hat{\alpha}_k^* &= (X^{*\prime}X^* + kI_p)^{-1}X^{*\prime}Y \\ &= (\Lambda + kI_p)^{-1}X^{*\prime}Y \quad . \end{aligned} \tag{2.21}$$

If c_i denotes the i'th element of the vector $X^{*\prime}Y$, then the individual components of $\hat{\alpha}_k^*$ may be written as:

$$\hat{\alpha}_k^* = \left(\frac{c_1}{\lambda_1 + k}, \frac{c_2}{\lambda_2 + k}, \ldots, \frac{c_p}{\lambda_p + k} \right). \qquad (2.22)$$

Since orthogonal transformations preserve lengths, inequality (2.19) is equivalent to $\|\hat{\alpha}_t^*\| < \|\hat{\alpha}_s^*\|$. From the definition of $\hat{\alpha}_s^*$:

$$\begin{aligned} \|\hat{\alpha}_s^*\|^2 &= \hat{\alpha}_s^{*\prime}\hat{\alpha}_s^* \\ &= \sum_{i=1}^{p} \hat{\alpha}_{s_i}^{*\,2} \\ &= \sum_{i=1}^{p} \frac{c_i^2}{(\lambda_i + s)^2} \quad . \end{aligned} \qquad (2.23)$$

A similar result holds for $\|\hat{\alpha}_t^*\|^2$. Since $s<t$, it follows that $(\lambda_i + t)^{-1} < (\lambda_i + s)^{-1}$ for all λ_i. Therefore, $\|\hat{\alpha}_t^*\| < \|\hat{\alpha}_s^*\|^2$ from which the required follows.

<u>Lemma 2.2</u>: Denote the set of estimators $\{b : \|b\| \leq r\}$ by B_r. Let $r < \|\hat{\beta}\|$ and suppose β^* minimizes $\Phi(b)$ subject to $\|b\| \leq r$. Then:

$$\|\beta^*\| = r \quad . \qquad (2.24)$$

<u>Proof</u>: Suppose that $\|\beta^*\| < r$ so that $\beta^* \in B_r$. There exists a neighbourhood about β^* such that β^* minimizes $\Phi(b)$ in this neighbourhood. The gradient for $\Phi(b)$ is given by:

$$\nabla\Phi(b) = 2X'Xb - 2X'Y \qquad (2.25)$$

so that β^* must satisfy:

$$(X'X)\beta^* = X'Y \quad . \qquad (2.26)$$

(2.26) defines the ordinary least squares estimator. It follows from (2.26) that $\beta^* = \hat{\beta} \notin B_r$ which is a contradiction. Thus the assumption that $\|\beta^*\| < r$ is impossible and so the Euclidean length of

β^* must equal r.

Lemma 2.3: Suppose β^* minimizes $\Phi(b)$ subject to the constraint that $\|b\| = r$. Then there exists a real number λ such that:

$$(X'X + \lambda)\beta^* = X'Y \quad . \tag{2.27}$$

Proof: Let $h(b) = \|b\|^2$ so that $\Phi(b)$ is to be minimized subject to the constraint that $h(b) = r^2$. Defining the corresponding Lagrangian equation in terms of gradients, β^* satisfies:

$$\nabla\Phi(\beta^*) + \lambda\nabla h(\beta^*) = 0 \tag{2.28}$$

or:

$$2(X'X)\beta^* - 2X'Y + 2\lambda\beta^* = 0 \tag{2.29}$$

from which (2.27) follows.

Lemma 2.4: Suppose that $\lambda < \mu$, and suppose a and b are two distinct vectors such that:

i) $\|a\| = \|b\| = r > 0$

ii) $(X'X + \lambda)a = X'Y$

iii) $(X'X + \mu)b = X'Y$.

Then:

$$\Phi(a) > \Phi(b) \quad . \tag{2.30}$$

Proof: Applying (i) and (ii), the residual sum of squares for $\Phi(a)$ may be expanded as follows:

$$\begin{aligned}\Phi(a) &= (Y - Xa)'(Y - Xa) \\ &= Y'Y - 2a'X'Y + a'X'Xa \\ &= Y'Y - 2a'X'Y + a'X'Y - a'\lambda a \\ &= Y'Y - a'X'Y - \lambda r^2 \quad .\end{aligned} \tag{2.31}$$

In the same manner:

$$\Phi(b) = Y'Y - b'X'Y - \mu r^2 \quad . \tag{2.32}$$

Inequality (2.30) will be proven if it is shown that:

$$a'X'Y - b'X'Y < (\mu - \lambda)r^2 \quad . \tag{2.33}$$

However,

$$a'X'Y - b'X'Y = a'(X'X + \mu)b - b'(X'X + \lambda)a$$

$$= (\mu - \lambda)a'b \quad . \tag{2.34}$$

Schwarz's inequality provides that:

$$|a'b| \leq \|a\| \cdot \|b\| = r^2 \tag{2.35}$$

with inequality holding if and only if a=tb for some real number t. Two cases must be considered separately. First, suppose that strict inequality holds for (2.35). By assumption $\mu>\lambda$ so that:

$$(\mu - \lambda)a'b < (\mu - \lambda)r^2 \quad . \tag{2.36}$$

Secondly, suppose that $\|a\| = \|b\|$. Since a and b are assumed to be distinct, it follows that t equals -1 and:

$$(\mu - \lambda)a'b = -(\mu - \lambda)r^2$$

$$< (\mu - \lambda)r^2 \quad . \tag{2.37}$$

Inequality (2.33) follows by substituting (2.36) and (2.37) into (2.34).

Theorem 2.1 may be proven by invoking the four lemmas provided above. Since $\Phi(b)$ is a continuous function of b and $B_r \subset R^p$ is compact[1], $\Phi(b)$ must have a minimum value in B_r. Let β^* be the point at which the constrained minimum value of $\Phi(b)$ in R^p is obtained. As a result of Lemma 2.2:

$$\|\beta^*\| = r \quad . \tag{2.38}$$

Lemma 2.3 provides for the existence of a real λ such that:

$$(X'X + \lambda)\beta^* = X'Y \quad . \tag{2.39}$$

By assumption, β^* is the ridge estimator with an Euclidean length equal to r. If $k<\lambda$, a contradiction arises since:

$$\|\beta^*\| < \|\hat{\beta}_k^*\| = r \tag{2.40}$$

would be true by Lemma 2.1. As a result of (2.37) and (2.38), $\lambda<k$ can not be true either. In this case:

$$\|\beta^*\| = \|\hat{\beta}_k^*\| \quad . \tag{2.41}$$

(1.) Royden, H.L.: Real Analysis. Toronto: Collier-MacMillan Canada, Ltd., 1968, p. 158.

along with Lemma 2.4 would imply:

$$\Phi(\beta^*) < \Phi(\hat{\beta}_k^*) \tag{2.42}$$

which contradicts the minimality assumption for β^*. Therefore λ equals k so that $\hat{\beta}_k^*$ satisfies equation (2.40). Since $X'X$ is assumed non-singular, $(X'X + kI_p)$ is also non-singular and

$$\beta^* = \hat{\beta}_k^* \tag{2.43}$$

so that Theorem 2.1 is proven. As a result of Theorem 2.1, it is observed that:

$$\hat{\beta}_k^{*\prime}\hat{\beta}_k^* < \hat{\beta}'\hat{\beta} \tag{2.44}$$

for all positive values of k.

In Chapter 1, it was shown that if the unobservable disturbances are assumed normally distributed and satisfy conditions (1.2), β and σ^2 will have a likelihood function given by:

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

$$= L(\beta,\sigma^2|Y,X). \tag{2.45}$$

Substituting (2.12) into (2.45), the likelihood function for any estimator b of β becomes:

$$L(b,\sigma^2|Y,X) = \frac{1}{(2\pi)^{n/2}\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}(Y - X\hat{\beta})'(Y - X\hat{\beta}) - \frac{1}{2\sigma^2}(\hat{\beta} - b)'X'X(\hat{\beta} - b)\right\}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}(\Phi_{min} + \Phi_0)\right\}. \tag{2.46}$$

From (2.46), it can be seen that an increase in the residual sum of squares results in a decrease in the likelihood of the estimator b given a fixed σ^2. It follows that the ridge estimator maximizes

the likelihood function subject to the constraint that $\|\hat{\beta}_k^*\| \leq r$. Therefore, an increase in the value of k corresponds to a decrease in the likelihood function.

It was mentioned in Chapter 1 that the ridge estimator is a biased estimator for which it is hoped that the introduction of bias into the estimator will reduce the variances of the resulting parameter estimates. This compromise between bias and variance may be quantified by the mean squared error for $\hat{\beta}_k^*$:

$$\begin{aligned}
MSE(\hat{\beta}_k^*) &= E(L^2(\hat{\beta}_k^*)) \\
&= E((\hat{\beta}_k^* - \beta)'(\hat{\beta}_k^* - \beta)) \\
&= E((Z_k\hat{\beta} - \beta)'(Z_k\hat{\beta} - \beta)) \\
&= E((Z_k\hat{\beta} - Z_k\beta + Z_k\beta - \beta)'(Z_k\hat{\beta} - Z_k\beta + Z_k\beta - \beta)) \\
&= E((\hat{\beta} - \beta)'Z_k'Z_k(\hat{\beta} - \beta)) \\
&\qquad + \beta'(Z_k - I_p)'(Z_k - I_p)\beta \quad . \qquad (2.47)
\end{aligned}$$

Applying the same sequence of arguements used to simplify the expression for $EL^2(\hat{\beta})$, the first term of (2.47) becomes:

$$\begin{aligned}
&E((\hat{\beta} - \beta)'Z_k'Z_k(\hat{\beta} - \beta)) \\
&= tr(Z_k'Z_k\sigma^2(X'X)^{-1}) \\
&= \sigma^2 tr((I_p + k(X'X)^{-1})^{-2}(X'X)^{-1}) \\
&= \sigma^2 tr((X'X + kI_p)^{-1}(I_p + k(X'X)^{-1})^{-1}) \quad . \qquad (2.48)
\end{aligned}$$

Invoking the binomial inverse theorem:

$$\begin{aligned}
&(I_p - k(X'X)^{-1})^{-1} \\
&= I_p + k(X'X)^{-1}(k(X'X)^{-1} + k^2(X'X)^{-2})^{-1}k(X'X)^{-1}
\end{aligned}$$

$$= I_p - k(X'X + kI_p)^{-1} \quad . \tag{2.49}$$

Substituting (2.49) into (2.48):

$$E((\hat{\beta} - \beta)'Z_k'Z_k(\hat{\beta} - \beta))$$

$$= \sigma^2 tr((X'X + kI_p)^{-1}) - k\sigma^2 tr((X'X + kI_p)^{-2}) \quad . \tag{2.50}$$

Letting P denote the orthogonal transformation which diagonalizes the matrix $X'X$,

$$E((\hat{\beta} - \beta)'Z_k'Z_k(\hat{\beta} - \beta))$$

$$= \sigma^2 tr((P'(X'X + kI_p)P)^{-1}) - k\sigma^2 tr((P'(X'X + kI_p)P)^{-2})$$

$$= \sigma^2 tr((\Lambda + kI_p)^{-1}) - k\sigma^2 tr((\Lambda + kI_p)^{-2})$$

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

$$= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} \quad . \tag{2.51}$$

The constant term in (2.47) may be simplified by applying (2.49):

$$\beta'(Z_k - I_p)'(Z_k - I_p)\beta$$

$$= \beta'(I_p - k(X'X + kI_p)^{-1} - I_p)'(I_p - k(X'X + kI_p)^{-1} - I_p)\beta$$

$$= k^2\beta'(X'X + kI_p)^{-2}\beta$$

$$= k^2\beta'(PP'(X'X + kI_p)PP')^{-2}\beta$$

$$= k^2\beta'(P(\Lambda + kI_p)P')^{-2}\beta$$

$$= k^2\beta'P(\Lambda + kI_p)^{-2}P'\alpha$$

$$= k^2\alpha'(\Lambda + kI_p)^{-2}\alpha$$

$$= k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2} \quad . \tag{2.52}$$

Substituting (2.51) and (2.52) into (2.47), the mean squared error for the ridge regression estimator becomes:

$$MSE(\hat{\beta}_k^*) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2} \quad . \qquad (2.53)$$

From the above, it can be seen that the mean squared error for $\hat{\beta}_k^*$ is the sum of the two functions:

$$\gamma_1(k) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2}$$

$$\gamma_2(k) = k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2} \quad . \qquad (2.54)$$

$\gamma_1(k)$ represents the total sum of the variances of the estimates for each of the p parameters. $\gamma_1(k)$ tends to the sum of the variances for the least squares estimates as k tends to zero. As k gets large and $\hat{\beta}_k^*$ shrinks, $\gamma_1(k)$ decays to zero. $\gamma_2(k)$ equals the sum of the squares of the biases in the ridge regression estimator components. The squared bias equals zero for the least squares solution but tends to $\beta'\beta$ as k grows large. Figure 1 displays the relationship between $\gamma_1(k)$ and $\gamma_2(k)$ for various values of k.

Consider the limit of the first derivative of $\gamma_1(k)$ as k tends to zero:

$$\lim_{k \to 0^+} \frac{d}{dk} \gamma_1(k) = \lim_{k \to 0^+} \sigma^2 \sum_{i=1}^{p} \frac{-2\lambda_i}{(\lambda_i + k)^3}$$

$$= -2\sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i^2} \quad . \qquad (2.55)$$

$\gamma_1(k)$ has a negative derivative near zero which depends upon the conditioning of the design matrix. If the design matrix X is ill-

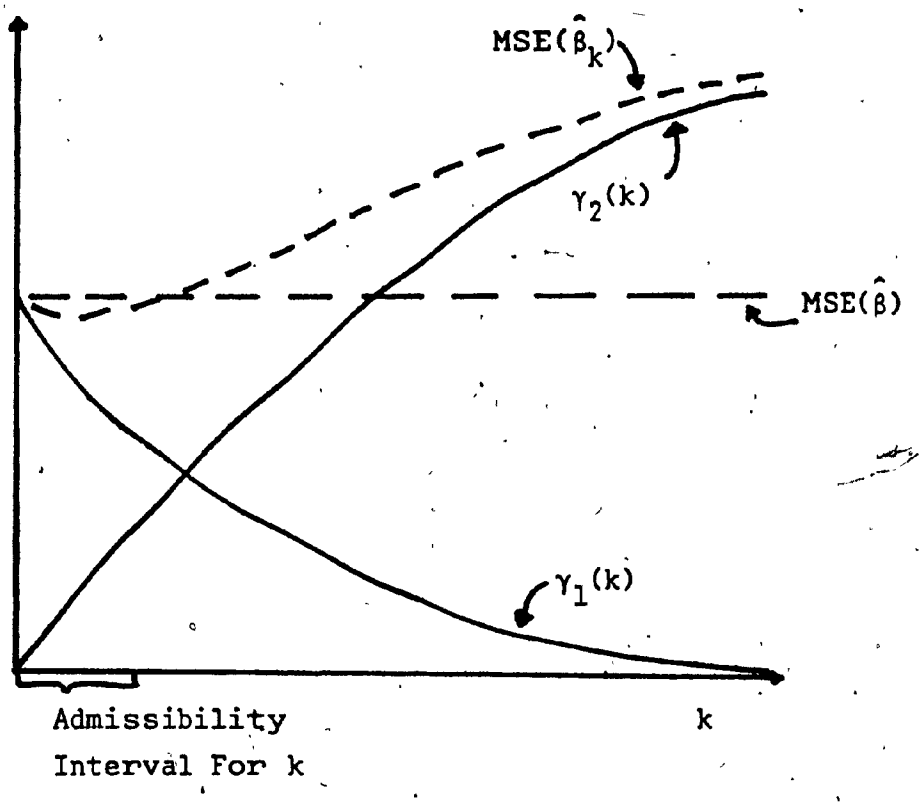


Figure 1 - The Relationship Between The Mean Squared Error Functions For The Ordinary Least Squares Estimator ($\hat{\beta}$) And The Ridge Regression Estimator ($\hat{\beta}_k$)

conditioned so that the $X'X$ matrix has some small eigenvalues, the decrease in the sum of the parameter variances will be large as k moves away from zero. On the other hand,

$$\lim_{k\to 0^+} \frac{d}{dk}\gamma_2(k) = \lim_{k\to 0^+} \sum_{i=1}^{p} \left\{ \frac{2k\alpha_i^2}{(\lambda_i + k)^2} - \frac{2k^2\alpha_i}{(\lambda_i + k)^3} \right\}$$

$$= 0 \qquad (2.56)$$

so that $\gamma_2(k)$ is flat for k near zero. These properties for $\gamma_1(k)$ and $\gamma_2(k)$ would suggest the possibility of a reduction in the overall mean squared error if the ridge estimator is used for $k>0$ instead of the ordinary least squares estimator.

Hoerl and Kennard (1970a) summarized a number of the properties of the mean squared error function for $\hat{\beta}_k^*$. They presented the following theorems and corollaries:

<u>Theorem 2.2</u>: The total variance $\gamma_1(k)$ is a continuous, monotonically decreasing function of k.

<u>Proof</u>: Since the $X'X$ matrix is assumed to be non-singular and k is non-negative, it follows that $(\lambda_i + k) > 0$ for all i so that $\gamma_1(k)$ can not have any singularity points. Furthermore,

$$\lim_{k\to 0^+} \gamma_1(k) = \lim_{k\to 0^+} \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2}$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}$$

$$= \gamma_1(0) \quad . \qquad (2.57)$$

Therefore $\gamma_1(k)$ is a continuous function for $k \geq 0$. For all $k \geq 0$,

$$\frac{d}{dk}\gamma_1(k) = -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^3} < 0 \qquad (2.58)$$

and:

$$\frac{d^2}{dk^2}\gamma_1(k) = 6\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i + k)^4} > 0 \qquad (2.59)$$

Thus $\gamma_1(k)$ is a monotonically increasing function of k.

<u>Corollary 2.2.1</u>: The first derivative of $\gamma_1(k)$ with respect to k approachs $+\infty$ as $k \to 0^+$ and $\lambda_p \to 0^+$.

<u>Proof</u>: By definition of $\gamma_1(k)$:

$$\lim_{\lambda_p \to 0^+}\lim_{k \to 0^+}\frac{d}{dk}\gamma_1(k) = \lim_{\lambda_p \to 0^+}\lim_{k \to 0^+} -2\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i + k)^3}$$

$$= \lim_{\lambda_p \to 0^+} -2\sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i^2}$$

$$= +\infty \qquad (2.60)$$

<u>Theorem 2.3</u>: The squared bias is a continuous monotonically increasing function of k.

<u>Proof</u>: By assumption, $k \geq 0$ and $\lambda_i > 0$ for all i so that $\gamma_2(k)$ has no singularity points. Furthermore,

$$\lim_{k \to 0^+}\gamma_2(k) = \lim_{k \to 0^+} k^2\sum_{i=1}^{p}\frac{\alpha_i^2}{(\lambda_i + k)^2}$$

$$= \gamma_2(0) \qquad (2.61)$$

so that $\gamma_2(k)$ is a continuous function for all $k \geq 0$. Suppose that $g_i(k)$ represents the i'th term of the summation defined by $\gamma_2(k)$. That $g_i(k)$ is a monotonically increasing function of k follows from:

$$\frac{d}{dk}g_i(k) = \frac{2k\lambda_i\alpha_i^2}{(\lambda_i + k)^3} > 0 \quad . \qquad (2.62)$$

Since $\gamma_2(k)$ is composed of a sum of monotonically increasing functions of k, it follows that $\gamma_2(k)$ is a monotonically increasing function.

<u>Corollary 2.3.1</u>: The squared bias $\gamma_2(k)$ approaches $\beta'\beta$ as an upper limit.

<u>Proof</u>: By definition of $\gamma_2(k)$:

$$\lim_{k\to+\infty} \gamma_2(k) = \lim_{k\to+\infty} k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2}$$

$$= \sum_{i=1}^{p} \alpha_i^2$$

$$= \alpha'\alpha$$

$$= (P'\beta)'(P'\beta)$$

$$= \beta'PP'\beta$$

$$= \beta'\beta \quad . \tag{2.63}$$

Hoerl and Kennard (1970a) originally demonstrated the existence of an interval for k such that the ridge regression esimator is mean squared error admissible by means of the theorem:

<u>Theorem 2.4</u>: For all $0<k<\dfrac{\sigma^2}{\alpha_{max}^2}$,

$$MSE(\hat{\beta}_k^*) < MSE(\hat{\beta})$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \quad . \tag{2.64}$$

<u>Proof</u>: For any $k>0$,

$$\sum_{i=1}^{p} \frac{1}{\lambda_i + k} < \sum_{i=1}^{p} \frac{1}{\lambda_i} \tag{2.65}$$

so that it suffices to show the existence of an interval for k such that:

$$MSE(\hat{\beta}_k^*) < \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i + k} \quad . \tag{2.66}$$

From (2.53),

$$MSE(\hat{\beta}_k^*) = \gamma_1(k) + \gamma_2(k)$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2} \quad . \qquad (2.67)$$

Inequality (2.66) holds for a positive value of k if for each i:

$$\sigma^2 \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \frac{\alpha_i^2}{(\lambda_i + k)^2} < \sigma^2 \frac{1}{\lambda_i + k} \qquad (2.68)$$

or:

$$k < \frac{\sigma^2}{\alpha_i^2} \quad . \qquad (2.69)$$

Therefore $\hat{\beta}_k^*$ is mean squared error admissible if inequality (2.69) is satisfied for all i so that:

$$k < \frac{\sigma^2}{\alpha_{max}^2} \quad . \qquad (2.70)$$

The use of Euclidean distance to measure the deviation of the ridge regression estimates from the true parameter vector β was criticized by Needler (1972). He noted that the mean squared errors for the individual components of $\hat{\beta}_k^*$ are lumped together in an unweighted sum. Swindel and Chapman (1974) showed that $\hat{\beta}_k^*$ provides a strictly smaller mean squared error estimator of any non-null linear combination of the components of β in comparision to $\hat{\beta}$ if k lies in an open interval which depends upon X, β and σ^2.

Theobald (1974) generalized the conditions for Theorem 2.4 by showing the existence of a $k>0$ such that $\hat{\beta}_k^*$ is admissible using any weighted mean squared error function. Suppose that b_i is any estimator of β. The second-order moment and weighted mean squared error matrices for b_i are:

$$M(b_i) = E(b_i - \beta)(b_i - \beta)'$$
$$m(b_i) = E(b_i - \beta)'B(b_i - \beta) \qquad (2.71)$$

where B is any non-negative definite matrix. For any estimators b_i

and b_2 of β, Theobald (1974) proved that the two conditions:

i) $M(b_1) - M(b_2)$ is non-negative definite

ii) $m(b_1) - m(b_2) = 0$ for all non-negative definite matrices B

are equivalent.

The existence of a $k>0$ such that $\hat{\beta}_k^*$ is mean squared error admissible follows immediately from Theobald's (1974) theorem:

<u>Theorem 2.5</u>: There exists a $K>0$ such that $M(\hat{\beta}) - M(\hat{\beta}_k^*)$ is positive definite whenever $0<k<K$.

<u>Proof</u>: The second-order moment matrix is by definition:

$$\begin{aligned} M(b_i) &= E(b_i - \beta)(b_i - \beta)' \\ &= E(b_i - E(b_i))(b_i - E(b_i))' + (E(b_i) - \beta)(E(b_i) - \beta)' \\ &= Var(b_i) + (E(b_i) - \beta)(E(b_i) - \beta)' \quad . \qquad (2.72) \end{aligned}$$

Since the ordinary least squares estimator is unbiased, the second-order moment matrix for $\hat{\beta}$ is:

$$\begin{aligned} M(\hat{\beta}) &= Var(\hat{\beta}) \\ &= \sigma^2(X'X)^{-1} \quad . \qquad (2.73) \end{aligned}$$

From (2.10) and (2.11),

$$\begin{aligned} M(\hat{\beta}_k^*) &= Var(\hat{\beta}_k^*) + (E(\hat{\beta}_k^*) - \beta)(E(\hat{\beta}_k^*) - \beta)' \\ &= \sigma^2(X'X + kI_p)^{-2}(X'X) \\ &\quad + k^2(X'X + kI_p)^{-1}\beta\beta'(X'X + kI_p)^{-1} \quad . \qquad (2.74) \end{aligned}$$

Therefore, the difference be the second-order moment matrices of $\hat{\beta}$ and $\hat{\beta}_k^*$ equals:

$$M(\hat{\beta}) - M(\hat{\beta}_k^*)$$

$$= \sigma^2(X'X)^{-1} - \sigma^2(X'X + I_p)^{-2}(X'X)$$
$$- k^2(X'X + kI_p)^{-1}\beta\beta'(X'X + kI_p)^{-1}$$
$$= \sigma^2(X'X + kI_p)^{-1}(I_p + k(X'X)^{-1} - (X'X + kI_p)^{-1}X'X)$$
$$- k^2(X'X + kI_p)^{-1}\beta\beta'(X'X + kI_p)^{-1}$$

$$= \sigma^2(X'X + kI_p)^{-1}(k(X'X)^{-1} + k(X'X)^{-1}(I_p + k(X'X)^{-1})^{-1})$$
$$- k^2(X'X + kI_p)^{-1}\beta\beta'(X'X + kI_p)^{-1}$$

$$= k(X'X + kI_p)^{-1}(\sigma^2(2I_p + k(X'X)^{-1}) - k\beta\beta')(X'X + kI_p)^{-1}. \quad (2.75)$$

(2.75) is a positive definite matrix if:

$$2\sigma^2 I_p + \sigma^2 k(X'X)^{-1} - k\beta\beta' \quad (2.76)$$

is positive definite. Since the $X'X$ matrix is assumed to be of full rank, $M_1 - M_2$ is positive definite if:

$$2\sigma^2 I_p - k\beta\beta' \quad (2.77)$$

is non-negative definite. Expression (2.77) has p-1 roots equal to $2\sigma^2$ and one root equal to $(2\sigma^2 - k\beta'\beta)$ Therefore, a sufficient condition for $M(\hat{\beta}) - M(\hat{\beta}_k^*)$ to be positive definite is for k to satisfy:

$$0 < k < \frac{2\sigma^2}{\beta'\beta} \quad . \quad (2.78)$$

As a result of Theobald's (1974) equivalence statement, any comparisons on the basis of a weighted mean squared error will favour $\hat{\beta}_k^*$ over $\hat{\beta}$ if condition (2.78) is satisfied. It should be noted that Hoerl and Kennard's (1970a) admissibility interval for k is approximately p/2 times longer than Theobald's (1974) interval when B is a p x p identity matrix.

In the assessment of the accuracy of the ridge regression estimator, the mean squared error for $\hat{\beta}_k^*$ has been compared with:

$$\sigma^2 tr((X'X)^{-1}) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \tag{2.79}$$

which is the mean squared error for $\hat{\beta}$. Banerjee and Carr (1971) argued that the relative accuracy of $\hat{\beta}_k^*$ would be more meaningful if the decrease in the mean squared error is compared with:

$$\sigma^2 tr((X'X + kI_p)^{-1}) = \sigma^2 \sum_{i=1}^{p} \frac{1}{(\lambda_i + k)} . \tag{2.80}$$

In order to compare the mean squared error of $\hat{\beta}_k^*$ with (2.80), Banerjee and Carr (1971) introduced the augmented model:

$$\begin{bmatrix} Y_X \\ \hline Y_A \end{bmatrix} = \begin{bmatrix} X \\ \hline \sqrt{k} I_p \end{bmatrix} \beta + \varepsilon \tag{2.81}$$

which was originally proposed by Marquardt (1970). The original design matrix is augmented with a p x p diagonal matrix whose diagonal elements are all equal to $\sqrt{k}$. The vector of dependent variables is augmented to become a (n + p) x 1 vector. The expected values of the components of the model defined by (2.81) are:

$$E(Y_X) = X\beta \tag{2.82}$$

and:

$$\begin{aligned} E(Y_A) &= \sqrt{k} I_p \beta \\ &= \sqrt{k}\beta \end{aligned} \tag{2.83}$$

respectively. The ordinary least squares estimator corresponding to the augmented model is:

$$\hat{\beta}_A = (X'X + kI_p)^{-1}(X'Y_X + \sqrt{k}Y_A) . \tag{2.84}$$

Banerjee and Carr (1971) argued that $\hat{\beta}_k^*$ should be compared with $\hat{\beta}_A$ instead of $\hat{\beta}$. $\hat{\beta}_k^*$ is obtained from $\hat{\beta}_A$ by omitting from the estimation proceedure some observations Y_A which if observable and available would have been included. Since $\hat{\beta}_A$ is an ordinary least squares estimator:

$$MSE(\hat{\beta}_A) = E(L^2(\hat{\beta}_A))$$

$$= \sigma^2 tr((X'X + kI_p)^{-1})$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i + k} \quad . \qquad (2.85)$$

Thus $\hat{\beta}_k^*$ is mean squared error admissible if and only if there exists a $k>0$ such that:

$$MSE(\hat{\beta}_k^*) < MSE(\hat{\beta}_A)$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i + k} \quad . \qquad (2.86)$$

However, this last condition is provided by (2.68) in the proof of Theorem 2.4. Thus the ridge estimator is mean squared error admissible when compared with it's corresponding unbiased estimator.

Up to this point, no mention has been made of how a choice of $k>0$ should be made given any set of data. Hoerl (1962) suggested plotting the residual sum of squares as a function of the sum of squares of the coefficients. He stated that the ridge estimates should be choosen from an interval of k where the individual estimates are stable and the sum of squares of the residuals is increasing rapidly. Functionally, these points correspond to the interval on the curve where the derivative of the sum of squares of the residuals with respect to the sum of squares of the coefficients has a maximum. In order to determine this point, Hoerl (1962) suggested computing various values of:

$$\frac{d}{dR} \Phi(k) = \frac{d}{dR} (Y'Y - \hat{\beta}_k^{*\prime} X'Y + k\hat{\beta}_k^{*\prime}\hat{\beta}_k^*) \qquad (2.87)$$

where:

$$R = \hat{\beta}_k^{*\prime}\hat{\beta}_k^* \quad . \qquad (2.88)$$

and choosing a value of $\hat{\beta}_k^*$ which maximizes the derivative.

Later, Hoerl and Kennard (1970a) clarified the problem of

choosing a value of k through the introduction the the ridge trace. The ridge trace is a graph of the ridge regression parameters and the resulting residual sums of squares as functions of k. Hoerl and Kennard (1970a) provided the following suggestions for choosing a value of k using the ridge trace:

i) The coefficients will have stabilized with respect to their signs and absolute vadues.

ii) The system will resemble an orthogonal system.

iii) The residual sum of squares will not be greatly inflated with respect to the least squares solution.

Hoerl and Kennard (1970a) suggested that systems which show unreasonably large coefficient values or incorrect signs for k=0 will probably correct themselves as k increases.

A number of other algorithms for choosing a value of k have been proposed in the ridge regression literature and tested by means of Monte Carlo simulations. According to Hoerl and Kennard's (1970a) mean squared error admissibility criterion, k should be choosen so that:

$$0 < k < \frac{\sigma^2}{\alpha_i^2} = r_i \qquad (2.89)$$

is satisfied for all i. It is clear that an estimate of k should be made by combining estimates of each r_i. Hoerl, Kennard and Baldwin (1975) argued that an ordinary average of the r_i's would not be appropriate. Such an average would give too much weight to the smallest α_i's which have little predictive power. As a result, they argued

that too much bias would be introduced into the estimate of k. Instead, Hoerl, Kennard and Baldwin (1975) suggested that k be formed from the harmonic mean of the r_i's. In this case, a combined value of k would be:

$$\frac{1}{k_h} = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{r_i}$$

$$= \frac{\alpha'\alpha}{p\sigma^2}$$

$$= \frac{\beta'\beta}{p\sigma^2} \quad . \qquad (2.90)$$

They proposed estimating k_h by substituting ordinary least squares estimates of β and σ^2 into (2.90) so that:

$$\hat{k}_h = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \quad . \qquad (2.91)$$

Hoerl, Kennard and Baldwin (1975) carried out a large number of simulated ridge regressions using $\hat{k}_h$. Provided that the multicollinearity was severe enough, they found substantial improvements in the mean squared errors for the ridge regression estimates using $\hat{k}_h$ compared with the corresponding least squares estimates.

In a later paper, Hoerl and Kennard (1976) noted that $\hat{\beta}'\hat{\beta}$ tends to over-estimate $\beta'\beta$ so that the resulting estimates of k_h using (2.91) will often be too small. They suggested that an iterative procedure be adopted to obtain better estimates of k_h. Initial estimates of k_h could be obtained using (2.91) and the resulting ridge regression estimates of β calculated. Since $\hat{\beta}_k^*$ is assumed to be closer

to β than $\hat{\beta}$, a new estimate of k_h could be formed by replacing $\hat{\beta}$ with $\hat{\beta}_h^*$ in (2.91). Hoerl and Kennard (1976) suggested that the procedure be continued until the convergence of the $\hat{k}_h$'s is obtained. Hoerl and Kennard (1976) incorporated this iterative procedure in simulations similar to those described in their earlier paper. Based upon the simulation results, they concluded that the iterative procedure for estimating k_h leads to mean squared errors with smaller means and variances than the mean squared errors which result from the ordinary least squares estimator and the ridge estimator using a single value of k_h.

McDonald and Galarneau (1975) introduced a mechanical rule for choosing k which employes an unbiased estimator of $\beta'\beta$. Equation (1.26) gives the expected squared length of the ordinary least squares estimator of β. It can be seen from this equation that:

$$\begin{aligned} d &= \hat{\alpha}'\hat{\alpha} - \hat{\sigma}^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \\ &= \hat{\beta}'\hat{\beta} - \hat{\sigma}^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \end{aligned} \tag{2.92}$$

is an unbiased estimator of the squared length of β. McDonald and Galarneau (1975) suggested that a value of k be choosen to satisfy:

$$\hat{\alpha}_k^{*\prime}\hat{\alpha}_k^* = d \tag{2.93}$$

In situations where (2.92) is negative, they recommended that the parameters be estimated using the ordinary least squares estimator.

Besides the two studies mentioned above, Newhouse and Oman (1971), McDonald and Galarneau (1975), Lawless and Wang (1976) and Wichern and Churchill (1978) tested the potential usefulness of ridge estimators using Monte Carlo simulations. In each case, they reported that the improvement in mean squared error using $\hat{\beta}_k^*$ tends to be greater as the number of explanatory variables, spread in the eigenvalues of the $X'X$ matrix or magnitude of σ^2 increases. Wichern and Churchill (1978) carried out an extensive simulation study of the ridge estimator using various rules for choosing k including the ridge trace. They found that the values of k produced by mechanical rules were not always consistent with those produced using the ridge trace. The values of k obtained from the ridge trace tended to be larger. Wichern and Churchill (1978) recommended that the ridge trace only be used in conjunction with the mechanical procedures.

Hocking (1972) reviewed some criteria for choosing subset regressions. He suggested that ridge regression might provide a useful tool for choosing an appropriate equation. Furnival and Wilson (1972) described an efficient technique for computing large numbers of subset regressions. They conjectured that evaluating all subset regressions using their technique might be preferable to ridge regression when the cost of measuring the explanatory variables is small. Hocking (1976) provided an extensive review of various procedures for selecting subset regressions. In addition, he considered a number of biased estimators including the ridge estimator. Hocking (1976) concluded that the ridge estimator compares favourably with any of the procedures he considered for

selecting subset regressions.

Hawkins (1975) presented an efficient technique for evaluating the ridge estimates corresponding to various values of k. In an earlier paper, Hawkins (1973) demonstrated that the ordinary least squares estimator may be constructed using weighted sums of the eigenvalues for an augmented correlation matrix of the dependent and independent variables. The technique for evaluating the ridge estimates proposed by Hawkins (1975) is a direct application of this result.

Consider the augmented matrix:

$$W = (Y \mid X) \tag{2.94}$$

of independent and dependent variables. If it is assumed that each of the variables in (2.94) has been scaled to have a mean equal to zero and a unit standard deviation, the matrix:

$$\begin{aligned} S &= W'W \\ &= \begin{bmatrix} Y'Y & Y'X \\ X'Y & X'X \end{bmatrix} \end{aligned} \tag{2.95}$$

becomes a correlation matrix. Let D denote the $(p + 1) \times (p + 1)$ matrix which satisfies the condition:

$$D'SD = I_{(p + 1)} \tag{2.96}$$

The (i,j)'th element of D is given by:

$$d_{ij} = \lambda_i^{-\frac{1}{2}} a_{ij} \tag{2.97}$$

where a_{ij} is the j'th element of the i'th eigenvector and λ_i the i'th eigenvalue for the matrix S. Hawkins (1973) noted that the transformation:

$$Z = D\,W \tag{2.98}$$

produces (p + 1) mutually uncorrelated random variables.

Suppose that W_i and Z_i represent the observations of the variables which make up the columns of the matrices W and Z. Hawkins (1973) derived the ordinary least squares estimators by considering equations of the form:

$$\sum_{i=1}^{p+1} \gamma_i Z_i = 0 \tag{2.99}$$

which satisfy:

$$\sum_{i=1}^{p+1} \gamma_i d_{i1} = 1 \tag{2.100}$$

Expressing (2.99) in terms of the columns of the original augmented matrix gives:

$$\sum_{i=1}^{p+1} \gamma_i \sum_{j=1}^{p+1} d_{ij} W_j = \sum_{j=1}^{p+1} \left\{ \sum_{i=1}^{p+1} \gamma_i d_{ij} \right\} W_j$$

$$= 0 \tag{2.101}$$

Hawkins (1973) noted that a multiple regression equation may be formed by substituting (2.100) into (2.101) so that:

$$Y = W_1$$

$$= - \sum_{j=2}^{p+1} \left\{ \sum_{i=1}^{p+1} \gamma_i d_{ij} \right\} W_j \tag{2.102}$$

Since the columns of W have been standardized, the residual variance of (2.102) as an estimator of the independent variable Y is given by:

$$\sum_{i=2}^{p+1} \gamma_i^2 \, \mathrm{Var}\left(\sum_{j=1}^{p+1} d_{ij} W_j \right) = \sum_{i=2}^{p+1} \gamma_i^2 \sum_{j=1}^{p+1} \frac{a_{ij}^2}{\lambda_i} \mathrm{Var}(W_j)$$

$$= \sum_{i=2}^{p+1} \gamma_i^2 \tag{2.103}$$

Hawkins (1973) observed that the ordinary least squares estimator of γ corresponds to that estimator which minimizes (2.103) subject to the constraints defined by (2.100). Therefore, the ordinary least squares estimators of γ_i and β_{i-1} are given by:

$$\hat{\gamma}_i = - d_{i1} / \left(\sum_{j=1}^{p+1} d_{j1}^2 \right)$$

$$= - a_{i1} \lambda_i^{-\frac{1}{2}} \left\{ \sum_{j=1}^{p+1} \frac{a_{j1}^2}{\lambda_j} \right\}^{-\frac{1}{2}} \tag{2.104}$$

and:

$$\hat{\beta}_{i-1} = \sum_{j=1}^{p+1} \hat{\gamma}_i \, d_{ij}$$

$$= - \left\{ \sum_{j=1}^{p+1} \lambda_i^{-1} a_{i1} \, a_{ij} \right\} \left\{ \sum_{j=1}^{p+1} \lambda_i^{-1} a_{j1}^2 \right\}^{-1} \tag{2.105}$$

respectively.

Equation (2.105) expresses the ordinary least squares estimator of β_i in terms of weighted sums of the eigenvalues for the augmented correlation matrix S. Hawkins (1975) provided a similar formulation of the ridge estimator. To this end, he utilized the augmented model:

$$\begin{bmatrix} Y \\ \hline A \end{bmatrix} = \begin{bmatrix} X \\ \hline B \end{bmatrix} \beta + \varepsilon \tag{2.106}$$

where:

$$A'A = k$$
$$B'B = kI_p \qquad (2.107)$$
$$A'B = B'A = 0$$

which was originally proposed by Allen (1974). As a result of assumptions (2.107), the ordinary least squares estimator for the augmented model may be expressed as:

$$\hat{\beta}_A = (X'X + B'B)^{-1}(X'Y + B'A)$$
$$= (X'X + kI_p)^{-1}X'Y. \qquad (2.108)$$

It can be seen from (2.108) that the ordinary least squares estimator for the augmented model corresponds to the ridge regression estimator for the standard model.

Since the ridge estimator for the standard model corresponds to the ordinary least squares estimator for the augmented model defined by (2.106), Hawkins (1975) suggested utilizing (2.105) to evaluate the ridge estimates. A correlation matrix similar to S may be constructed for the augmented model according to:

$$S(k) = \begin{bmatrix} Y'Y + A'A & Y'X + A'B \\ X'Y + B'A & X'X + B'B \end{bmatrix}$$
$$= \begin{bmatrix} Y'Y + k & Y'X \\ X'Y & X'X + kI_p \end{bmatrix}$$
$$= S + kI_{(p+1)} \qquad (2.109)$$

The eigenvalues for S(k) can be obtained by diagonalizing the matrix S and adding k. In this case, the ridge estimator of β_{i-1} becomes:

$$\hat{\beta}^*_{i-1}(k) = \left\{ \sum_{j=1}^{p+1} \frac{a_{i1}\, a_{ij}}{(\lambda_i + k)} \right\} \left\{ \sum_{j=1}^{p+1} \frac{a_{j1}^{\,2}}{(\lambda_j + k)} \right\}^{-1} \qquad (2.110)$$

Hawkins (1976) argued that equation (2.110) provides an efficient technique for estimating the ridge regression estimates corresponding to different values of k. He pointed out that only simple averaging operations are required to determine the estimates corresponding to different values of k once the a_{ij}'s are initially computed.

A number of numerical examples have been provided in the ridge regression literature to demonstrate the utility of ridge regression. Hoerl and Kennard (1970b) employed ridge regression analysis in two multiple factor problems. One was a model describing the comprehesive strength of pitprops as a function of thirteen factors which can be measured on the props. In the original solutions for this problem proposed by Jeffers (1967), it was shown that there are significant decreases in the amount of variation in the comprehensive strength explained by the independent variables when some of the independent variables are dropped to break the correlation bonds. Hoerl and Kennard (1970b) showed that while the ridge regression solutions reduce the amount of variation in the comprehesive strength explained; the reduction is not as large as would be incurred if some of the factors are dropped or principal components employed. McDonald and Schwing (1973) employed the ridge estimator in analysing mortality rates by regressing on various socio-economic, weather and pollution variables. Ridge regressions were utilized by Goode (1975) to analyse professional football data.

Obenchain and Vinod (1974) studied the estimation of partial derivatives using ridge regression. In a later paper, Vinod (1976c) utilized these results to extend the ridge regression technique to the problem of canonical correlation analysis. Vinod (1974) considered the estimation of a trans-log production function when multicollinearity and autocorrelated disturbances are present. Brown and Beattie (1975) discussed some of the advantages and limitations of the ridge estimator in the context of economic models. They employed the ridge estimator to estimate the marginal value productivity of irregation water.

Bolding and Houston (1974) described a fortran program for ridge regression estimation. More elaborate programs were developed by the National Bureau Of Economic Research (1975).

Farebrother (1975) examined the relationship between ridge regression and minimum mean squared error estimators. He proposed two consistent estimators of β which are variations of the minimum mean squared error estimators. Farebrother (1975) suggested that the consistent estimators are preferable to ridge regression. Dwivedi and Srivastava (1978) analysed the properties of Farebrother's (1975) estimators. They provided iterative procedures for estimating β based upon the consistent estimators.

Brown (1977) noted that the multiple linear regression model is often of the form:

$$Y = \beta_0 \underline{1} + X\beta_1 + \varepsilon \tag{2.111}$$

where β_0 is a constant and $\underline{1}$ a unit vector. If β_0 and β_1 are not separated, the general linear model becomes:

$$Y = Z\beta + \varepsilon \tag{2.112}$$

where $Z=(\underline{1},X)$ and $\beta'=(\beta_0,\beta_1')$. The usual assumption which is made in the ridge regression literature is that the observations for each independent variable are standardized by subtracting the sample mean from the observations and dividing by the standard deviation. The resultant $X'X$ matrix is a correlation matrix. Brown (1977) observed that the constant β_0 in (2.112) serves only to center the function with respect to the average value of Y when all the independent variables are set to zero. As a result, Brown (1977) considered the simpler standardization of the explanatory variables which is obtained by subtracting the sample mean from the observations for each variable. He pointed out that the resultant ridge estimator is of the form:

$$\begin{aligned}\hat{\beta}_k^* &= (Z'Z + k\,\mathrm{Diag}(0,1,1,\ldots,1))^{-1}Z'Y \\ &= \left\{\frac{1}{n}\sum_{i=1}^{n} y_i\ ,\ (X\,X + kI_p)^{-1}X'Y\right\} .\end{aligned} \tag{2.113}$$

Brown (1977) utilized this form of the ridge estimator to demonstrate the location invariance of the estimator.

Farebrother (1978) considered the standardized ridge estimator defined by (2.113). He establised that this estimator is a special case of the more general ridge estimator:

$$\tilde{\beta}^* = (X'X + kA)^{-1}X'Y \tag{2.114}$$

where A is a positive semi-definite matrix. Farebrother (1978)

derived the mean squared error admissibility conditions for $\tilde{\beta}_k^*$ in terms of the original explanatory variables.

In order to illustrate one of the situations in which the use of ridge regression greatly enhances the credibility of the estimated parameters, consider the following example extracted from the French national accounts for the years 1949 to 1959[1]. The total imports (Y), gross domestic production (X_1), stock-formation (X_2) and consumption (X_3) for these years are summarized in Table 3. Suppose that a model relating total imports to the other three variables is required. It may be required that the total imports corresponding to different levels of each of the three explanatory variables be estimated. The model will be of the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \qquad (2.115)$$

The sample correlation matrix calculated from the data in is:

	Y	X_1	X_2	X_3
Y	1.0000			
X_1	0.9721	1.0000		
X_2	0.3311	0.1687	1.0000	
X_3	0.9753	0.9973	0.1545	1.0000

(2.116)

(1.) Malinvaud, E.: Statistical Methods Of Econometrics. Chicago: Rand, McNally & Company, 1966, p. 17.

Table 3 - Imports, Production, Stock Formation And Consumption In France

(In Millards Of New Francs At 1956 Prices)

Year	Imports	Gross Domestic Production	Stock Formation	Consumption
1949	12.6	117.0	3.1	84.5
1950	13.1	126.3	3.6	89.7
1951	15.1	134.4	2.3	96.2
1952	15.1	137.5	2.3	99.1
1953	14.9	141.7	0.9	103.2
1954	16.1	149.4	2.1	107.5
1955	17.9	158.4	1.5	114.1
1956	21.0	166.5	3.8	120.4
1957	22.3	177.1	3.6	126.8
1958	21.9	179.8	4.1	127.2
1959	21.0	183.8	1.9	128.7
Average -	17.36	151.99	2.65	108.85
Standard Deviation -	3.61	22.77	1.05	15.69

It can be seen that both gross domestic production and consumption are highly correlated with total imports. At the same time, there is almost 'perfect' correlation between gross domestic production and consumption.

The parameters in model (2.115) are estimated by standardizing the data. Each observation is replaced by the difference between the observation and the sample mean for the variable divided by the sample estimate of the variable's standard deviation. As a result of this transformation, the $X'X$ matrix corresponds to the sample correlation matrix for the independent variables while $X'Y$ is a vector of correlations between the dependent variable and each of the independent variables. It should be noted that there is no constant term in the standardized model. The ordinary least squares solution for the standardized model is:

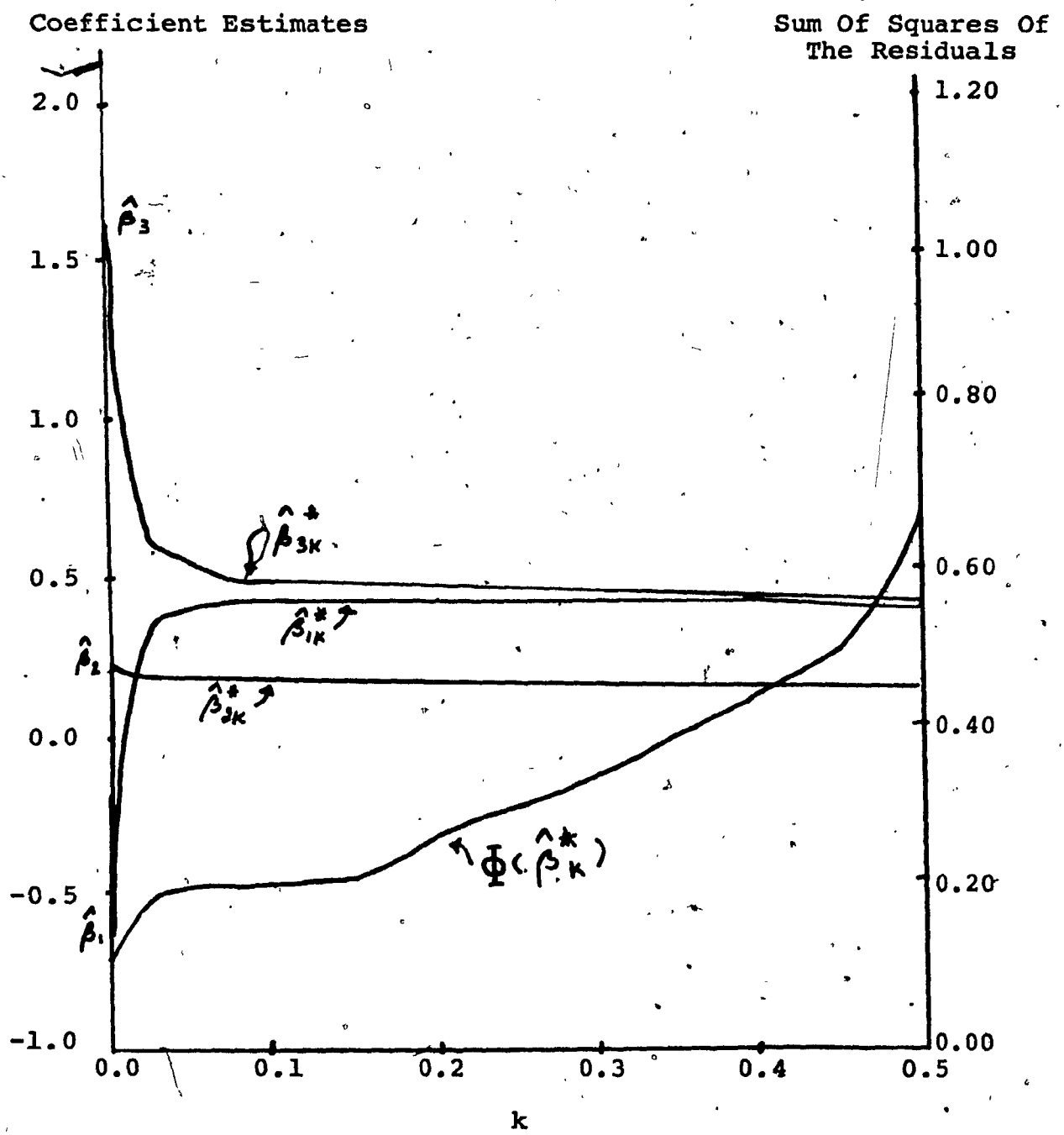
$$\hat{\beta} = \begin{pmatrix} -0.6453 \\ 0.1945 \\ 1.5889 \end{pmatrix} . \qquad (2.117)$$

This solution is clearly nonsensical. Even though gross domestic production and consumption are highly correlated, the estimates for the corresponding parameters are drastically different. The sign for the parameter corresponding to gross domestic production is wrong.

The ridge regression estimates for various values of k between 0.0 and 0.5 are plotted in Figure 2. From this graph, it can be seen that the sign of the parameter corresponding to gross domestic production corrects itself in the interval (0.0, 0.2) of k. For any value of k in the interval (0.09,0.16), the three coefficients are reasonably stable yet the sum of squares of the residuals has not increased significantly. Any value of k in this interval would be a reasonable choice. Taking k to be equal to 0.10, the resulting ridge regression estimates are:

$$\hat{\beta}^*_{0.10} = \begin{pmatrix} 0.4239 \\ 0.1688 \\ 0.4700 \end{pmatrix} . \qquad (2.118)$$

Figure 2 - The Ridge Regression Estimates And Residual Sum Of Squares Function For The Model Of The Total French Imports



At this point, the coefficients corresponding to gross domestic production and consumption are of similar magnitudes reflecting their approximately equal correlations with total imports. Based upon these estimates for the standardized model, the estimated non-standardized model becomes:

$$Y = -6.40 + 0.0672\ X_1 + 0.5832\ X_2 + 0.1102\ X_3 . \quad (2.119)$$

For comparison purposes, a summary of the ordinary least squares estimates for each of the possible subset regressions is provided in Table 4. The parameter estimates and resulting residual sum of squares shown in the table are based upon the correlation form of the model. From Table 4 it can be seen that it is necessary to drop either the variable representing gross domestic production or consumption from the import model in order that the remaining least squares estimates will have the correct signs. The remaining coefficient estimate for either gross domestic production or consumption is approximately equal to the sum of the coefficient estimates corresponding to both variables when k is taken to be 0.10 in the ridge regression solution. The residual sum of squares for this ridge regression solution is 0.2146. This figure compares favourably with the residual sum of squares for the subset regressions when one variable is dropped. Therefore, it can be seen from the above that ridge regressions can be employed to calculate parameter estimates which are both 'sensible' and stable for the French imports model without having to drop any explanatory variables.

Table 4 - A Summary Of The Standardized Ordinary Least Squares Estimates For The French Import Model Corresponding To Each Subset Regression

Explanatory Variables	Estimated Standardized Coefficients $\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	Sum Of Squares Of The Residuals
1	0.9721	-	-	0.5499
2	-	0.3311	-	8.9037
3	-	-	0.9753	0.4870
1,2	0.9431	0.1720	-	0.2624
1,3	-0.1143	-	1.0894	0.4863
2,3	-	0.1849	0.9468	0.1534
1,2,3	-0.6453	0.1945	1.5889	0.1320

Chapter 3

The Relationships Between The Ridge Estimator And Other Biased Estimators

The instability of the ordinary least squares estimator when there is a large degree of multicollinearity between the explanatory variables has led to the development of a number of biased estimators. Two different approachs to the construction of these biased estimators are exemplified by the generalized least squares estimators and shrinkage estimators. In the case of the generalized least squares estimator, it is no longer assumed that the $X'X$ matrix is of rank p. Rather, an estimator is obtained by replacing the inverse of the $X'X$ matrix in (1.6) by a generalized inverse of a rank less than p. In contrast, the shrinkage estimators retain the assumption that the $X'X$ matrix of full rank but cope with the effects of multicollinearity between the explanatory variables by applying a linear transformation to shrink the least squares estimates. The form of the linear transformation is dependent upon the norm with respect to which the accuracy of the estimator is to be measured and possibly an additional constraint upon the estimated length of the parameters. In the remainder of this section, the properties of these two types of estimators are surveyed and compared with the ridge estimator.

Before considering in detail the properties of the generalized least squares estimator, the concept of a generalized inverse is introduced. A generalized inverse for an arbitrary p x q matrix A is defined as any matrix A^+ which satisfies:

$$A A^{+} A = A \qquad (3.1)$$

Matrices A^{+} satisfying (3.1) and variants of such matrices are also called conditional inverses, pseudo inverses, g-inverses and Rao inverses in the mathematical literature. It should be noted that if the matrix A is square and of full rank, then the usual inverse of the matrix A which is denoted by A^{-1} exists and is also a generalized inverse.

The following argument due to Searle (1970) shows that the generalized inverse for A is not unique. Suppose that the matrix A is of rank r. There always exists square matrices P and Q such that:

$$PAQ = \Delta = \begin{bmatrix} D_{r \times r} & 0_{r \times (q-r)} \\ 0_{(p-r) \times r} & 0_{(p-r) \times (q-r)} \end{bmatrix} \qquad (3.2)$$

where $D_{r \times r}$ is a diagonal matrix and the remaining elements of Δ are equal to zero. The matrices P and Q are products of elementary row and column operations. Since the matrix A was assumed to be of rank r, all the diagonal elements of $D_{r \times r}$ are nonzero and the inverse of $D_{r \times r}$ exists. Therefore it is possible to define a matrix G equal to:

$$G = Q\Delta^{-}P \qquad (3.3)$$

where:

$$\Delta^{-} = \begin{bmatrix} D_{r \times r}^{-1} & 0_{r \times (p-r)} \\ 0_{(q-r) \times r} & 0_{(q-r) \times (p-r)} \end{bmatrix} \qquad (3.4)$$

That G is a generalized inverse of the matrix A follows from:

$$\begin{aligned} AGA &= A(Q\Delta^{-}P)A \\ &= (P^{-1}\Delta Q^{-1})(Q\Delta^{-}P)(P^{-1}\Delta Q^{-1}) \\ &= P^{-1}\Delta Q^{-1} \\ &= A \quad . \end{aligned} \tag{3.5}$$

Since by definition neither P nor Q are unique, it follows that neither Δ^{-} nor G are unique. Therefore, the generalized inverse of A defined by (3.3) is not unique.

A number of algorithms have been developed for computing generalized inverses for an arbitrary p x q matrix A. The following algorithm which was described by Rao (1963) and Searle (1970) is based upon the argument employed above to show that generalized inverses are not unique. Suppose that a generalized inverse of rank r is required for the matrix A. The rank of A should be greater than or equal to r. Let B_{11} be a non-singular r x r minor of A. If B_{11} is not the leading r x r minor, there exist column and row operators which transform the matrix A such that B_{11} becomes the leading r x r minor. Letting R and S denote the corresponding transformation matrices, it follows that:

$$\begin{aligned} RAS &= B \\ &= \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \quad . \end{aligned} \tag{3.6}$$

If B_{11} is already the leading r x r minor of A, R and S will be identity matrices.

It is possible to define a matrix F similar to the matrix

Δ^- such that:

$$F = \begin{bmatrix} B_{11}^{-1} & 0_{r \times (q-r)} \\ 0_{(p-r) \times r} & 0_{(p-r) \times (q-r)} \end{bmatrix} . \quad (3.7)$$

Consider the matrix expression:

$$BFB = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{21}B_{11}^{-1}B_{12} \end{bmatrix} . \quad (3.8)$$

As a result of the partition of the matrix B, it follows that there exists a matrix K such that:

$$\begin{bmatrix} B_{21} & B_{22} \end{bmatrix} = K \begin{bmatrix} B_{11} & B_{12} \end{bmatrix} . \quad (3.9)$$

It follows from (3.9) that:

$$\begin{aligned} B_{22} &= KB_{12} \\ &= B_{21}B_{11}^{-1}B_{12} . \end{aligned} \quad (3.10)$$

Substituting (3.10) into (3.8), it may be seen that:

$$BFB = B \quad (3.11)$$

so that F is a generalized inverse of the transformed matrix B. Combining (3.6) and (3.11),

$$\begin{aligned} A &= R^{-1}BS^{-1} \\ &= R^{-1}(BFB)S^{-1} \\ &= A(SFR)A . \end{aligned} \quad (3.12)$$

Therefore, a generalized inverse of the matrix A is given by:

$$A^+ = SFR . \quad (3.13)$$

Searle (1970) provided a number of numerical examples to illustrate this algorithm.

Marquardt (1970) noted that the eigenvalues for the X'X matrix may be classified into the three groups: eigenvalues substantially greater than zero; eigenvalues slightly greater

than zero; and eigenvalues equal to zero. In practice, it is usually impossible to separate the last two groups of eigenvalues due to round-off errors. Marquardt (1970) suggested that a criterion could be adopted for establishing the number of essentially zero eigenvalues for any practical problem. He argued that a criterion should be employed which will assign a rank to the $X'X$ matrix such that the assigned rank includes 'substantially' all the variation in X. To this end, Marquardt (1970) recommended that the $X'X$ matrix be assigned the rank r where r is the largest integer such that:

$$(\mathrm{tr}(\Lambda))^{-1} \sum_{j=1}^{r-1} \lambda_j < \omega \tag{3.14}$$

where ω is an arbitrary constant. He argued that in most practical applications ω would lie in the range of 10^{-1} to 10^{-7}.

Suppose that the $X'X$ matrix is assigned a rank r which is less than or equal to p. In this case, it is assumed that:

$$\lambda_1 = \lambda_2 = \ldots = \lambda_{p-r} = 0 \tag{3.15}$$

and:

$$0 < \lambda_{p-r+1} \leq \lambda_{p-r+2} \leq \ldots \leq \lambda_p \ . \tag{3.16}$$

(3.15) and (3.16) imply a partition of the diagonal matrix Λ of the form:

$$\Lambda = \begin{bmatrix} \Lambda_{(p-r)} & 0 \\ 0 & \Lambda_r \end{bmatrix} \tag{3.17}$$

where: $\Lambda_{(p-r)}$ is a $(p-r) \times (p-r)$ matrix of zeros and Λ_r is a $r \times r$ diagonal matrix containing the nonzero eigenvalues. In addition, suppose that the orthogonal matrix P is partitioned such that:

$$P = \left[P_{(p-r)} \quad P_r \right] \tag{3.18}$$

where $P_{(p-r)}$ is a p x (p − r) matrix and P_r is a p x r matrix. A generalized inverse of rank r for the X′X matrix may be formed in an analogous manner to (3.13). In this case:

$$\begin{aligned}
(X'X)_r^+ &= P \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_r^{-1} \end{bmatrix} P' \\
&= \left[P_{(p-r)} \quad P_r \right] \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_r^{-1} \end{bmatrix} \begin{bmatrix} P_{(p-r)}' \\ P_r' \end{bmatrix} \\
&= P_r \Lambda_r^{-1} P_r' \\
&= \sum_{j=p-r+1}^{p} \frac{1}{\lambda_j} S_j S_j'
\end{aligned} \tag{3.19}$$

where S_j is the j'th column of the orthogonal matrix P.

In order to illustrate the generalized inverse $(X'X)_r^+$ defined by (3.19), consider the French imports model presented in Chapter 2. Before calculating the ordinary least squares and ridge regression estimates of the parameters in the model, the X′X matrix was standardized so that the X′X matrix was in it's correlation form. The X′X matrix was calculated to be:

$$X'X = \begin{bmatrix} 1.0000 & 0.1687 & 0.9973 \\ 0.1687 & 1.0000 & 0.1545 \\ 0.9973 & 0.1545 & 1.0000 \end{bmatrix} \tag{3.20}$$

The three eigenvalues for this X′X matrix are: 2.0472, 0.9502 and 0.0026. The two largest eigenvalues account for 99.91% of the sum of the three eigenvalues for the X′X matrix. The orthogonal matrix containing the corresponding eigenvectors is:

$$P = \begin{bmatrix} 0.6916 & -0.1434 & -0.7079 \\ 0.2132 & 0.9770 & 0.0104 \\ 0.6901 & -0.1581 & 0.7062 \end{bmatrix} \quad . \tag{3.21}$$

Based upon (3.19), the generalized inverses corresponding to the three different possible ranks for the $X'X$ matrix are:

$$(X'X)_1^+ = \begin{bmatrix} 0.2336 & 0.0720 & 0.2331 \\ 0.0720 & 0.0222 & 0.0719 \\ 0.2331 & 0.0719 & 0.2327 \end{bmatrix} \tag{3.22}$$

$$(X'X)_2^+ = \begin{bmatrix} 0.2553 & -0.0754 & 0.2570 \\ -0.0754 & 1.0266 & -0.0907 \\ 0.2570 & -0.0907 & 0.2589 \end{bmatrix} \tag{3.23}$$

and:

$$(X'X)_3^+ = \begin{bmatrix} 194.5176 & -2.9155 & -193.5452 \\ -2.9155 & 1.0681 & 2.7429 \\ -193.5452 & 2.7427 & 193.6022 \end{bmatrix} \quad . \tag{3.24}$$

It can be seen from the above that there are large reductions in the magnitudes of the elements of $(X'X)_r^+$ when the assigned rank for the $X'X$ matrix is reduced from three to two. When the assigned rank is reduced to one, the changes in the elements of the generalized inverses are not nearly as dramatic. This behavour is indicative of the presence of one very small eigenvalue and two moderate to large sized eigenvalues.

Marquardt (1970) utilized the generalized inverse $(X'X)_r^+$ to define a class of generalized least squares estimators:

$$\hat{\beta}_r^+ = (X'X)_r^+ X'Y \quad . \tag{3.25}$$

$\hat{\beta}_r^+$ shares a number of properties in common with the ordinary least squares estimator. $\hat{\beta}_r^+$ defaults to $\hat{\beta}$ when the $X'X$ matrix is assigned the rank p. It was noted earlier that $\hat{\beta}$ minimizes the residual sum of squares function for all p-dimensional vectors b. An analogous

result for the generalized least squares estimator is provided by:

Theorem 3.1: The generalized least squares estimator $\hat{\beta}_r^+$ minimizes the residual sum of squares function $\Phi(b)$ for all estimators b of β within the r-dimensional subspace spanned by P_r.

Proof: Suppose that:

$$\xi = XP_r \tag{3.26}$$

denotes the projection of X onto the eigenvectors which form the last r columns of P. The normal equation (1.5) may be transformed so that:

$$(P_r'X'XP_r)P_r'b = P_r'X'Y \tag{3.27}$$

or:

$$(\xi'\xi)P_r'b = \xi'Y \quad . \tag{3.28}$$

In addition, let:

$$\beta^* = P_r'b \tag{3.29}$$

denote the projection of b onto the eigenvector coordinates so that:

$$(\xi'\xi)\beta^* = \xi'Y \quad . \tag{3.30}$$

Equation (3.30) yields the ordinary least squares solution. This solution minimizes the residual sum of squares within the subspace spanned by P_r. It should be noted that:

$$(\xi'\xi)\beta^* = \Lambda_r\beta^* \tag{3.31}$$

so that:

$$\begin{aligned}\beta^* &= \Lambda_r^{-1}\xi'Y \\ &= \Lambda_r^{-1}P_r'X'Y \quad .\end{aligned} \tag{3.32}$$

The solution for (3.30) may be expressed in terms of the original

coordinates as:

$$\begin{aligned} b &= P_r\beta^* \\ &= P_r\Lambda_r^{-1}P_r'X'Y \\ &= (X'X)_r^+X'Y \\ &= \hat{\beta}_r^+ \end{aligned} \tag{3.33}$$

As a result of (3.33), it can be seen that $\hat{\beta}_r^+$ minimizes the residual sum of squares function within the subspace spanned by P_r.

The generalized least squares estimator may be expressed as:

$$\begin{aligned} \hat{\beta}_r^+ &= (X'X)_r^+X'Y \\ &= (X'X)_r^+(X'X)\hat{\beta} \\ &= W_r\hat{\beta} \end{aligned} \tag{3.34}$$

where:

$$\begin{aligned} W_r &= (X'X)_r^+(X'X) \\ &= (P_r\Lambda_r^{-1}P_r')(P\Lambda P') \\ &= P_r\Lambda_r^{-1}P_r'(P_{(p-r)}\Lambda_{(p-r)}P_{(p-r)}' + P_r\Lambda_rP_r') \end{aligned} \tag{3.35}$$

It follows from (3.35) that:

$$\begin{aligned} E(\hat{\beta}_r^+) &= W_rE(\hat{\beta}) \\ &= P_r\Lambda_r^{-1}P_r'(P_{(p-r)}\Lambda_{(p-r)}P_{(p-r)}' \\ &\qquad + P_r\Lambda_rP_r')\beta \end{aligned} \tag{3.36}$$

As a result of (3.36), it can be seen that $\hat{\beta}_r^+$ is a biased estimator of β if $\Lambda_{(p-r)}$ is a non-null matrix

Chipman (1964) noted that if the matrix X has a rank less than p, then there cannot exist an unbiased linear estimator of the form:

$$b = AY \tag{3.37}$$

for β. By the linearity property of the expectation operator:

$$E(b) = AXE(Y)$$
$$= AX\beta \quad . \tag{3.38}$$

Chipman (1964) pointed out that b is an unbiased estimator of β if and only if (3.38) equals β for all values of β. This implies that:

$$AX = I_p \tag{3.39}$$

or:

$$\text{rank}(AX) = p \quad . \tag{3.40}$$

However, the rank of the product of two matrices cannot exceed the rank of either matrix. Since the rank of the matrix X is assumed to be less than p, it follows that no unbiased estimator of the form (3.37) exists for β. In the same spirit as Chipman (1964), $\hat{\beta}_r^+$ is said to be conditionally unbiased relative to the constraints implied by the columns of $P_{(p-r)}$ if $\Lambda_{(p-r)}$ is a null matrix.

The variance-covariance matrix for $\hat{\beta}_r^+$ may be derived by utilizing (3.34) and (3.35):

$$\begin{aligned}
\text{Var}(\hat{\beta}_r^+) &= E(\hat{\beta}_r^+ - E(\hat{\beta}_r^+))(\hat{\beta}_r^+ - E(\hat{\beta}_r^+))' \\
&= W_r E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' W_r' \\
&= \sigma^2 W_r (X'X)^{-1} W_r' \\
&= \sigma^2 (P_r \Lambda_r^{-1} P_r')(X'X)^{-1} (P_r \Lambda_r^{-1} P_r')' \\
&= \sigma^2 P_r \Lambda_r^{-1} (P_r'(X'X)^{-1} P_r) \Lambda_r^{-1} P_r' \\
&\geq \sigma^2 P_r \Lambda_r^{-1} P_r' \quad .
\end{aligned} \tag{3.41}$$

Invoking (1.25), the mean squared error function for $\hat{\beta}_r^+$ may

be expressed as:

$$MSE(\hat{\beta}_r^+) = \sum_{i=1}^{p} Var(\hat{\beta}_{ir}^+) + (E(\hat{\beta}_r^+) - \beta)'(E(\hat{\beta}_r^+) - \beta)$$

$$= tr(Var(\hat{\beta}_r^+)) + \beta'(W_r - I_p)'(W_r - I_p)\beta \quad . \quad (3.42)$$

The behavior of the mean squared error function for $\hat{\beta}_r^+$ is summarized by the following two theorems due to Marquardt (1970):

<u>Theorem 3.2</u>: The variance term of the mean squared error function for $\hat{\beta}_r^+$ is an increasing function of the assigned rank r.

<u>Proof</u>: Appealling to (3.41), the variance term of the mean squared error function for $\hat{\beta}_r^+$ may be expressed as:

$$tr(Var(\hat{\beta}_r^+)) = \sigma^2 tr(P_r \Lambda_r^{-1} P_r')$$

$$= \sigma^2 tr(\Lambda_r^{-1})$$

$$= \sigma^2 \sum_{i=p-r+1}^{p} \frac{1}{\lambda_i} \quad . \quad (3.43)$$

Since the largest r eigenvalues of the X'X matrix are assumed to be nonzero, it follows that (3.43) increases monotonically with r.

<u>Theorem 3.3</u>: The bias term in $MSE(\hat{\beta}_r^+)$ is a monotonically decreasing function of r.

<u>Proof</u>: Substituting (3.36) into (3.42), the squared bias term in the mean squared error function becomes:

$$(E(\hat{\beta}_r^+) - \beta)'(E(\hat{\beta}_r^+) - \beta)$$

$$= \beta'(P_r P_r' - I_p)'(P_r P_r' - I_p)\beta$$

$$= \beta'(-P_{(p-r)} P_{(p-r)}')'(-P_{(p-r)} P_{(p-r)}')\beta$$

$$= \beta' P_{(p-r)} P_{(p-r)}' \beta \quad . \quad (3.44)$$

since:

$$PP' = P_r P_r' + P_{(p-r)} P_{(p-r)}' \quad . \quad (3.45)$$

In the same manner as (3.26), let:

$$\psi = P_{(p-r)}'\beta \tag{3.46}$$

denote the projection of β onto the subspace of R^p spanned by $P_{(p-r)}$. By definition, ψ is a $(p-r)$ element vector whose i'th element is α_i. Therefore, the squared bias for $\hat{\beta}_r^+$ becomes:

$$(E(\hat{\beta}_r^+) - \beta)'(E(\hat{\beta}_r^+) - \beta) = \psi'\psi$$

$$= \sum_{i=1}^{p-r} \alpha_i^2 \quad . \tag{3.47}$$

Since each component of ψ is independent of r, it follows that the bias term in the mean squared error function for $\hat{\beta}_r^+$ is a monotonically decreasing function of r.

As a direct consequence of Theorems 3.2 and 3.3, it is possible to derive the required condition for the mean squared error admissibility of $\hat{\beta}_r^+$:

<u>Theorem 3.4</u>: The generalized least squares estimator $\hat{\beta}_r^+$ is mean squared error admissible if:

$$\sum_{i=1}^{p-r} \frac{1}{\lambda_i} > \frac{1}{\sigma^2} \sum_{j=1}^{p-r} \alpha_i^2 \quad . \tag{3.48}$$

<u>Proof</u>: Combining (3.43) and (3.47), the mean squared error function for $\hat{\beta}_r^+$ becomes:

$$MSE(\hat{\beta}_r^+) = tr(Var(\hat{\beta}_r^+)) + \beta'(W_r - I_p)'(W_r - I_p)\beta$$

$$= \sigma^2 \sum_{i=p-r+1}^{p} \frac{1}{\lambda_i} + \sum_{i=1}^{p-r} \alpha_i^2 \quad . \tag{3.49}$$

The mean squared error function for $\hat{\beta}$ is provided by (1.28). Therefore, $\hat{\beta}_r^+$ will be mean squared error admissible if and only if:

$$MSE(\hat{\beta}_r^+) = \sigma^2 \sum_{i=p-r+1}^{p} \frac{1}{\lambda_i} + \sum_{i=1}^{p-r} \alpha_i^2$$

$$< \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}$$

$$= MSE(\hat{\beta}) \tag{3.50}$$

or:

$$\sum_{i=1}^{p-r} \frac{1}{\lambda_i} > \frac{1}{\sigma^2} \sum_{i=1}^{p-r} \alpha_i^2 \quad . \tag{3.51}$$

A more restrictive but more useful mean squared error admissibility condition for $\hat{\beta}_r^+$ is provided by:

<u>Corollary 3.4.1</u>: A sufficient condition for $\hat{\beta}_r^+$ to be mean squared error admissible is:

$$\sum_{i=1}^{p-r} \frac{1}{\lambda_i} > \frac{1}{\sigma^2} \beta'\beta \tag{3.52}$$

<u>Proof</u>: By definition ψ is the projection of β onto the subspace of R^p spanned by $P_{(p-r)}$. Therefore,

$$\begin{aligned} \beta'\beta &= \beta'PP'\beta \\ &\geq \beta'P_{(p-r)}P_{(p-r)}'\beta \\ &= \psi'\psi \\ &= \sum_{i=1}^{p-r} \alpha_i^2 \quad . \end{aligned} \tag{3.53}$$

As a result of this last inequality, if:

$$\sum_{i=1}^{p-r} \frac{1}{\lambda_i} > \frac{1}{\sigma^2} \beta'\beta \ , \tag{3.54}$$

it also follows that:

$$\sum_{i=1}^{p-r} \frac{1}{\lambda_i} > \frac{1}{\sigma^2} \sum_{i=1}^{p-r} \alpha_i^2 \tag{3.55}$$

so that (3.52) provides a sufficient condition for the mean squared admissibility of $\hat{\beta}_r^+$.

It was mentioned earlier that Marquardt (1970) proposed assigning the $X'X$ matrix a rank by examining the relative sizes of each eigenvalue λ_i. In a later paper, Marquardt and Snee (1975) suggested that the rank may be choosen by constructing graphs of the generalized least squares estimates as functions of the assigned rank r. These graphs would be interpreted in much the same manner as the ridge trace. It should be noted that this approach is probably only practical for multi-factor problems. Hocking, Speed and Lynn (1976) suggested that the assigned rank be choosen by minimizing estimates of the mean squared error for $\hat{\beta}_r^+$. To this end, they recommended evaluating (3.49) for various values of r using the ordinary least squares estimates of α and σ^2.

In the proof of Lemma 2.1, it was shown that:

$$|\hat{\beta}_t^*| < |\hat{\beta}_s^*| \tag{3.56}$$

for all:

$$0 \leq s < t \quad . \tag{3.57}$$

An analogous result for the generalized least squares estimator is provided by the theorem:

<u>Theorem 3.5</u>: Suppose that $\hat{\beta}_r^+$ is defined as in (3.25). $\|\hat{\beta}_r^+\|$ is a stepwise increasing function of r.

<u>Proof</u>: Consider any assigned rank r less than or equal to p:

$$\|\hat{\beta}_r^+\|^2 = \hat{\beta}_r^{+\prime}\hat{\beta}_r^+$$

$$= (P_r\Lambda_r^{-1}P_r'X'Y)'(P_r\Lambda_r^{-1}P_r'X'Y)$$

$$= Y'XP_r\Lambda_r^{-2}P_r'X'Y$$

$$= \Lambda_r^{-2}(Y'XP_rP_r'X'Y) \qquad (3.58)$$

Let g_i denote the i'th element of the vector $X'Y$ and P_{ij} denote the (i,j)'th element of the orthogonal matrix P. Adopting this notation, (3.58) becomes:

$$\|\hat{\beta}_r^+\| = \sum_{j=p-r+1}^{p} \lambda_j^{-2}\{ \sum_{i=1}^{p} (g_iP_{ij})^2 \} \quad . \qquad (3.59)$$

It can be seen from (3.59) that $\|\hat{\beta}_r^+\|^2$ may be expressed as the sum of r terms each of which is independent of the assumed rank r. The j'th term of the summation is dependent upon λ_j, $X'Y$ and the j'th eigenvalue. Therefore, it follows that $\|\hat{\beta}_r^+\|$ is a stepwise increasing function of r.

Marquardt (1970) argued that the assumption of an integral rank for the $X'X$ matrix often imposes an unrealistic constraint upon the generalized least squares estimator $\hat{\beta}_r^+$. He suggested that the optimum rank is usually non-integral. As a result, Marquardt (1970) proposed extending the definition of $\hat{\beta}_r^+$ to allow the assigned rank to be a continuous variable in the interval (0,p]. Suppose that the $X'X$ matrix is assigned the rank:

$$r = k + t \qquad (3.60)$$

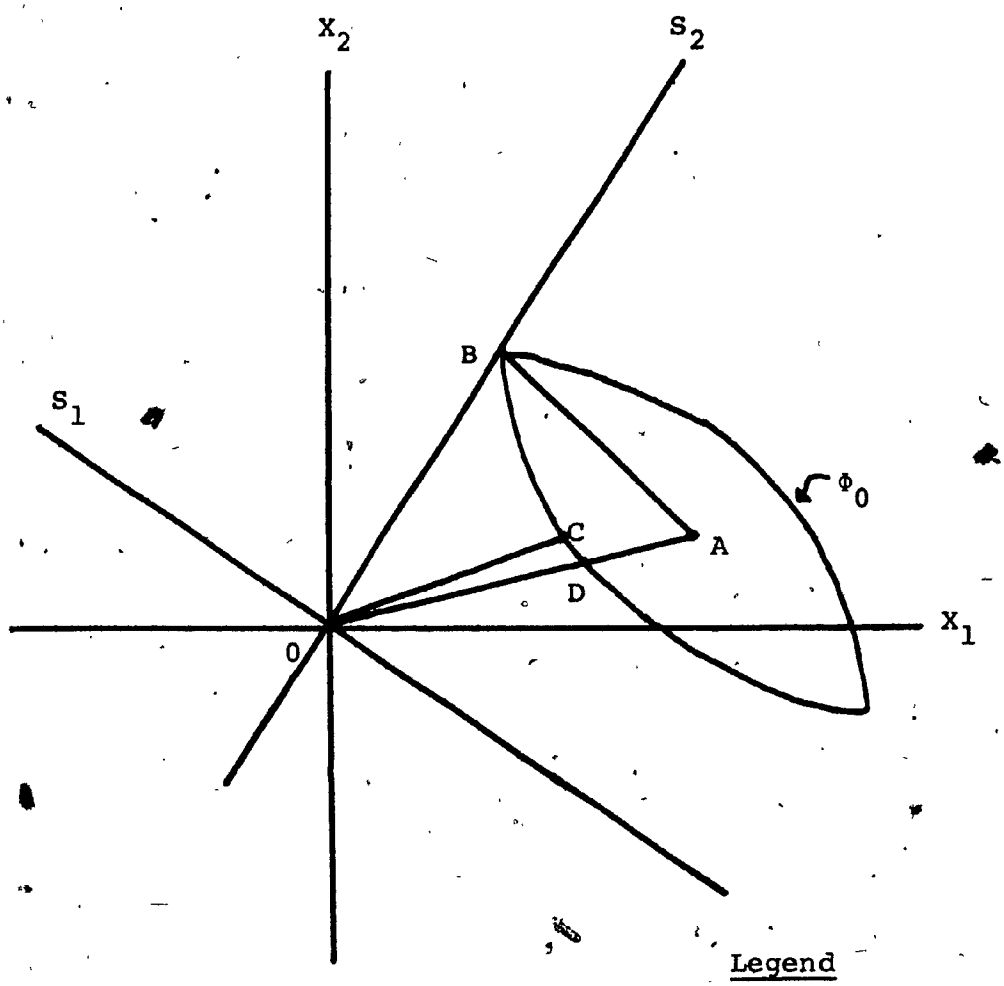
where k is a non-negative integer less than or equal to p and t lies in the interval [0,1). Marquardt (1970) proposed that the generalized inverse for (3.25) be defined as:

$$(X'X)_r^+ = \sum_{j=p-k+1}^{p} \frac{1}{\lambda_j} S_jS_j' + \frac{t}{\lambda_{(p-k)}} S_{(p-k)}S_{(p-k)}' \quad . \qquad \ldots (3.61)$$

In order to illustrate the geometrical properties of the generalized least squares estimator, consider the hypothetical two-dimensional sum of squares contour which is provided in Figure 3. Suppose that the point A represents the minimum of the residual sum of squares function for all estimates b. A corresponds to the ordinary least squares solution $\hat{\beta}$. Let Φ_0 denote a contour of constant residual sum of squares. Since C is the point on Φ_0 which is closest to the origin, C represents the ridge estimate corresponding to this contour. Suppose that S_1 and S_2 represent the normalized eigenvectors for the $X'X$ matrix. If the assigned rank for the $X'X$ matrix is one, the generalized least squares solution is given by the point on S_2 which results in the minimum residual sum of squares. This point is denoted by B in Figure 3. On the other hand, if the $X'X$ matrix is assigned the rank two, the generalized least squares estimator corresponds to the minimum of the residual sum of squares function in the plane defined by S_1 and S_2. In this case, $\hat{\beta}_2^+$ equals $\hat{\beta}$. If a continuous rank between one and two is allowed, $\hat{\beta}_r^+$ follows the path defined by BA.

Hocking, Speed and Lynn (1976) investigated the properties of the generalized least squares estimator in their survey of biased estimators. They referred to the estimator as a 'principal component' estimator and a 'fractional rank' estimator depending upon whether an integral or a continuous rank is assumed for the $X'X$ matrix. Adopting the canonical form of the general model,

Figure 3 - A Geometrical Representation Of The Biased Estimators



Legend

Point	Estimator
A	Ordinary Least Squares
B	Generalized Least Squares (Rank = 1)
C	Ridge
D	Shrinkage

Hocking, Speed and Lynn (1976) expressed the generalized least squares estimator of rank k as:

$$\begin{aligned}\hat{\alpha}_k^+ &= P'\hat{\beta}_k^+ \\ &= P'(P_k\Lambda_k^{-1}P_k')X'Y \\ &= \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_k^{-1} \end{bmatrix} X^{*\prime}Y \\ &= \hat{\theta}_k \end{aligned} \tag{3.62}$$

where $\hat{\theta}_k$ is a vector whose first $(p - k)$ components are zero and last k components agree with $\hat{\alpha}$. Combining (3.61) and (3.62), it can be seen that the canonical form of the generalized least squares estimator may be expressed as:

$$\hat{\alpha}_r^+ = (1 - t)\hat{\theta}_k + t\hat{\theta}_{(k-1)} \tag{3.63}$$

when a continuous rank of the form (3.60) is assumed for the $X'X$ matrix. Suppose that B is a diagonal matrix whose i'th element is given by:

$$b_i = \begin{cases} 0 & \text{for } i = 1,2,3, \ldots ,(p - k - 1) \\ t & \text{for } i = p - k \\ 1 & \text{otherwise} \end{cases} \tag{3.64}$$

Hocking, Speed and Lynn (1976) observed that:

$$\hat{\alpha}_r^+ = B\hat{\alpha} \tag{3.65}$$

Adopting Hocking, Speed and Lynn's (1976) formulation of the generalized least squares estimator, the mean squared error function for $\hat{\alpha}_r^+$ may be written as:

$$MSE(\hat{\alpha}_r^+) = MSE(B\hat{\alpha})$$

$$= \sum_{i=1}^{p} b_i^2 \, \mathrm{Var}(\hat{\alpha}_i) + \sum_{i=1}^{p} (b_i - 1)^2 \alpha_i^2$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{b_i^2}{\lambda_i} + \sum_{i=1}^{p} (b_i - 1)^2 \alpha_i^2 \qquad (3.66)$$

since:

$$\begin{aligned} \mathrm{Var}(\hat{\alpha}) &= \mathrm{Var}(P'\hat{\beta}) \\ &= \sigma^2 P'(X'X)^{-1} P \\ &= \sigma^2 \Lambda^{-1} \quad . \end{aligned} \qquad (3.67)$$

Suppose that the $X'X$ matrix is assigned an integral rank k, so that:

$$b_i = \begin{cases} 0 & \text{for } i = 1,2,3, \ldots, (p-k) \\ 1 & \text{otherwise} \end{cases} \quad . \qquad (3.68)$$

In this case, (3.66) reduces to:

$$\mathrm{MSE}(\hat{\alpha}_r^+) = \sigma^2 \sum_{i=p-k+1}^{p} \frac{1}{\lambda_i} + \sum_{i=1}^{p-k} \alpha_i^2 \quad . \qquad (3.69)$$

It can be seen from (3.69) that the resultant increase in the mean squared error function when the assigned rank for the $X'X$ matrix is decreased from p to k is given by:

$$\mathrm{MSE}(\hat{\alpha}_r^+) - \mathrm{MSE}(\hat{\alpha}) = \sum_{i=1}^{p-k} \{\alpha_i^2 - \lambda_i^{-1}\sigma^2\} \quad . \qquad (3.70)$$

Hocking, Speed and Lynn (1976) proposed a two-stage procedure for determining fractional ranks of the form (3.60). They suggested that the integral portion of the rank be determined by choosing the value of k which minimizes the estimated increase in the mean squared error function which results when $\hat{\alpha}_r^+$ is employed instead of $\hat{\alpha}$. For this purpose, Hocking, Speed and Lynn

(1976) recommended that the ordinary least squares estimates of α and σ^2 be substituted into (3.70). Assuming that the integral portion of the rank is choosen to be k, the mean squared error function for the generalized least squares estimator $\hat{\alpha}_r^+$ becomes:

$$MSE(\hat{\alpha}_r^+) = \sigma^2 \frac{t}{\lambda_{(p-r)}} + \sigma^2 \sum_{i=p-k+1}^{p} \frac{1}{\lambda_i} + \sum_{i=1}^{p-k+1} \alpha_i^2 + (t-1)^2 \alpha_{(p-k)}^2 \quad . \tag{3.71}$$

Minimizing (3.71) with respect to t leads to the optimum t for a fixed value of k:

$$t = \frac{\lambda_{(p-k)} \alpha_{(p-k)}^2}{\sigma^2 + \lambda_{(p-k)} \alpha_{(p-k)}^2} \tag{3.72}$$

Hocking, Speed and Lynn (1976) proposed an iterative procedure for estimating t based upon (3.72). Let $\hat{t}_{(j)}$ denote the estimate of t corresponding to the j'th iteration. Substituting (3.65) and the ordinary least squares estimate of σ^2 into (3.72), Hocking, Speed and Lynn (1976) constructed the iterative formula:

$$\hat{t}_{(j+1)} = \frac{\lambda_{(p-k)} \hat{t}_{(j)}^2 \hat{\alpha}_{(p-k)}^2}{\sigma^2 + \lambda_{(p-k)} \hat{t}_{(j)}^2 \hat{\alpha}_{(p-k)}^2}$$

$$= \frac{\hat{t}_{(j)}^2}{\hat{t}_{(j)}^2 + L} \tag{3.73}$$

where:

$$L = \hat{\sigma}^2 \lambda_{(p-k)}^{-1} \hat{\alpha}_{(p-k)}^{-2} \tag{3.74}$$

Suppose that $t_{(0)}$ is assigned the value $(1 + L)^{-1}$. The convergence theorem from the Appendix may be invoked to obtain the limiting values of $t_{(j)}$:

$$t^* = \begin{cases} 0 & \text{if } L > \frac{1}{4} \\ \frac{1}{2} + (\frac{1}{4} - L)^{\frac{1}{2}} & \text{otherwise} \end{cases} \quad . \tag{3.75}$$

Utilizing (3.75), Hocking, Speed and Lynn's (1976) limiting solution for the generalized least squares estimator becomes:

$$\hat{\alpha}^{+}_{(k + t^*)} = (1 - t^*)\hat{\theta}_k + t^*\hat{\theta}_{(k - 1)} \quad . \tag{3.76}$$

Marquardt and Snee (1975) provided a number of numerical examples to compare the use of the continuous rank generalized least squares and ridge estimators. They described a model in which the percentage conversion of n-heptane to acetylene was considered a function of the reactor tempature, the mole ratio of H_2 to n-heptane and the contact time. A number of quadratic and cross-product terms were included in the model. Marquardt and Snee (1975) also presented the results of a simulated 2^3 factorial experiment. They found that both the generalized least squares and the ridge estimators produced smaller predictive standard errors than the ordinary least squares estimator as the degree of multicollinearity in the experiments was increased. Marquardt and Snee (1975) employed their examples to demonstrate that the continuous rank generalized least squares and ridge estimators can often be utilized to produce similar parameter estimates.

The generalized least squares estimator attempts to cope with the effects of multicollinearity between the explanatory variables by replacing the inverse of the $X'X$ matrix with a generalized inverse of a rank less than p. It was mentioned earlier that a second approach to dealing with the effects of the multicollinearity is to retain the assumption that the $X'X$ matrix is of full rank but to apply a linear transformation to shrink the ordinary least squares estimates. Suppose that C represents the matrix corresponding to a transformation which shrinks the ordinary least squares estimates. It is possible to define a class of shrinkage estimators:

$$\begin{aligned} b(C) &= C(X'X)^{-1}X'Y \\ &= C\hat{\beta} \end{aligned} \tag{3.77}$$

for β. For example, $b(Z_k)$ corresponds to the class of ridge estimators when Z_k is defined by (2.5).

In most of the remainder of this section, the matrix C in (3.77) will be assumed to be a diagonal matrix of the form:

$$C = cI_p \tag{3.78}$$

where the shrinkage factor c satisfies:

$$0 \leq c \leq 1 \tag{3.79}$$

In order to simplify the notation, let the shrinkage estimator which is obtained by substituting (3.78) into (3.77) be denoted by:

$$\tilde{\beta}_s = c\hat{\beta} \tag{3.80}$$

A geometrical comparison of the shrinkage estimator $\tilde{\beta}_s$ with the

other biased estimators described earlier is provided in Figure 3. As c varies between zero and one, $\tilde{\beta}_s$ follows the straight line path between the origin and the ordinary least squares estimate which is denoted by A. It was mentioned above that Φ_0 denotes a contour of constant residual sum of squares. The point D corresponds to the shrinkage estimate whose residual sum of squares is equal to the residual sum of squares of all the other estimates on that contour. In particular, the shrinkage estimate corresponding to D has the same residual sum of squares as the generalized least squares and ridge estimates denoted by B and C respectively.

Mayer and Willke (1973) proposed a classification of the shrinkage estimators according to the form of their shrinkage factors. They called $\tilde{\beta}_s$ a 'deterministically' shrunken estimator if c is a fixed scalar so that c is independent of the observed values of Y. Otherwise, Mayer and Willke (1973) called $\tilde{\beta}_s$ a 'stochastically' shrunken estimator. If c is fixed with repect to Y, the first two moments of $\tilde{\beta}_s$ are given by:

$$E(\tilde{\beta}_s) = E(c\hat{\beta})$$
$$= c\beta \tag{3.81}$$

and:

$$Var(\tilde{\beta}_s) = Var(c\hat{\beta})$$
$$= \sigma^2 c^2 (X'X)^{-1} \tag{3.82}$$

respectively. It should be noted that if $\tilde{\beta}_s$ is a 'stochastically' shrunken estimator, the moments of $\tilde{\beta}_s$ depend upon the form of the relationship between c and Y so that they cannot be given in

general.

Mayer and Willke (1973) provided the following theorem to demonstrate the existence of a mean squared error admissible 'deterministically' shrunken estimator for β:

<u>Theorem 3.6</u>: For every value of β, there exists a c satisfying (3.79) such that $\tilde{\beta}_s$ is mean squared error admissible.

<u>Proof</u>: By definition of the mean squared error function:

$$\begin{aligned} MSE(\tilde{\beta}_s) &= E((c\hat{\beta} - \beta)'(c\hat{\beta} - \beta)) \\ &= E((c\hat{\beta} - c\beta)'(c\hat{\beta} - c\beta)) + (c\beta - \beta)'(c\beta - \beta) \\ &= c^2 MSE(\hat{\beta}) + (c - 1)^2 \beta'\beta \end{aligned} \tag{3.83}$$

$\tilde{\beta}_s$ is mean squared error admissible if and only if:

$$\begin{aligned} MSE(\tilde{\beta}_s) &= c^2 MSE(\hat{\beta}) + (c - 1)^2 \beta'\beta \\ &< MSE(\hat{\beta}) \end{aligned} \tag{3.84}$$

or:

$$c > \frac{\beta'\beta - MSE(\hat{\beta})}{\beta'\beta + MSE(\hat{\beta})} \tag{3.85}$$

Since:

$$\beta'\beta - MSE(\hat{\beta}) < \beta'\beta + MSE(\hat{\beta}) \tag{3.86}$$

it follows that there always exists an interval for c in $[0,1]$ such that β_s is a mean squared error admissible estimator of β.

It was shown in Chapter 2 that the ridge estimator minimizes the residual sum of squares for a fixed Euclidean parameter length. Mayer and Willke (1973) provided a similar characterization for the 'deterministically' shrunken estimator. Instead of the Euclidean norm, they utilized the design dependent norm:

$$m_d(b) = b'(X'X)b \tag{3.87}$$

to measure the length of the estimated parameters. In this case, Lagrangian equation (2.16) becomes:

$$F^*(b) = (Y - Xb)'(Y - Xb) + k(b'(X'X)b - c^2) \quad . \tag{3.88}$$

$F^*(b)$ is minimized when:

$$\begin{aligned} \frac{d}{db} F^*(b) &= -2X'Y + 2X'Xb + 2kX'Xb \\ &= 2(1 + k)X'Xb - 2X'Y \\ &= 0 \end{aligned} \tag{3.89}$$

or:

$$\begin{aligned} b &= (1 + k)^{-1}(X'X)^{-1}X'Y \\ &= (1 + k)^{-1} \hat{\beta} \\ &= \tilde{\beta}_s \quad . \end{aligned} \tag{3.90}$$

It follows from (3.90) that the 'deterministically' shrunken estimator minimizes the residual sum of squares for a fixed value of the design dependent norm.

Up to this point in the discussion of shrinkage estimators, no mention has been made of how the shrinkage factors should be determined. James and Stein (1961) considered the problem of

estimating the location parameter θ for a p-dimensional normal variate Z whose unknown variance matrix is of the form $\sigma^2 I_p$. Further, they assumed that a single observation of Z and another random variable which is distributed as a $\sigma^2\chi_n^2$ variate independent of Z are available. Suppose that $\psi_i(Z)$ denotes an estimator of θ. James and Stein (1961) utilized the unweighted quadratic loss function:

$$L_1(\theta,\psi_i(Z)) = (\psi_i(Z) - \theta)'(\psi_i(Z) - \theta) \tag{3.91}$$

to measure the accuracy of $\psi_i(Z)$ as an estimator of θ. Adopting the decision-theoretic approach for selecting an estimator, $\psi_i(Z)$ is said to be admissible if there is no other estimator $\psi_j(Z)$ of θ whose expected quadratic loss or risk satisfies:

$$\begin{aligned} R(\theta;\psi_j(Z)) &= E(L_1(\theta,\psi_j(Z))) \\ &\leq E(L_1(\theta,\psi_i(Z))) \\ &= R(\theta;\psi_i(Z)) \end{aligned} \tag{3.92}$$

for all values of θ with strict inequality holding for at least one value of θ. Stein (1956) demonstrated that the usual estimator of θ:

$$\psi_1(Z) = Z \tag{3.93}$$

is admissible if and only if p is less than or equal to two.

James and Stein (1961) studied a class of estimators for θ of the form:

$$\psi_2(Z,S) = \left(1 - \frac{dS}{Z'Z}\right) Z \tag{3.94}$$

Where: d is a non-negative constant; Z is normally distributed

as above; and S is an observation of the $\sigma^2\chi_n^2$ variable which is independent of Z. They showed that $\psi_2(Z,S)$ has an expected unweighted quadratic loss equal to:

$$R(\theta;\psi_2(Z,S)) = E(L_1(\theta,\psi_2(Z,S)))$$
$$= \sigma^2 \{p - 2dn(p-2)E(\frac{1}{p-2+2K})$$
$$+ d^2n(n+2)E(\frac{1}{p-2+2K})\} \quad (3.95)$$

where K has a Poisson distribution with mean $(\theta'\theta)/2\sigma^2$. Minimizing (3.95) with respect to d results in the estimator:

$$\psi_3(Z,S) = (1 - \frac{p-2}{n+2}\frac{S}{Z'Z})Z \quad (3.96)$$

The corresponding risk for $\psi_3(Z,S)$ is given by:

$$R(\theta;\psi_3(Z,S)) = \sigma^2\{p - \frac{n}{n+2}(p-2)^2E(\frac{1}{p-2+2K})\}. \quad (3.97)$$

In comparison, the risk for $\psi_1(Z)$ is given by:

$$R(\theta;\psi_1(Z)) = E((\psi_1(Z) - \theta)'(\psi_1(Z) - \theta))$$
$$= E((Z-\theta)'(Z-\theta))$$
$$= tr(Var(Z))$$
$$= p\sigma^2 \quad (3.98)$$

By comparing (3.97) and (3.98), James and Stein (1961) showed that $\psi_3(Z,S)$ always has a smaller risk than $\psi_1(Z)$ assuming an unweighted quadratic loss function.

Bhattacharya (1966) generalized James and Stein's (1961) results by considering weighted loss functions of the form:

$$L_2(\theta,\psi_i(Z)) = (\psi_i(Z) - \theta)'D(\psi_i(Z) - \theta) \quad (3.99)$$

where D is a known p x p positive definite, symmetric matrix.

In addition, he utilized James and Stein's (1961) shrinkage estimator to provide an improvement over the ordinary least squares estimator. Suppose that the unobservable disturbances in (1.1) are assumed to be normally distributed and satisfy conditions (1.2). In order to simplify the notation, consider the canonical form of the general model defined by (2.20). In this case, the p-dimensional random vector Y is normally distributed with mean $X^*\alpha$ and variance $\sigma^2 I_n$. Further, assume that α is to be estimated by $\xi_i(Y)$ subject to the design dependent loss function:

$$L_2(\alpha, \xi_i(Y)) = (\xi_i(Y) - \alpha)'\Lambda(\xi_i(Y) - \alpha) \tag{3.100}$$

where Λ is the diagonal matrix defined by (1.27).

Since X^* is assumed to be a n x p matrix of rank p, it follows that there exists an orthogonal matrix A such that:

$$X^{*\prime}A = 0 \tag{3.101}$$

Consider the transformations defined by:

$$W = A'Y \tag{3.102}$$

and:

$$V = W'W \tag{3.103}$$

respectively. V is distributed as a $\sigma^2\chi^2_{(n-p)}$ variate independent of $\hat{\alpha}$. Suppose that an estimator $\hat{\alpha}^*$ is defined by:

$$\hat{\alpha}^* = \Lambda^{\frac{1}{2}}\hat{\alpha} \tag{3.104}$$

where $\Lambda^{\frac{1}{2}}$ is a diagonal matrix whose diagonal elements are equal to the square roots of the eigenvalues for the $X'X$ matrix. $\hat{\alpha}^*$

is normally distributed with mean:

$$\theta = \Lambda^{\frac{1}{2}} \hat{\alpha} \tag{3.105}$$

and variance $\sigma^2 I_p$. Since $\hat{\alpha}^*$ is formed by applying a linear transformation which does not depend upon V to $\hat{\alpha}$, $\hat{\alpha}^*$ and V are independent.

Now consider the problem of estimating θ subject to the unweighted quadratic loss function:

$$L_1(\theta,\psi_i(Y)) = (\psi_i(Y) - \theta)'(\psi_i(Y) - \theta) \quad . \tag{3.106}$$

For any estimator $\psi_i(Y)$ of θ, it is possible to define an estimator $\xi_i(Y)$ of α according to:

$$\xi_i(Y) = \Lambda^{-\frac{1}{2}} \psi_i(Y) \quad . \tag{3.107}$$

Substituting (3.107) into (3.106), it follows that:

$$\begin{aligned} L_2(\alpha,\xi_i(Y)) &= (\xi_i(Y) - \alpha)'\Lambda(\xi_i(Y) - \alpha) \\ &= (\psi_i(Y) - \theta)'(\psi_i(Y) - \theta) \\ &= L_1(\theta,\psi_i(Y)) \quad . \end{aligned} \tag{3.108}$$

It can be seen from (3.108) that the problem of estimating α from $(\hat{\alpha},V)$ is equivalent to estimating θ from $(\Lambda^{\frac{1}{2}}\hat{\alpha},V)$. Since V is distributed as a $\sigma^2\chi^2_{(n-p)}$ variate independent of $\hat{\alpha}^*$, James and Stein's (1961) results may be invoked to conclude that:

$$\hat{\alpha}^*_{JS} = \left(1 - \frac{1}{n-p+2}\frac{V}{\hat{\alpha}^{*\prime}\hat{\alpha}^*}\right)\hat{\alpha}^* \tag{3.109}$$

has a smaller risk than $\hat{\alpha}^*$ assuming an unweighted quadratic loss function. As a result of the equivalence between estimates of α and θ which is defined by (3.108), it follows that:

$$\hat{\alpha}_{JS} = \Lambda^{-\frac{1}{2}} \hat{\alpha}^*_{JS}$$

$$= \left(1 - \frac{p-2}{n-p+2} \frac{V}{\hat{\alpha}'\Lambda\hat{\alpha}}\right) \Lambda^{-1} \hat{\alpha} \qquad (3.110)$$

is an improvement over the ordinary least squares estimator assuming a design dependent loss function.

Mayer and Willke (1973) examined some of the problems associated with estimating appropriate shrinkage factors for $\tilde{\beta}_s$. They noted that since the absolute value of each component of $\tilde{\beta}_s$ is linearly increasing in c, $\tilde{\beta}_s$ will not stabilize as c increases. As a result, Mayer and Willke (1973) argued that a scheme similar to the ridge trace would not be appropriate for determining the shrinkage factor c. Instead, they proposed that the shrinkage factor be determined by fixing an upper bound on the increase in the residual sum of squares which results when $\tilde{\beta}_s$ is employed instead of $\hat{\beta}$ and solving for c. Suppose that the increase in the residual sum of squares is assumed to be 100M%. In this case, c would be assigned the value:

$$c = 1 - \left\{\frac{M(Y - X\hat{\beta})'(Y - X\hat{\beta})}{\hat{\beta}'X'X\hat{\beta}}\right\}^{\frac{1}{2}} \qquad (3.111)$$

In addition, Mayer and Willke (1973) attempted to develop an estimator which minimizes the sum of the variances of the individual components of the estimator among all estimators of β having a fixed residual sum of squares. They argued that such an estimator would have the advantage of being independent of the norm choosen to measure the length of the parameter estimates. Later, Mayer (*)

and Dwivedi, Srivastava and Richardson pointed out several errors in Mayer and Willke's (1973) solution.

Hocking, Speed and Lynn (1976) derived an explicit solution for the shrinkage factor by estimating the value of c which minimizes the mean squared error function for $\tilde{\beta}_s$. Consider the mean squared error function for $\tilde{\beta}_s$:

$$\begin{aligned} MSE(\tilde{\beta}_s) &= E((\tilde{\beta}_s - \beta)'(\tilde{\beta}_s - \beta)) \\ &= E((c\hat{\beta} - \beta)'(c\hat{\beta} - \beta)) \\ &= E((c\hat{\alpha} - \alpha)'(c\hat{\alpha} - \alpha)) \\ &= c^2E((\hat{\alpha} - \alpha)'(\hat{\alpha} - \alpha)) + (c\alpha - \alpha)'(c\alpha - \alpha) \\ &= c^2\sum_{i=1}^{p} Var(\hat{\alpha}_i) + (c - 1)^2\alpha'\alpha \end{aligned} \tag{3.112}$$

Substituting (3.67) into (3.112), the mean squared error function for $\tilde{\beta}_s$ becomes:

$$MSE(\tilde{\beta}_s) = c^2\sigma^2\sum_{i=1}^{p} \frac{1}{\lambda_i} + (c - 1)^2\alpha'\alpha \quad . \tag{3.113}$$

Minimizing (3.113) with respect to the shrinkage factor c gives:

$$c_1 = \frac{\alpha'\alpha}{\alpha'\alpha + \sigma^2\sum_{i=1}^{p} \frac{1}{\lambda_i}} \quad . \tag{3.114}$$

Let $\tilde{\beta}_{s1}$ denote the shrinkage estimator of β which is obtained using the shrinkage factor c_1.

Hocking, Speed and Lynn (1976) proposed that a value for c_1 be determined using an iterative procedure similar to the one

they gave for the continuous rank generalized least squares estimator. Let $c_{1(j)}$ denote the j'th iterate of the sequence of estimates of c_1. Assume that σ^2 is estimated by it's ordinary least squares estimator. In the calculation of $c_{1(j+1)}$, $\alpha'\alpha$ would be estimated by $c_{1(j)}^2\hat{\alpha}'\hat{\alpha}$. Hocking, Speed and Lynn's (1976) iterative formula for c_1 is given by:

$$c_{1(j+1)} = \frac{c_{i(j)}^2\hat{\alpha}'\hat{\alpha}}{c_{i(j)}^2\hat{\alpha}'\hat{\alpha} + \hat{\sigma}^2\sum_{i=1}^{p}\frac{1}{\lambda_i}}$$

$$= \frac{c_{1(j)}^2}{c_{1(j)}^2 + L_{s1}} \qquad (3.115)$$

where:

$$L_{s1} = \frac{\hat{\sigma}^2}{\alpha'\alpha}\sum_{i=1}^{p}\frac{1}{\lambda_i}$$

$$= \frac{\hat{\sigma}^2}{\beta'\beta}\sum_{i=1}^{p}\frac{1}{\lambda_i} \qquad (3.116)$$

Suppose that $c_{1(0)}$ is assigned the value $(1 + L_{s1})^{-1}$. Invoking the convergence theorem from the Appendix, the limiting values of $c_{i(j)}$ are found to be:

$$c_1^* = \begin{cases} 0 & \text{if } L_{s1} > \frac{1}{4} \\ \frac{1}{2} + (\frac{1}{4} - L_{s1})^{\frac{1}{2}} & \text{otherwise} \end{cases} \qquad (3.117)$$

Substituting (3.117) into (3.80), the limiting solution for $\tilde{\beta}_{s1}$ becomes:

$$\tilde{\beta}_{s1}^* = \begin{cases} 0 & \text{if } L_{s1} > \frac{1}{4} \\ (\frac{1}{2} + (\frac{1}{4} - L_{s1})^{\frac{1}{2}})\hat{\beta} & \text{otherwise} \end{cases} \qquad (3.118)$$

In addition to $\tilde{\beta}_{s1}^*$, Hocking, Speed and Lynn (1976) consider-

ed the shrinkage estimator which is obtained by minimizing the design dependent criterion:

$$E(L_D^2(b)) = E((b - \beta)'X'X(b - \beta)) \tag{3.119}$$

instead of the mean squared error function. Adopting procedures similar to those utilized above:

$$\begin{aligned} E(L_D^2(\tilde{\beta}_s)) &= E((\tilde{\beta}_s - \beta)'X'X(\tilde{\beta}_s - \beta)) \\ &= E((c\hat{\alpha} - \alpha)'\Lambda(c\hat{\alpha} - \alpha)) \\ &= \sum_{i=1}^{p} \lambda_i E(c\hat{\alpha}_i - \alpha_i)^2 \\ &= \sum_{i=1}^{p} \lambda_i \{c^2 Var(\hat{\alpha}_i) + (c - 1)^2 \alpha_i^2\} \\ &= pc^2\sigma^2 + (c - 1)^2 \alpha'\Lambda\alpha \end{aligned} \tag{3.120}$$

Minimizing (3.120) with respect to c leads to:

$$c_2 = \frac{\alpha'\Lambda\alpha}{p\sigma^2 + \alpha'\Lambda\alpha} \tag{3.121}$$

Let $\tilde{\beta}_{s2}$ denote the shrinkage estimator which is obtained by substituting c_2 into (3.80). Hocking, Speed and Lynn (1976) employed a sequence of estimators of c_2 analogous to the $c_{i(j)}$'s to obtain an explicit solution for a shrinkage estimator of β based upon $\tilde{\beta}_{s2}$. They defined the sequence of estimators:

$$\begin{aligned} c_{2(j+1)} &= \frac{c_{2(j)}^2 \hat{\alpha}'\Lambda\hat{\alpha}}{p\hat{\sigma}^2 + c_{2(j)}^2 \hat{\alpha}'\Lambda\hat{\alpha}} \\ &= \frac{c_{2(j)}^2}{c_{2(j)}^2 + L_{s2}} \end{aligned} \tag{3.122}$$

where:

$$L_{s2} = \frac{p\hat{\sigma}^2}{\hat{\alpha}'\Lambda\hat{\alpha}} \tag{3.123}$$

for c_2. Appealling to the convergence theorem from the Appendix, Hocking, Speed and Lynn (1976) obtained the shrinkage estimator:

$$\tilde{\beta}_{s2}^{*} = c_2^{*}\hat{\beta}$$

$$= \begin{cases} 0 & \text{if } L_{s2} > \frac{1}{2} \\ (\frac{1}{2} + (\frac{1}{4} - L_{s2})^{\frac{1}{2}})\hat{\beta} & \text{otherwise} \end{cases} \quad . \tag{3.124}$$

It should be noted that both $\tilde{\beta}_{s1}^{*}$ and $\tilde{\beta}_{s2}^{*}$ are 'stochastically' shrunken estimators because they depend upon L_{s1} and L_{s2} respectively which are in turn functions of $\hat{\beta}$.

Sclove (1968) proposed a modification to the shrinkage estimator $\tilde{\beta}_s$ by suggesting that the shrinkage factor only be applied to the components of the canonical form of the ordinary least squares estimator corresponding to the smallest eigenvalues of the $X'X$ matrix. Consider the canonical form of the shrinkage estimator:

$$\tilde{\alpha}_s = P'\tilde{\beta}_s \quad . \tag{3.125}$$

Suppose that the shrinkage factor is only applied to the $(p - r)$ components of $\hat{\alpha}$ which correspond to the smallest λ_i's. In this case, Sclove's (1968) shrinkage estimator becomes:

$$\hat{\alpha}_s = \begin{bmatrix} c_{ss}I(p - r) & 0 \\ 0 & I_r \end{bmatrix} \hat{\alpha} \tag{3.126}$$

where:

$$0 \leq c_{ss} \leq 1 \tag{3.127}$$

If c_{ss} is assigned the value zero, Sclove's (1968) estimator reduces to the generalized least squares estimator of rank r. Optimum values

for c_{ss} may be derived using techniques similar to those employed to determine c_1^* and c_2^*. Suppose that:

$$L_{ss1} = \hat{\sigma}^2 \{\sum_{i=1}^{p-r} \frac{1}{\lambda_i}\} \{\sum_{i=1}^{p-r} \alpha_i^2\}^{-1} \tag{3.128}$$

and:

$$L_{ss2} = (p - r)\hat{\sigma}^2 \{\sum_{i=1}^{p-r} \lambda_i \alpha_i^2\}^{-1} . \tag{3.129}$$

In addition, let c_{ss1} and c_{ss2} represent the respective shrinkage factors for $\tilde{\alpha}_{ss}$ which are obtained when the mean squared error function and design dependent norm criterion are minimized. Iterative sequences similar to (3.115) and (3.122) may be constructed for c_{ss1} and c_{ss2}. It is easily shown that the limiting values of these sequences are of the form:

$$c_{ssi} = \begin{cases} 0 & \text{if } L_{ssi} > \frac{1}{4} \\ (\frac{1}{2} + (\frac{1}{4} - L_{ssi})^{\frac{1}{2}}) & \text{otherwise} \end{cases} . \tag{3.130}$$

Therefore, the limiting solutions for Sclove's (1968) estimator of β using c_{ss1} and c_{ss2} are given by:

$$\tilde{\beta}_{ssi}^* = \begin{cases} P\hat{\theta}_r & \text{if } L_{ssi} > \frac{1}{4} \\ (\frac{1}{2} + (\frac{1}{4} - L_{ssl})^{\frac{1}{2}})P(\hat{\theta}_{r'} - \hat{\alpha}) + P\hat{\theta}_r & \text{otherwise} \end{cases} \tag{3.131}$$

where $\hat{\theta}_k$ is a vector whose first $(p - k)$ components are zero and last k components agree with $\hat{\alpha}$.

It was noted earlier that any matrix A^+ which satisfies (3.1) is defined as a generalized inverse for the matrix A. Suppose that in addition to condition (3.1), the matrix A^+ satisfies:

$$A^+ A A^+ = A^+ \tag{3.132}$$

$$(A A^{+})' = A A^{+} \tag{3.133}$$

and:

$$(A^{+} A)' = A^{+} A \tag{3.134}$$

In this case, A^{+} is said to be the Moore-Penrose inverse for the matrix A and is denoted by $A^{(p)}$. In contrast to a generalized inverse, a Moore-Penrose inverse is unique if it exists[1]. Lowerre (1974) utilized the concept of a Moore-Penrose inverse to present a more general family of estimators defined by:

$$\hat{b}^{*} = (X'X + C)^{(p)} X'Y \tag{3.135}$$

where C is a symmetric matrix which commutes with the $X'X$ matrix. He showed that the class of estimators defined by (3.135) includes the ridge, shrinkage and generalized least squares estimators by assigning C the values kI_p, $(c^{-1} - 1)X'X$ and $-P_r \Lambda_r P_r'$ respectively. Lowerre (1974) employed $\hat{b}^{*}$ to derive conditions upon C which ensure that $\hat{b}^{*}$ is component-wise mean squared error admissible. In addition, he constructed an example to demonstrate that $\hat{b}^{*}$ might be quite different from β even when all the eigenvalues of C are positive and close to zero.

Goldstein and Smith (1974) argued that James and Stein's (1961) shrinkage estimator $\hat{\alpha}_{JS}$ would be inappropriate for use in situations where severe multicollinearity between the explanatory variables exists. They noted that James and Stein's (1961) shrinkage esti-

(1.) Searle, S.R., Linear Models, New York: John Wiley & Sons, Inc., 1971, p. 18.

mator was derived using the design dependent loss function. This loss function weighs the deviations of the individual components of the estimator from the components of the parameter vector by the corresponding eigenvalues from the $X'X$ matrix. As a result, the design dependent loss function implicitly takes less account of the loss corresponding to those directions where estimation is most inaccurate. In addition, Goldstein and Smith (1974) pointed out that nothing can be said about the accuracy of the individual components of $\hat{\alpha}_{JS}$. Goldstein and Smith (1974) provided an alternative formulation of the shrinkage estimator which they suggested addresses these problems.

Goldstein and Smith (1974) noted the existence of an orthogonal n x n matrix Q such that:

$$QXP' = \begin{bmatrix} \Lambda^{\frac{1}{2}} \\ \hline 0 \end{bmatrix} \tag{3.136}$$

where $\Lambda^{\frac{1}{2}}$ is a p x p diagonal matrix and 0 is a (n - p) x p matrix of zeros. It is assumed in (3.136) that the i'th diagonal element of $\Lambda^{\frac{1}{2}}$ equals the square root of λ_i. Consider the following transformation of the canonical form of the general model:

$$\begin{aligned} Z &= QY \\ &= QX^{*}\alpha + Q\varepsilon \end{aligned} \tag{3.137}$$

In this case, Z has an expected value given by:

$$E(Z) = \begin{cases} \lambda_i^{\frac{1}{2}}\alpha_i & \text{for } i = 1,2, \ldots p \\ 0 & \text{for } i = p+1, p+2, \ldots n \end{cases} \tag{3.138}$$

and a variance-covariance matrix equal to:

$$Var(Z) = Var(Q\varepsilon)$$

$$= Q \, Var(\varepsilon) \, Q'$$

$$= \sigma^2 I_n \qquad (3.139)$$

Adopting Goldstein and Smith's (1974) formulation of the general model, the ordinary least squares estimator of α is given by:

$$\hat{\alpha} = (X^{*\prime} Q' Q X^*)^{-1} X^{*\prime} Q' Z$$

$$= \Lambda^{-1} P' X' Q' Z$$

$$= \lambda_i^{-\frac{1}{2}} z_i \quad \text{for } i = 1,2, \ldots p \qquad (3.140)$$

It was noted earlier that ill-conditioning in the design matrix has the effect of inflating the Euclidean length of the ordinary least squares estimates. In order to compensate for this tendency, Goldstein and Smith (1974) introduced a class of linear shrinkage estimators of the form:

$$\hat{\alpha}_{GSi} = c_i \, z_i$$

$$= c(\lambda_i^{\frac{1}{2}}, k) \, z_i \qquad (3.141)$$

where:

$$|c_i \, z_i| < |\lambda_i^{-\frac{1}{2}} \, z_i|$$

$$= |\hat{\alpha}_i| \qquad (3.142)$$

for $k > 0$. Goldstein and Smith (1974) introduced the adjustable factor k so that (3.141) defines a class of estimators for α. In order to ensure that $\hat{\alpha}_{GSi}$ is a shrinkage estimator, they imposed the following conditions upon $c(\lambda_i^{\frac{1}{2}}, k)$:

i) $c(\lambda_i^{\frac{1}{2}}, 0) z_i$ corresponds to the ordinary least squares estimator $\hat{\alpha}_i$.

ii) For a fixed $\lambda_i^{\frac{1}{2}}$, $|c(\lambda_i^{\frac{1}{2}}, k)|$ is a continuous, monotonic,

decreasing function of k as k increases from 0 to $+\infty$.

iii) $c(\lambda_i^{\frac{1}{2}},k)$ has the same sign as $\lambda_i^{\frac{1}{2}}$ for all values of k.

These three conditions may be summarized by assuming that $c(\lambda_i^{\frac{1}{2}},k)$ satisfies:

$$c(\lambda_i^{\frac{1}{2}},0) = \frac{1}{\lambda_i^{\frac{1}{2}}} \tag{3.143}$$

and:

$$\frac{1}{\lambda_i^{\frac{1}{2}}} \frac{\delta}{\delta k} c(\lambda_i^{\frac{1}{2}},k) < 0 \tag{3.144}$$

for all non-negative values of k.

Goldstein and Smith (1974) provided the following theorem regarding the mean squared error functions of the individual components of $\hat{\alpha}_{GS}$:

<u>Theorem 3.7</u>: For each component i, there exists a k>0 such that $\hat{\alpha}_{GSi}$ has a smaller mean squared error than $\hat{\alpha}_i$.

<u>Proof</u>: By definition, the mean squared error function for $\hat{\alpha}_{GSi}$ is given by:

$$\begin{aligned} MSE(\hat{\alpha}_{GSi}) &= E(c_i z_i - \alpha_i)^2 \\ &= E(c_i z_i - c_i\lambda_i^{\frac{1}{2}}\alpha_i)^2 + (c_i\lambda_i^{\frac{1}{2}}\alpha_i - \alpha_i)^2 \\ &= c_i^2 Var(z_i) + \alpha_i^2(c_i\lambda_i^{\frac{1}{2}} - 1)^2 \\ &= c_i^2\sigma^2 + \alpha_i^2(c_i\lambda_i^{\frac{1}{2}} - 1)^2 \quad . \end{aligned} \tag{3.145}$$

The variance for $\hat{\alpha}_i$ is given as (3.67). Therefore, $\hat{\alpha}_{GSi}$ is mean squared error admissible if:

$$c_i^2\sigma^2 + \alpha_i^2(c_i\lambda_i^{\frac{1}{2}} - 1)^2 < \frac{\sigma^2}{\lambda_i} \tag{3.146}$$

or:

$$\alpha_i^2 \leq \frac{\sigma^2}{\lambda_i} \frac{(\lambda_i^{-1} - c_i^2)}{(c_i - \lambda_i^{-\frac{1}{2}})^2}$$

$$= \frac{\sigma^2}{\lambda_i} \frac{(\lambda_i^{-\frac{1}{2}} + c_i)}{(\lambda_i^{-\frac{1}{2}} - c_i)}$$

$$= \frac{\sigma^2}{\lambda_i} \frac{c(\lambda_i^{\frac{1}{2}},0) + c(\lambda_i^{\frac{1}{2}},k)}{c(\lambda_i^{\frac{1}{2}},0) - c(\lambda_i^{\frac{1}{2}},k)} \tag{3.147}$$

Invoking (3.143) and (3.144), it can be seen that the right hand side of (3.147) is positive. Therefore, there exists a positive k such that the i'th component of $\hat{\alpha}_{GS}$ has a smaller mean squared error than $\hat{\alpha}_i$.

It was noted above that James and Stein's (1961) estimator was constructed using a weighted mean squared error function. As a result, there is no guarantee as to the performance of the individual components of $\hat{\alpha}_{JS}$. In contrast, Theorem 3.7 provides for the mean squared error admissibility of each component of $\hat{\alpha}_{GS}$. Consider the estimator of α derived using Goldstein and Smith's (1974) estimator of α:

$$\hat{\beta}_{GS} = P\,\hat{\alpha}_{GS} \tag{3.148}$$

where:

$$\hat{\alpha}_{GSi} = c(\lambda_i^{\frac{1}{2}},k)\, z_i \tag{3.149}$$

The component-wise mean squared error admissibility of $\hat{\beta}_{GS}$ is provided by the theorem:

<u>Theorem 3.8</u>: For any β, there exists a $k>0$ such that each component of $\hat{\beta}_{GS}$ has a smaller mean squared error

than the corresponding component of the ordinary least squares estimator $\hat{\beta}$.

The proof of this theorem is an extension of the proof of Theorem 3.7.

Goldstein and Smith (1974) examined a number of possible choices for $c(\lambda_i^{\frac{1}{2}},k)$. For example, they considered the function:

$$c(\lambda_i^{\frac{1}{2}},k) = \lambda_i^{-\frac{1}{2}} - k \quad . \tag{3.150}$$

(3.150) satisfies conditions (3.143) and (3.144). However, this function has an effect opposite to that desired. (3.150) has the most effect when λ_i is large and the least effect when λ_i is small. Goldstein and Smith (1974) argued that the simplest suitable $c(\lambda_i^{\frac{1}{2}},k)$ is of the form:

$$c(\lambda_i^{\frac{1}{2}},k) = \frac{\lambda_i^{\frac{1}{2}}}{\lambda_i + k} \quad . \tag{3.151}$$

In this case,

$$\hat{\alpha}_{GSi} = \frac{\lambda_i^{\frac{1}{2}}}{\lambda_i + k} z_i \tag{3.152}$$

and:

$$\hat{\alpha}_{GS} = (\Lambda + kI_p)^{-1}P'X'Q'Z \quad . \tag{3.153}$$

The corresponding estimator of β is given by:

$$\begin{aligned}\hat{\beta}_{GS} &= P(\Lambda + kI_p)^{-1}P'X'Q'Z \\ &= (P\Lambda P' + kI_p)^{-1}X'Y \\ &= (X'X + kI_p)^{-1}X'Y \\ &= \hat{\beta}_k^* \quad .\end{aligned} \tag{3.154}$$

It can be seen from (3.154) that the ridge estimator may be

viewed as a subclass of the shrinkage estimators defined by (3.148) and (3.149). Goldstein and Smith (1974) utilized their formulation of the shrinkage estimator to provide a greater insight into ridge regression. Suppose that P_{ij} denotes the (i,j)'th element of the orthogonal matrix P. Goldstein and Smith (1974) showed that the largest potential decrease in the mean squared error function occurs when $|P_{ij}|$ is large and λ_i is small. This is the situation when β_i has large components in the canonical space in directions where the estimation is the most inaccurate. Goldstein and Smith (1974) argued that this is a justification for the claim that coefficient estimates with incorrect signs when k equals zero tend to change to the correct sign as k increases.

Several numerical examples and simulation experiments have been provided in the statistical literature to compare the use of least squares, shrinkage and other biased estimators. Mayer and Willke (1973) utilized a multifactor problem to compare their 'deterministically' and 'stochastically' shrunken estimators with the ridge estimator. Hocking, Speed and Lynn (1976) employed the pitprop data described by Jeffers (1967) and the air pollution data provided by McDonald and Schwing (1973) to compare their iterative shrinkage estimators with the ridge regression estimator. They found that the solutions produced by their iterative shrinkage estimators had significantly smaller R^2 values than the corresponding ridge regression solutions. Gunst and Mason (1974) provided

a detailed series of simulations to evaluate the performance of the ordinary least squares, generalized least squares, ridge and James and Stein (1961) shrinkage estimators. Based upon the results of the simulations, they concluded that the James and Stein's (1961) estimator is substantially better than the least squares estimator only when the $X'X$ matrix is nearly orthogonal. In addition, Gunst and Mason (1974) found that the ridge estimator almost always yielded a smaller mean squared error than James and Stein's (1961) estimator. Another series of simulation experiments comparing the efficiencies of different ridge and shrinkage estimators was described by Dempster, Schatzoff and Wermuth (1977). In addition to the simulation results, the comments of a number of reviewers of Dempster, Schalzoff and Wermuth's (1977) work were published.

Chapter 4

Criticisms Of The Ridge Estimator

Coniffe and Stone (1973) provided an extensive critique of the ridge regression procedures outlined by Hoerl and Kennard (1970a). A number of mathematical and statistical objections to the procedures which Coniffe and Stone (1973) felt limited the usefulness of the estimator were presented. Their most serious objection concerned the mean squared error admissibility of the ridge estimates calculated from sampled data. Hoerl and Kennard (1970a) demonstrated the existence of an interval for k dependent upon α and σ^2 such that the ridge estimator is mean squared error admissible. In addition, Hoerl and Kennard (1970a) suggested several procedures for choosing appropriate values of k. Coniffe and Stone (1973) pointed out that Hoerl and Kennard's (1970a) arguments assumed that both α and σ^2 are known quantities. In reality, these parameters must be estimated from the sampled data. Therefore, Coniffe and Stone (1973) argued that there is no guarantee that any of the ridge regression procedures proposed by Hoerl and Kennard (1970a) will produce values of k which lead to improved parameter estimates in any practical problem.

In order to illustrate their criticism of Hoerl and Kennard's (1970a) mean squared error admissibility condition, Coniffe and Stone (1973) considered the estimation of the mean for a population of normal variates. Suppose that the population mean and variance are denoted by μ and σ^2 respectively. A sample of size n from the population is assumed. First, Coniffe and Stone (1973) considered

the sample mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{4.1}$$

as an estimator of μ. Since the sample mean is an unbiased estimator of μ, the mean squared error of $\bar{x}$ is given by:

$$\begin{aligned} MSE(\bar{x}) &= E(\bar{x} - \mu)^2 \\ &= \sigma^2/n \end{aligned} \tag{4.2}$$

Coniffe and Stone (1973) also considered $k\bar{x}$ where:

$$k = (1 + \sigma^2/(n\mu^2))^{-1} \tag{4.3}$$

as an estimator of μ. They noted that the mean squared error of $k\bar{x}$ is given by:

$$\begin{aligned} MSE(k\bar{x}) &= E(k\bar{x} - \mu)^2 \\ &= k^2 Var(\bar{x}) + (k - 1)^2\mu^2 \\ &= \mu^2\sigma^2/(n\mu^2 + \sigma^2) \end{aligned} \tag{4.4}$$

provided that the true value of k is known. Utilizing (4.2) and (4.4), $k\bar{x}$ is mean squared error admissible if and only if:

$$\mu^2\sigma^2/(n\mu^2 + \sigma^2) < \sigma^2/n \tag{4.5}$$

Condition (4.5) is satisfied for all values of μ and n if $\sigma^2>0$. Therefore, Coniffe and Stone (1973) concluded that $k\bar{x}$ is a mean squared error admissible estimator of μ whenever $\sigma^2>0$ and k is known.

In light of the above, Coniffe and Stone (1973) suggested that it might seem reasonable to consider an estimator for μ of the form:

$$\hat{\mu} = k\bar{x}$$

$$= \bar{x}(1 + s^2/(nx^2))^{-1} \tag{4.6}$$

whenever the true value of k is unknown. However, Hodges and Lehmann (1951) showed that $\bar{x}$ has the smallest mean squared error of all estimators of μ whenever both σ^2 and σ^2/μ^2 are unknown. As a result, no estimator of the form (4.6) can be mean squared error admissible if k is unknown. In the same manner, Coniffe and Stone (1973) argued that Hoerl and Kennard's (1970a) admissibility condition is only valid if α and σ^2 are known quantities.

Most of the other criticisms raised by Coniffe and Stone (1973) concerned the stability of the ridge estimator and the criteria proposed by Hoerl and Kennard (1970a) for choosing appropriate values of k. Coniffe and Stone (1973) noted that ridge regression consists of inflating the diagonal elements of the $X'X$ matrix so that $\hat{\beta}_k^{*\prime}\hat{\beta}_k^*$ is less than $\hat{\beta}'\hat{\beta}$. Hoerl and Kennard (1970a) argued that since the coefficients of the least squares solution are often inflated; the ridge estimator should produce better estimates of β than the ordinary least squares estimator. Coniffe and Stone (1973) rejected this arguement pointing out that unless one knows which components of β are overestimated and which are underestimated; there is no reason to believe that any components of the ridge estimates should be closer to the true parameters than the corresponding components of the ordinary least squares estimates. In addition, they argued that it is not self-evident that all the criteria proposed by Hoerl and Kennard (1970a) for choosing

a value of k from a ridge trace can be simultaneously satisfied in any problem.

Hoerl and Kennard (1970a) emphasized the property of the ridge estimator that the ridge estimates tend to stabilize for some positive values of k. Coniffe and Stone (1973) argued that this tendency is a direct result of the form of the ridge estimator and not a property indicating that the ridge estimator is superior to the ordinary least squares estimator. To make their point, they considered the case where the $X'X$ matrix is an identity matrix and so perfectly conditioned. Invoking (2.2), the ridge estimator for this $X'X$ matrix may be written as:

$$\hat{\beta}_k^* = (I_p + k(X'X)^{-1})^{-1} \hat{\beta}$$
$$= (1 + k)^{-1} \hat{\beta} \quad . \tag{4.7}$$

In this case:

$$\frac{d}{dk} \hat{\beta}_k^* = - (1 + k)^{-2} \hat{\beta} \quad . \tag{4.8}$$

It can be seen from (4.8) that the ridge estimator will change more slowly for increasing k even in the case when the $X'X$ matrix is perfectly conditioned. Therefore, Coniffe and Stone (1973) concluded that the tendency of the ridge estimator to stabilize is the result of the definition of the estimator and not the ill-conditioning in the data.

Finally, Coniffe and Stone (1973) pointed out that an ill-conditioned $X'X$ matrix indicates that the data set is inadequate

or that some of the dependent variables are redundant. Instead of using a biased estimator such as the ridge estimator, they suggested that it would be more proper to collect more data or drop any redundant variables. Coniffe and Stone (1973) pointed to the extreme case where one of the eigenvalues for the $X'X$ matrix is actually zero to justify this assertion. It would not be proper to utilize the ridge estimator in this situation since the ridge estimator would re-introduce the effect of an eigenvalue known to be zero into the estimation process. However, Coniffe and Stone (1973) pointed out that the smallest eigenvalue may not be exactly equal to zero due to errors in the measurement of the independent variables. As a result, all the eigenvalues would be incorrectly assumed to be non-zero and ridge regression applied.

Smith and Goldstein (1975) responded to Coniffe and Stone's (1973) criticisms of the ridge estimator. They agreed that mean squared error function arguments for $\hat{\beta}_k^*$ can often be misleading. Smith and Goldstein (1975) emphasized that one should be aware of when to apply the ridge estimator to shrink the estimates of the coefficients and when not to. However, they pointed out that a value for k which improves each component of the estimates can always be found. The mean squared error admissibility condition presented by Hoerl and Kennard (1970a) gives a sufficient condition for such a value of k.

The condition under which the greatest improvement in the mean squared errors of the ridge estimates can be expected was presented by Smith and Goldstein (1975). The mean squared error for the ridge estimator of β_i can be expressed as:

$$\begin{aligned} MSE(\hat{\beta}^*_{ik}) &= E(\hat{\beta}^*_{ik} - \hat{\beta}_i)^2 \\ &= E(\sum_{j=1}^{p} P_{ij}(\hat{\alpha}^*_{jk} - \alpha_j))^2 \quad . \end{aligned} \tag{4.9}$$

Smith and Goldstein (1975) employed the form of the canonical model defined by (3.137) to simplify (4.9). Invoking (3.138), (3.139) and (3.152), the mean squared error of the ridge estimate of β_i may be rewritten as:

$$\begin{aligned} MSE(\hat{\beta}^*_{ik}) = &\sum_{j=1}^{p} P_{ij} \left\{\frac{\lambda_j\sigma^2 - k^2\alpha_i^2}{(\lambda_j + k)^2}\right\} \\ &+ 2\sum_{j=1}^{p-1}\sum_{m=j+1}^{p} \frac{k^2 P_{ij}P_{im}\alpha_j\alpha_m}{(\lambda_j + k)(\lambda_m + k)} \quad . \end{aligned} \tag{4.10}$$

Therefore, it follows that:

$$\frac{d}{dk} MSE(\hat{\beta}^*_{ik})\Big|_{k=0} = -2\sigma^2 \sum_{j=1}^{p} \lambda_j^{-2} P_{ij}^2 \quad . \tag{4.11}$$

As a result of (4.11), Smith and Goldstein (1975) concluded that the greatest potential for improvement in the estimated coefficients using the ridge estimator occurs when P_{ij} is large and λ_j small. Further, they argued that k will have little effect upon the estimated coefficients when λ_j is at all large. They argued that the ridge estimates of the coefficients should be close to the ordinary least squares estimates in this case.

Smith and Goldstein (1975) argued that k should be considered as a function of the $X'X$ matrix instead of Y. They suggested that the mean squared error function for the ridge estimator may be viewed as an expectation conditional upon the $X'X$ matrix. In this case, k would be a constant with respect to the expectation operator. Smith and Goldstein (1975) rejected Coniffe and Stone's (1973) assertion that Hoerl and Kennard's (1970a) proof of the existence of mean squared error admissible ridge estimators requires α and σ^2 to be known. In fact, Goldstein and Smith (1975) recalled one of Hoerl and Kennard's (1970a) original criteria for choosing a value for k using the ridge trace to suggest an algorithm for selecting k. Based upon Hoerl and Kennard's (1970a) argument that the ridge trace should resemble an orthogonal system near the value of k choosen, Smith and Goldstein (1975) proposed that k be selected to satisfy:

$$Z_k(X'X)^{-1}Z_k = (1 + k)^{-2}I_p \quad . \tag{4.11}$$

Smith and Goldstein (1975) suggested that the analysis of a ridge trace may be viewed as a numerical inspection procedure for solving (4.11). In practice, such a procedure may or may not give reasonable results.

Smith and Goldstein (1975) accused Coniffe and Stone (1973) of being naive in recommending that more data be collected or dependent variables dropped when the data set is inadequate rather than utilizing a biased estimator such as the ridge estimator. Smith

and Goldstein (1975) noted that collecting more data should almost always be preferable. However, they suggested that data sets are often encountered for which the collection of more data is not practical. In the same manner, Smith and Goldstein (1975) noted that problems arise for which it is necessary to estimate the parameters for the full model. Smith and Goldstein (1975) asserted that ridge regression is preferable to ordinary least squares estimation in precisely these situations.

Later, Coniffe and Stone (1975) replied to Smith and Goldstein's (1975) discussion of their critique. In addition to restating their original criticisms of ridge regression, Coniffe and Stone (1975) noted that Smith and Goldstein (1975) did not reject their arguement that Hoerl and Kennard (1970a) have not shown that any specific algorithm for choosing k necessarily leads to a mean squared error admissible estimator. Furthermore, Coniffe and Stone (1975) argued that k is a function of Y irregardless of whether it is estimated by $\hat{\sigma}^2/\hat{\alpha}_i^2$ or selected using a ridge trace. They pointed out that condition (4.11) cannot be satisfied by any finite value of k and suggested that Smith and Goldstein (1975) had misunderstood the stability criteria proposed by Hoerl and Kennard (1970a). Coniffe and Stone (1975) observed that Mayer and Willke (1973) raised objections to the ridge estimator similar to theirs.

Bacon and Hausman (1974) outlined several difficulties inherent

in ridge regression in addition to those mentioned above. They noted that the value of σ^2 required to estimate the variances of the ridge estimates must be estimated using the ordinary least squares estimator. Bacon and Hausman (1974) pointed out that no direct meaning can be assigned to k and that the algorithms for choosing k are imprecise. In addition, they noted that ridge regression is only suitable for models whose error structure can efficiently be handled by the ordinary least squares estimator.

In order to circumvent some of the difficulties mentioned above, Bacon and Hausman (1974) employed Chapman's (1964) minimum mean squared error estimator to treat the biasing parameter k as a variable. k was assigned a prior mean and variance. Bacon and Hausman (1974) noted that more general error structures may be handled using Chipman's (1964) estimator if a covariance matrix is assumed for the unobservable disturbances. Bacon and Hausman (1974) also pointed out that the choice of the k for the ridge estimator in their formulation does not depend upon the estimation results but rather upon prior assumptions which are known before the data is collected.

Vinod (1976b) described several problems associated with the ridge regression procedures outlined by Hoerl and Kennard (1970a). He noted that it was difficult to utilize the constant k to monitor the degree of multicollinearity in any particular problem since it has an infinite range. In addition, Vinod (1976b) mentioned the

property first pointed out by Coniffe and Stone (1973) that the stability of the ridge trace for increasing k results directly from the form of the estimator. In order to alleviate these two problems, Vinod (1976b) introduced a new scale for the horizontal axis of the ridge trace which he called the multicollinearity allowance. He defined the multicollinearity allowance by:

$$m = p - \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + k} \quad . \tag{4.12}$$

The multicollinearity allowance takes on the extreme values 0 when k equals 0 and p when k is set to $+\infty$. Vinod (1976b) interpreted the multicollinearity allowance as the assigned deficiency in the $X'X$ matrix.

Unlike the k scale, the multicollinearity allowance scale does not have the property that the ridge trace appears more stable for larger m even if the data is completely orthogonal. This can be illustrated by assuming that the $X'X$ matrix is a p x p identity matrix. In this case:

$$\begin{aligned} \hat{\beta}_k^* &= (I_p + k(X'X)^{-1})^{-1}\,\hat{\beta} \\ &= (1 + k)^{-1}\,\hat{\beta} \quad , \end{aligned} \tag{4.13}$$

so that:

$$\frac{d}{dk}\,\hat{\beta}_k^* = -\,(1 + k)^{-2}\,\hat{\beta} \tag{4.14}$$

and:

$$\begin{aligned} \frac{d}{dk}\,m &= \frac{d}{dk}\left(p - \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + k}\right) \\ &= p/(1 + k)^2 \quad . \end{aligned} \tag{4.15}$$

Therefore, the rate of change in the ridge estimator with respect to m is given by:

$$\frac{d}{dm}\hat{\beta}_k^* = \frac{d}{dk}\hat{\beta}_k^* \cdot \frac{dk}{dm}$$

$$= -\hat{\beta}/p \qquad (4.16)$$

which is independent of m. It follows from (4.16) that the ridge trace will not necessarily stabilize as m increases.

Vinod (1976b) argued that the use of the m scale should lead to ridge traces which are more easily interpreted than those produced using the k scale. He illustrated this conjecture using two economy of scale functions which exhibited serious multicollinearity in their exogenous variables.

Obenchain (1975b) reviewed a number of the properties of the residuals which result from the use of the ordinary least squares, weighted least squares and biased estimators. He examined various residual optimality properties. Obenchain (1975b) noted that, by definition, the ordinary least squares residual vector is the shortest. Therefore, he argued that it is intuitively obvious that no improvement in the mean squared residual error can result from the use of a ridge estimator.

Chapter 5

Generalizations Of The Ridge Estimator

Hoerl and Kennard (1970a) derived a sufficient condition for the mean squared error admissibility of the ridge estimator $\hat{\beta}_k^*$. They showed that for any constant k which satisfies:

$$0 < k < \frac{\sigma^2}{\alpha_i^2} \tag{5.1}$$

for each component i; the mean squared error of $\hat{\beta}_k^*$ is less than the sum of the variances of the components of $\hat{\beta}$. Hoerl and Kennard (1970a) proposed a generalization of the ridge estimator $\hat{\beta}_k^*$ based upon inequality (5.1). Instead of choosing a single biasing parameter k, they suggested the use of a separate k_i corresponding to each component of β. As a result, the generalized ridge estimator becomes:

$$\hat{\beta}^*(K) = (X'X + K)^{-1}X'Y \tag{5.2}$$

where K is a diagonal matrix with non-negative diagonal elements. In some situations, it will be more convenient to consider the canonical form of the general model:

$$\begin{aligned} Y &= X\beta + \varepsilon \\ &= X^*\alpha + \varepsilon \end{aligned} \tag{5.3}$$

where X^* and α are defined in (2.20). Adopting the canonical form of the model, the generalized ridge estimator becomes:

$$\begin{aligned} \hat{\alpha}^*(K) &= (X^{*\prime}X^* + K)^{-1}X^{*\prime}Y \\ &= (\Lambda + K)^{-1}X^{*\prime}Y \end{aligned} \tag{5.4}$$

Goldstein and Smith (1974) noted the existence of an orthogonal matrix Q such that:

$$QXP = \begin{bmatrix} \Lambda^{\frac{1}{2}} \\ \hline 0 \end{bmatrix} \tag{5.5}$$

where $\Lambda^{\frac{1}{2}}$ is a p x p diagonal matrix and 0 is a (n - p) x p matrix of zeros. The i'th diagonal element of $\Lambda^{\frac{1}{2}}$ equals the square root of λ_i. Using the orthogonal matrix Q, they derived a more explicit form of the estimator $\hat{\alpha}^*(K)$. Consider the following transformation of the canonical form of the general model:

$$\begin{aligned} Z &= QY \\ &= QX^*\alpha + Q\varepsilon \quad . \end{aligned} \tag{5.6}$$

Adopting this form of the general model, Hoerl and Kennard's (1970a) generalized ridge regression estimator becomes:

$$\begin{aligned} \hat{\alpha}^*(K) &= (X^{*\prime}Q'QX^* + K)^{-1}X^{*\prime}Q'Z \\ &= (\Lambda + K)^{-1}X^{*\prime}Q'Z \end{aligned} \tag{5.7}$$

where:

$$\begin{aligned} X^{*\prime}Q' &= P'X'Q' \\ &= \begin{bmatrix} \Lambda^{\frac{1}{2}} \\ \hline 0 \end{bmatrix}' \quad . \end{aligned} \tag{5.8}$$

Combining (5.7) and (5.8), it can be seen that each component of $\hat{\alpha}^*(K)$ is of the form:

$$\tilde{\alpha}_i^*(K) = \frac{1}{\lambda_i + k_i} \lambda_i^{\frac{1}{2}} z_i \quad . \tag{5.9}$$

Properties analogous to those for the ridge estimator may be derived for the generalized ridge estimator. Since:

$$\begin{aligned} \hat{\beta}^*(K) &= (X'X + K)^{-1}X'Y \\ &= (X'X + K)^{-1}(X'X)\hat{\beta} \\ &= (I_p + K(X'X)^{-1})^{-1}\hat{\beta} \quad , \end{aligned} \tag{5.10}$$

it can be seen that $\hat{\beta}^*(K)$ is a linear transformation of the ordinary

least squares estimator $\hat{\beta}$. The expected value of $\hat{\beta}^*(K)$ is given by:

$$E(\hat{\beta}^*(K)) = (I_p + K(X'X)^{-1})^{-1}E(\hat{\beta})$$
$$= (I_p + K(X'X)^{-1})^{-1}\beta \quad . \tag{5.11}$$

Therefore $\hat{\beta}^*(K)$ is a biased estimator of β if any diagonal element of K is nonzero. The variance-covariance matrix for $\hat{\beta}^*(K)$ follows directly from the variance-covariance matrix for the least squares estimator:

$$\mathrm{Var}(\hat{\beta}^*(K)) = \mathrm{Var}((I_p + K(X'X)^{-1})^{-1}\hat{\beta})$$
$$= \sigma^2(I_p + K(X'X)^{-1})^{-1}(X'X)^{-1}(I_p + K(X'X)^{-1})^{-1}$$
$$= \sigma^2(X'X + K)^{-2}(X'X) \quad . \tag{5.12}$$

A sufficient condition for the mean squared error admissibility of the generalized ridge estimator is provided by the theorem:

<u>Theorem 5.1</u>: Suppose that for each component i:

$$0 < k_i < \frac{\sigma^2}{\alpha_i^2} \quad , \tag{5.13}$$

then $\hat{\beta}^*(K)$ is a mean squared error admissible estimator of β in (1.1).

<u>Proof</u>: Since orthogonal transformations preserve lengths and:

$$\hat{\beta}^*(K) = P\hat{\alpha}^*(K) \quad , \tag{5.14}$$

it is sufficient to show that $\hat{\alpha}^*(K)$ is a mean squared error admissible estimator of α when (5.13) is satisfied for each component i. Consider Goldstein and Smith's (1974) characterization of the generalized ridge estimator. Since the unobservable disturbances are assumed

to satisfy conditions (1.2), the expected value and variance-covariance matrix for Z are:

$$
\begin{aligned}
E(Z) &= E(QX^{*}\alpha + Q\varepsilon) \\
&= QX^{*}\alpha + QE(\varepsilon) \\
&= \begin{bmatrix} \Lambda^{\frac{1}{2}} \\ \hline 0 \end{bmatrix} \alpha \\
&= \begin{cases} \lambda^{\frac{1}{2}}\alpha_i & \text{for } i = 1,2, \ldots p \\ 0 & \text{for } i = p+1, p+2, \ldots n \end{cases}
\end{aligned}
\tag{5.15}
$$

and:

$$
\begin{aligned}
Var(Z) &= Var(Q\varepsilon) \\
&= Q\ Var(\varepsilon)\ Q' \\
&= \sigma^2 I_p
\end{aligned}
\tag{5.16}
$$

respectively. By definition, the i'th component of the mean squared error function for $\hat{\alpha}^{*}(K)$ is:

$$
\begin{aligned}
E(\hat{\alpha}_i^{*}(K) - \alpha_i)^2 &= \frac{1}{(\lambda_i + k_i)^2} E(\lambda_i^{\frac{1}{2}} z_i - \lambda_i\alpha_i - k_i\alpha_i)^2 \\
&= \frac{1}{(\lambda_i + k_i)^2} (\lambda_i Var(z_i) + k_i^2\alpha_i^2) \\
&= \frac{1}{(\lambda_i + k_i)^2} (\lambda_i\sigma^2 + k_i^2\alpha_i^2) \quad .
\end{aligned}
\tag{5.17}
$$

Therefore, the mean squared error function for $\hat{\alpha}^{*}(K)$ becomes:

$$
\begin{aligned}
MSE(\hat{\alpha}^{*}(K)) &= E((\hat{\alpha}^{*}(K) - \alpha)'(\hat{\alpha}^{*}(K) - \alpha)) \\
&= \sum_{i=1}^{p} E(\hat{\alpha}_i^{*}(K) - \alpha_i)^2 \\
&= \sum_{i=1}^{p} \frac{\lambda_i\sigma^2 + k_i^2\alpha_i^2}{(\lambda_i + k_i)^2} \quad .
\end{aligned}
\tag{5.18}
$$

Comparing (2.53) and (5.18) it can be seen that the mean squared error function for $\hat{\alpha}^*(K)$ or $\hat{\beta}^*(K)$ is analogous to the corresponding function for $\hat{\beta}_k^*$. Invoking similar arguements as were employed in the proof of Theorem 2.4, it can be shown that $\hat{\alpha}^*(K)$ is mean squared error admissible when inequality (5.13) is satisfied for each component i.

Banerjee and Carr (1971) proposed a characterization of the generalized ridge estimator similar to the one they gave for $\hat{\beta}_k^*$. They suggested the augmentation of the existing data set (Y_X, X) with the data set (Y_A, V). V is an arbitrary p x p design matrix satisfying:

$$V'V = K \tag{5.19}$$

Y_A is the corresponding p x 1 vector of dependent variables which if observable would have been included in Y_X. In this case, the augmented model becomes:

$$\begin{bmatrix} Y_X \\ \hline Y_A \end{bmatrix} = \begin{bmatrix} X \\ \hline V \end{bmatrix} \beta + \varepsilon \tag{5.20}$$

An unbiased estimator of β based upon the augmented model (5.20) is given by the ordinary least squares estimator:

$$\begin{aligned} \hat{\beta}_A &= (X'X + V'V)^{-1}(X'Y_X + V'Y_A) \\ &= (X'X + K)^{-1}(X'Y_X + V'Y_A) \\ &= \hat{\beta}^*(K) + (X'X + K)^{-1}V'Y_A \end{aligned} \tag{5.21}$$

The generalized ridge estimator $\hat{\beta}^*(K)$ corresponds to the estimator of β obtained by dropping the term involving Y_A from (5.21). As was the case for the ridge estimator, Banerjee and Carr (1971) argued

that the relative accuracy of $\hat{\beta}^*(K)$ should be compared with it's corresponding unbiased estimator $\hat{\beta}_A$. By repeating the arguments employed in the proofs of Theorems 2.2 and 4.1, it can be seen that $\hat{\beta}^*(K)$ is mean squared error admissible when compared with $\hat{\beta}_A$.

Vinod (1976b) observed that the form of the ridge trace proposed by Hoerl and Kennard (1970a) cannot be applied to the generalized ridge estimator. It would not be practical to plot the estimated parameters against the different biasing parameters k_i. Instead, Vinod (1976b) suggested that a ridge trace can be utilized by modifying the multicollinearity allowance scale defined by (4.12). Suppose that a scale is formed by replacing k in (4.12) with the constants k_i which correspond to the different components of $\hat{\alpha}$. The resultant scale is of the form:

$$m = p - \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + k_i} \quad . \tag{5.22}$$

Vinod (1976b) put forward several procedures for plotting the generalized ridge estimates against the modified multicollinearity allowance scale. The criteria are designed to monitor the relative stability of the parameter estimates as the k_i's change.

Hocking, Speed and Lynn (1976) considered various procedures for calculating values for the k_i's directly from the data. They noted that the mean squared error function (5.18) is minimized when:

$$k_i = \frac{\sigma^2}{\alpha_i^2} \qquad (5.23)$$

Therefore, it would seem reasonable to estimate appropriate k_i's by substituting suitable estimates of α_i and σ^2 into (5.23). Hoerl and Kennard (1970a) proposed substituting the ordinary least squares estimates of α_i and σ^2 into (5.23). Let $\hat{k}_{i(0)}$ denote the estimator of k_i obtained by subsituting $\hat{\alpha}_i$ and $\hat{\sigma}^2$ into (5.23). $\hat{k}_{i(0)}$ parallels the estimator $\hat{k}_h$ which Hoerl, Kennard and Baldwin (1975) considered for the biasing parameter in the ridge estimator $\hat{\beta}_k^*$. Due to the tendency of $\hat{\alpha}'\hat{\alpha}$ to over-estimate $\alpha'\alpha$, Hoerl and Kennard (1970a) argued that an iterative procedure could be employed to obtain better estimates of the k_i's. Therefore, they suggested that the i'th component of the generalized ridge estimator obtained using $k_{i(0)}$ should be utilized to produce an improved estimate of k_i. Hoerl and Kennard (1970a) recommended repeating this procedure until the estimates of all the k_i's converge. Let $K_{(j)}$ denote the p x p diagonal

matrix whose i'th diagonal element equals the estimate of k_i corresponding to the j'th iteration. The general formula for Hoerl and Kennard's (1970a) iterative estimator of k_i becomes:

$$k_{i(j)} = \begin{cases} \dfrac{\hat{\sigma}^2}{\hat{\alpha}_i^2} & \text{for } j = 0 \\ \dfrac{\hat{\sigma}^2}{\hat{\alpha}_i^*(k_{(j-1)})^2} & \text{for } j = 1,2,3, \ldots \end{cases} \tag{5.24}$$

where:

$$\hat{\alpha}^*(K_{i(j)}) = (\Lambda + K_{(j)})^{-1}X^{*\prime}Y \tag{5.25}$$

Brown and Rock (1975) indicated a method for choosing the k_i's in such a manner as to minimize an estimate of the predictive mean squared error. Vinod (1976a) examined the relationship between the generalized ridge regression estimator and an estimator due to Bhattacharya (1966) stemming from Stein's (1960) proposals. He considered the generalized ridge regression estimator which is obtained when the k_i's are assigned the values:

$$k_i = \frac{\lambda_i}{\Delta_i} - \lambda_i \tag{5.26}$$

where the Δ_i's are Stein-like shrinkage factors similar to those described in Chapter 3. A heuristic modification of Bhattacharya's (1966) estimator which is analogous to the generalized ridge regression estimator for decreasing Δ_i's was also analysed by Vinod (1976a). In addition, the results of a number of simulation experiments to rank the mean squared errors for the various estimators were reported.

Hemmerle (1975) presented an explicit solution for the limiting values of the generalized ridge estimator using Hoerl and Kennard's (1970a) iterative procedure. In order to simplify the notation, the canonical form of the model will be considered. Using the results above, $\hat{\alpha}^*(K_{(j)})$ may be expressed in terms of the ordinary least squares solution as follows:

$$\begin{aligned}\hat{\alpha}^*(K_{(j)}) &= (\Lambda + K_{(j)})^{-1}X^*Y \\ &= (\Lambda + K_{(j)})^{-1}(X^*X^*)\hat{\alpha} \\ &= (\Lambda + K_{(j)})^{-1}\Lambda\hat{\alpha} .\end{aligned} \tag{5.27}$$

Let $\hat{\alpha}^*_{i(j)}$ denote the i'th component of $\hat{\alpha}^*(K_{(j)})$. Combining (5.24) and (5.27), the i'th component of $\hat{\alpha}^*(K_{(j)})$ becomes:

$$\begin{aligned}\hat{\alpha}^*_{i(j)} &= \frac{\lambda_i}{\lambda_i + k_{i(j)}}\hat{\alpha}_i \\ &= b_{i(j)}\hat{\alpha}_i\end{aligned} \tag{5.28}$$

where:

$$b_{i(j)} = \frac{\hat{\alpha}^*_{i(j-1)}{}^2}{\hat{\alpha}^*_{i(j-1)}{}^2 + \frac{\sigma^2}{\lambda_i}} \tag{5.29}$$

Adopting the methodology of Hocking, Speed and Lynn (1976), $b_{i(j)}$ may be expressed as:

$$b_{i(j)} = \frac{b_{i(j-1)}{}^2\hat{\alpha}_i^2}{b_{i(j-1)}{}^2\hat{\alpha}_i^2 + \frac{\sigma^2}{\lambda_i}} \tag{5.30}$$

Let L denote the ratio $\sigma^2/(\lambda_i\hat{\alpha}_i^2)$. In this case, $b_{i(j)}$ becomes:

$$b_{i(j)} = \frac{b_{i(j-1)}{}^2}{b_{i(j-1)}{}^2 + L} \tag{5.31}$$

The convergence theorem from the Appendix may be invoked to find the limiting values of $b_{i(j)}$ and hence a limiting solution for Hoerl and Kennard's (1970a) iterative generalized ridge regression estimator. Hoerl and Kennard (1970a) argued that initial estimates of the k_i's should be obtained by substituting ordinary least squares estimates of α_i and σ^2 into (5.23). In this case,

$$\begin{aligned}\hat{\alpha}^*_{i(0)} &= (1 + \hat{\sigma}^2/(\lambda_i\hat{\alpha}_i^2))^{-1}\hat{\alpha}_i \\ &= (1 + L)^{-1}\hat{\alpha}_i \qquad (5.32)\end{aligned}$$

so that $b_{i(0)}$ should be assigned the value $(1 + L)^{-1}$. Applying the convergence theorem from the Appendix, it follows that the limiting values of $b_{i(j)}$ satisfy:

$$b_i^* = \begin{cases} 0 & \text{if } L > \frac{1}{4} \\ \frac{1}{2} + (\frac{1}{4} - L)^{\frac{1}{2}} & \text{otherwise} \end{cases} \qquad (5.33)$$

Combining (5.28) and (5.33), it can be seen that the limiting values of the i'th component of $\hat{\alpha}^*(K_{(j)})$ are given by:

$$\hat{\alpha}_i^* = \begin{cases} 0 & \text{if } L > \frac{1}{4} \\ (\frac{1}{2} + (\frac{1}{4} - L)^{\frac{1}{2}})\hat{\alpha}_i & \text{otherwise} \end{cases} \qquad (5.34)$$

Hemmerle (1975) expressed his solution for the limiting values of Hoerl and Kennard's (1970a) iterative generalized ridge estimator in a simpler form than (5.34). Instead of determining the limiting values of $b_{i(j)}$, he examined the behavour of the ratios:

$$\begin{aligned}e_{i(j)} &= \frac{\hat{\sigma}^2}{\lambda_i(\hat{\alpha}^*_{i(j-1)})^2} \\ &= \frac{\hat{\sigma}^2}{\lambda_i b_{i(j-1)}^2\hat{\alpha}_i^2}\end{aligned}$$

$$= \frac{L}{b_{i(j-1)}^{2}} \quad . \tag{5.35}$$

for all j greater than or equal to zero. In (5.35), it is assumed that $b_{i(-1)}$ equals 1 so that $e_{i(0)}$ becomes L. Comparing (5.35) and (5.33) it can be seen that:

$$\hat{\alpha}^{*}_{i(j)} = b_{i(j)} \hat{\alpha}_i$$

$$= (1 + e_{i(j)})^{-1} \hat{\alpha}_i \quad . \tag{5.36}$$

Therefore, the limiting values of $\hat{\alpha}^{*}_{i(j)}$, $b_{i(j)}$ and $e_{i(j)}$ satisfy the relationship:

$$\hat{\alpha}^{*}_i = b^{*}_i \hat{\alpha}_i$$

$$= (1 + e^{*}_i)^{-1} \hat{\alpha}_i \quad . \tag{5.37}$$

The long run values of $e_{i(j)}$ may be found by directly applying (5.33). Consider first the case where $e_{i(0)}$ or L is greater than 0.25. Since $e_{i(j)}$ is related to $b_{i(j-1)}$ by (5.35) and b^{*}_i equals zero, it follows that e^{*}_i is infinite. In the case where $e_{i(0)}$ is less than or equal to 0.25, substituting (5.33) into (5.37) gives:

$$(1 + e^{*}_i)^{-1} = \tfrac{1}{2} + (\tfrac{1}{4} - L)^{\frac{1}{2}}$$

$$= \tfrac{1}{2} + (\tfrac{1}{4} - e_{i(0)})^{\frac{1}{2}} \tag{5.38}$$

or:

$$e^{*}_i = \frac{1}{2e_{i(0)}} \left((1 - 2e_{i(0)}) - (1 - 4e_{i(0)})^{\frac{1}{2}} \right) \quad . \tag{5.39}$$

Therefore, Hemmerle's (1975) form of the solution for the limiting values of the individual components of $\hat{\alpha}^{*}(K_{(j)})$ is given by:

$$\hat{\alpha}^{*}_i = \begin{cases} 0 & \text{if } e_{i(0)} = L > \tfrac{1}{4} \\ (1 + e^{*}_i)^{-1} \hat{\alpha}_i & \text{otherwise} \end{cases} \tag{5.40}$$

where e_i^* satisfies (5.39) Based upon this formulation of the limiting values of the generalized ridge regression estimator, Hemmerle (1975) proposed a test of the hypothesis that α_i equals zero by comparing $1/e_{i(0)}$ with a $F_{1,n-p-1}$ variate.

The explicit solution for the limiting values of Hoerl and Kennard's (1970a) iterative generalized ridge estimator was developed by Hemmerle (1975) without regard to the resulting increase in the residual sum of squares. He noted that in most practical applications of ridge regression some constraint should be placed upon the increase in the residual sum of squares. To this end, Hemmerle (1975) defined $\Delta^*_{(j)}$ to be the increase in the residual sum of squares which results when $\hat{\alpha}^*(K_{(j)})$ is employed as an estimator of α instead of $\hat{\alpha}$. As a result of (2.12), it follows that $\Delta^*_{(j)}$ satisfies:

$$\begin{aligned}\Delta^*_{(j)} &= \Phi(\hat{\alpha}^*(K_{(j)})) - \Phi(\hat{\alpha}) \\ &= (\hat{\alpha} - \hat{\alpha}^*(K_{(j)}))'X^{*\prime}X^*(\hat{\alpha} - \hat{\alpha}^*(K_{(j)})) \\ &= (\hat{\alpha} - \hat{\alpha}^*(K_{(j)}))'\Lambda(\hat{\alpha} - \hat{\alpha}^*(K_{(j)})) \quad . \end{aligned} \tag{5.41}$$

Suppose that the i'th component of the increase in the residual sum of squares corresponding to the j'th iteration of Hoerl and Kennard's (1970a) iterative generalized ridge estimator is denoted by $C_{i(j)}$. Applying (5.35), it follows that:

$$\begin{aligned}C_{i(j)} &= (\hat{\alpha}^*_{i(j)} - \hat{\alpha}_i)^2 \lambda_i \\ &= \hat{\sigma}^2\left(\frac{1}{\sqrt{e_{i(j)}}} - \frac{1}{\sqrt{e_{i(0)}}}\right)^2 \\ &= \hat{\sigma}^2 \frac{e_{i(j-1)}}{e_{i(j)}}^2 \quad . \end{aligned} \tag{5.42}$$

Assuming that $e_{i(0)}$ is less than or equal to 0.25, the limiting value for the i'th component of the increase in the residual sum of squares is given by:

$$\lim_{j \to +\infty} C_{i(j)} = \hat{\sigma}^2 e_i^* \quad . \tag{5.43}$$

In the case where $e_{i(0)}$ is greater than 0.25,

$$\lim_{j \to +\infty} C_{i(j)} = \hat{\sigma}^2 \lim_{j \to +\infty} \frac{e_{i(j-1)}^2}{e_{i(0)}(1 + e_{i(j-1)})}$$

$$= \frac{\hat{\sigma}^2}{e_{i(0)}} \quad . \tag{5.44}$$

In order to calculate the limiting value of the total increase in the residual sum of squares, it is necessary to partition the explanatory variables into the two sets:

$$a = \{ i \mid e_{i(0)} \leq 0.25 \} \tag{5.45}$$

$$b = a^c \quad .$$

Combining (5.43) and (5.44) it follows that:

$$\lim_{j \to +\infty} \Delta^*_{(j)} = \lim_{j \to +\infty} \sum_{i=1}^{p} C_{i(j)}$$

$$= \lim_{j \to +\infty} \{ \sum_a C_{i(j)} + \sum_b C_{i(j)} \}$$

$$= \hat{\sigma}^2 \{ \sum_a e_i^* + \sum_b \frac{1}{e_{i(0)}} \} \quad . \tag{5.46}$$

Hemmerle (1975) considered the situation where it is desired to constrain the generalized ridge regression estimates so that the increase in the residual sum of squares is no more than 100M%. In other words, it is required that:

$$\Delta^* = \lim_{j \to +\infty} \Delta^*_{(j)}$$

$$\leq M(n - p - 1)\ \hat{\sigma}^2$$

$$= M^* \quad . \tag{5.47}$$

This implies that the limiting solution defined by (5.40) may only be accepted if M^* is not less than Δ^*. Otherwise, the solution must be modified so that (5.47) is satisfied.

Hemmerle (1975) put forward two different procedures for modifying (5.40) so that the increase in the residual sum of squares is less than or equal to 100M%. First, he suggested iterating $\hat{\alpha}^*(K_{(j)})$ until the condition:

$$\Delta^* > M^* \tag{5.48}$$

is reached. Alternatively, he proposed constraining each component of the increase in the residual sum of squares separately An upper bound M_i^* on the increase in the residual sum of squares corresponding to each component must be choosen such that:

$$\sum_{i=1}^{p} M_i^* = M^* \quad . \tag{5.49}$$

The i'th component of (5.40) would be accepted if:

$$C_i^* = \lim_{j \to +\infty} C_{i(j)}$$

$$\leq M_i^* \tag{5.50}$$

Clearly condition (5.50) can only be satisfied if $e_{i(0)}$ is less than or equal to 0.25. For all other cases, Hemmerle (1975) proposed that C_i^* be set to M_i^* and the resulting equation solved for an appropriate k_i. Adopting the notation employed in (2.22), the i'th component of $\hat{\alpha}^*(K)$ may be expressed as:

$$\hat{\alpha}_i^*(K) = \frac{c_i}{\lambda_i + k_i} \quad . \tag{5.51}$$

Combining (5.41) and (5.51), it can be seen that the increase in the i'th component of the residual sum of squares when $\hat{\alpha}$ is replaced by $\hat{\alpha}^*(K)$ is given by:

$$C_i = \lambda_i(\hat{\alpha}_i - \hat{\alpha}_i^*(K))^2$$

$$= \lambda_i\left(\frac{c_i}{\lambda_i} - \frac{c_i}{\lambda_i + k_i}\right)^2$$

$$= \frac{k_i^2}{\lambda_i(\lambda_i + k_i)^2} c_i^2$$

$$= k_i^2 \frac{\lambda_i}{(\lambda_i + k_i)^2} \hat{\alpha}_i^2$$

$$= \frac{k_i^2 \hat{\sigma}^2}{e_{i(0)}(\lambda_i + k_i)^2} \qquad (5.52)$$

It follows from (5.52) that k_i should be choosen such that:

$$M_i^* = \frac{k_i^2 \hat{\sigma}^2}{e_{i(0)}(\lambda_i + k_i)^2} \qquad (5.53)$$

or:

$$k_i = \frac{\lambda_i \sqrt{M_i^* e_{i(0)}}}{\hat{\sigma} - \sqrt{M_i^* e_{i(0)}}} \qquad (5.54)$$

whenever condition (5.50) is not satisfied.

It should be noted that Hemmerle's (1975) second approach to constraining the generalized ridge estimates is rather arbitrary in that choices for the M_i^*'s must be made. Hemmerle (1975) suggested that the M_i^*'s be made proportional to the limiting values of $C_{i(j)}$. In a later paper, Hocking, Speed and Lynn (1976) reviewed Hemmerle's

(1975) constrained solutions for the limiting values of $\hat{\alpha}^*(K_{(j)})$. They agreed that the unconstrained iterative solution is often inappropriate and that some constraint might be desirable. However, they objected to the constrained solutions put forward by Hemmerle (1975) since these solutions choose finite values for each k_i and thus force the estimates of all the components of α to be nonzero. Hocking, Speed and Lynn (1976) argued that the influence of a very small eigenvalue which is removed in the unconstrained solution is reintroduced using (5.54).

Allen (1972) described several procedures for selecting values of the k_i's for the generalized ridge estimators when the linear regression model is to be employed primarily to predict the dependent variable. In these situations, he argued the k_i's should be choosen to minimize some measure of the prediction error. Allen (1972) suggested that the criterion utilized to measure the prediction errors should be small when the predicted values are close to the observed values and include a penalty for increasing the variance of the predictor. Further, the criterion should only be based upon predictions of the actual observations of Y since the regression model is not necessarily valid outside the ranges of the dependent variables actually utilized in the calculation of the estimates of β.

Allen (1972) recommended two criteria as being appropriate for determining values for the k_i's when the regression equation

is to be used to predict the dependent variable. First, he considered the total mean squared error of prediction:

$$\begin{aligned} MSE(\hat{Y}) &= E((X\hat{\beta}^*(K) - Y)'(X\hat{\beta}^*(K) - Y)) \\ &= E((X\hat{\beta}^*(K) - X\beta - \varepsilon)'(X\hat{\beta}^*(K) - X\beta - \varepsilon)) \\ &= n\sigma^2 + \sigma^2 tr((X'X + K)^{-1}X'X)^2 \\ &\quad + \beta'X'(I_n - X(X'X + K)^{-1}X')X\beta \quad . \end{aligned} \tag{5.55}$$

The mean squared error function defined by (5.55) cannot be evaluated directly since it depends upon the unknown values of β and σ^2. Instead, Allen (1972) suggested that a diagonal matrix K be obtained by minimimizing an estimator of (5.55). To this end, he put forward the estimator:

$$M(D) = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2 + 2tr(I_n + (X'X)^{-1}K)S^2 \tag{5.56}$$

where S^2 is an estimator of σ^2.

In addition to the mean squared error criterion, Allen (1972) considered the Predictive Sum Of Squares (PRESS):

$$PRESS(K) = \sum_{i=1}^{n} (\hat{Y}_{(i)} - Y_i)^2 \tag{5.57}$$

where $\hat{Y}_{(i)}$ is the estimate of Y_i which is obtained by deleting the i'th observations of the dependent and independent variables from the estimation procedure for β. The PRESS criterion predicts each observation using the other (n - 1) observations. Allen (1972) noted that both criteria described above involve nonlinear functions of the matrix K. He recommended that local minimum points for these functions be obtained using a standard nonlinear regression technique

such as that described by Marquardt (1963).

Although Hemmerle (1975) presented an explicit solution for Hoerl and Kennard's (1970a) iterative generalized ridge regression estimator and the associated convergence conditions, he did not investigate the properties of the resulting estimators. Dwivedi, Srivastava and Hall (1976) derived exact expressions for the first and second moments for the first iterate of Hoerl and Kennard's (1970a) estimator. Assuming that the unobservable disturbances satisfy conditions (1.2), they showed that the first and second moments for $\hat{\alpha}_i^*(K_{(0)})$ are given by:

$$E(\hat{\alpha}_i^*(K_{(0)})) = \alpha_i e^{-\delta_i} \sum_{m=0}^{\infty} \sum_{j=0}^{\infty} \Gamma\left(\frac{\nu-1}{\nu}\right) \frac{m\Gamma(m+\frac{\nu}{2})\Gamma(j+\frac{\nu+3}{2})\Gamma(j+\frac{5}{2}) \quad \delta_i^j}{\Gamma(\frac{\nu}{2})\Gamma(\nu+j+\frac{\nu+5}{2})\Gamma(j+\frac{3}{2})\Gamma(j+1)} \tag{5.58}$$

$$E(\hat{\alpha}_i^*(K_{(0)})^2) = 2\frac{\sigma^2}{\lambda_i} e^{-\delta_i} \sum_{m=0}^{\infty} \sum_{j=0}^{\infty} (m+1)\left(\frac{\nu+1}{\nu}\right) \frac{m \ \Gamma(m+\frac{\nu}{2})\Gamma(j+\frac{\nu+3}{2})\Gamma(j+\frac{7}{2}) \ \delta_i^j}{\Gamma(\frac{\nu}{2})\Gamma(m+j+\frac{\nu+7}{2})\Gamma(j+\frac{1}{2})\Gamma(j+1)} \quad \ldots \tag{5.59}$$

where:

$$\delta_i = \frac{\lambda_i \beta_i^2}{\sigma^2} \qquad \nu = n - p \tag{5.60}$$

In addition, Dwivedi, Srivastava and Hall (1976) conjectured that similar results may be derived for the higher order iterates of the estimator by applying their methodology.

Dwivedi, Srivastava and Hall (1976) numerically evaluated (5.58) and (5.59) by calculating the relative bias, relative mean squared error and relative efficiency of $\hat{\alpha}^*(K_{(0)})$ with respect to the ordinary least squares estimator. Based upon these calculations, they concluded that:

i) $\hat{\alpha}_i^*(K_{(0)})$ is biased in the direction which is opposite to the sign of α_i.

ii) The relative bias is a decreasing function of δ_i and an increasing function of ν.

iii) The relative mean squared error decreases as δ_i increases. Provided that δ_i is less than or equal to one, the mean squared error decreases as ν grows large.

iv) $\hat{\alpha}_i^*(K_{(0)})$ is a more efficient estimator of α_i than $\hat{\alpha}_i^{\circ}$ provided that δ_i is less than or equal to one. There is a substantial gain in efficiency if δ_i is small and ν is large.

As a result of these conclusions, Dwivedi, Srivastava and Hall (1976) argued that the parameter estimates for the general model may be substantially improved if $\hat{\alpha}^*(K_{(0)})$ is employed instead of $\hat{\alpha}$.

Hemmerle and Brantle (1976) explored an alternative approach to calculating generalized ridge regression estimates. Instead of estimating the optimum k_i's defined by (5.23), they proposed mini-

mizing an estimator of the mean squared error function for $\hat{\alpha}^*(K)$. Again the canonical form of the general model will be utilized. Consider the expected sum of squares of the differences between the individual components of $\hat{\alpha}^*(K)$ and the corresponding components of $\hat{\alpha}$:

$$
\begin{aligned}
&E((\hat{\alpha}^*(K) - \hat{\alpha})'(\hat{\alpha}^*(K) - \hat{\alpha})) \\
&= E(Y'X^*((\Lambda + K)^{-1} - \Lambda^{-1})^2 X^{*\prime} Y) \\
&= \sigma^2 tr(X^*((\Lambda + K)^{-1} - \Lambda^{-1})^2 X^{*\prime}) \\
&\qquad + \alpha' X^{*\prime} X^* ((\Lambda + K)^{-1} - \Lambda^{-1})^2 X^{*\prime} X^* \alpha \\
&= \sigma^2 tr(X^{*\prime} X^* ((\Lambda + K)^{-1} - \Lambda^{-1})^2) + \alpha'((I_p + K\Lambda^{-1})^{-1} - I_p)^2 \alpha \\
&= \sigma^2 tr(\Lambda((\Lambda + K)^{-1} - \Lambda^{-1})^2) + \alpha'((I_p + K\Lambda^{-1})^{-1} - I_p)^2 \alpha \\
&= \sigma^2 \sum_{i=1}^{p} \frac{k_i^2}{(\lambda_i + k_i)^2 \lambda_i} + \sum_{i=1}^{p} \frac{k_i^2 \alpha_i^2}{(\lambda_i + k_i)^2} \qquad (5.61)
\end{aligned}
$$

Combining (5.18) and (5.61) it follows that:

$$
\begin{aligned}
&E((\hat{\alpha}^*(K) - \alpha)'(\hat{\alpha}^*(K) - \alpha)) - E((\hat{\alpha}^*(K) - \hat{\alpha})'(\hat{\alpha}^*(K) - \hat{\alpha})) \\
&= \sigma^2 \sum_{i=1}^{p} \frac{(\lambda_i - k_i)}{\lambda_i(\lambda_i + k_i)} \qquad (5.62)
\end{aligned}
$$

As a result of (5.62) Hemmerle and Brantle (1976) proposed the following unbiased estimator of the mean squared error function for the generalized ridge estimator:

$$
\begin{aligned}
\hat{L}^2(\hat{\alpha}^*(K)) &= E((\hat{\alpha}^*(K) - \hat{\alpha})'(\hat{\alpha}^*(K) - \hat{\alpha})) \\
&\quad + \hat{\sigma}^2 \sum_{i=1}^{p} \frac{(\lambda_i - k_i)}{\lambda_i(\lambda_i + k_i)} \qquad (5.63)
\end{aligned}
$$

Hemmerle and Brantle (1976) proposed that an estimator of α

be obtained by minimizing each component of $\hat{L}^2(\hat{\alpha}^*(K))$. Let

$$v_i = \frac{\lambda_i}{\lambda_i + k_i} \tag{5.64}$$

so that:

$$\hat{\alpha}_i^*(K) = v_i \hat{\alpha}_i \quad . \tag{5.65}$$

Therefore the i'th component of $\hat{L}^2(\hat{\alpha}^*(K))$ becomes:

$$\begin{aligned} M_i &= (\hat{\alpha}_i^*(K) - \hat{\alpha}_i)^2 + \hat{\sigma}^2 \frac{(\lambda_i - k_i)}{\lambda_i(\lambda_i + k_i)} \\ &= \hat{\alpha}_i^2 (v_i - 1)^2 + \frac{\hat{\sigma}^2}{\lambda_i} (2v_i - 1) \quad . \end{aligned} \tag{5.66}$$

Adopting Hemmerle's (1975) definition of $e_{i(0)}$, M_i may be rewritten as:

$$\begin{aligned} M_i &= \hat{\alpha}_i^2 (v_i - 1)^2 + e_{i(0)} \hat{\alpha}_i^2 (2v_i - 1) \\ &= \hat{\alpha}_i^2 \{ (v_i - 1 + e_{i(0)})^2 + e_{i(0)} - e_{i(0)}^2 \}^2 \end{aligned} \tag{5.67}$$

Minimizing M_i with respect to v_i subject to the constraint that v_i lies in the interval $[0,1]$ leads to:

$$v_i = \begin{cases} 1 - e_{i(0)} & \text{if } e_{i(0)} < 1 \\ 0 & \text{otherwise} \end{cases} \quad . \tag{5.68}$$

By substituting (5.68) into (5.65) Hemmerle and Brantle (1976) obtained their estimator:

$$\hat{\alpha}_i^*(K) = \begin{cases} (1 - e_{i(0)}) \hat{\alpha}_i & \text{if } e_{i(0)} < 1 \\ 0 & \text{otherwise} \end{cases} \tag{5.69}$$

for the i'th component of α. In a further application of generalized ridge regression, Hemmerle and Brantle (1976) applied quadratic programming methods to obtain an estimator which satisfies constraints of the form:

$$C \hat{\beta}^*(K) > b \quad . \tag{5.70}$$

Goldstein and Smith (1974) provided an extension to their shrinkage estimator which admits the generalized ridge estimator as one of it's forms. Instead of considering the shrinkage estimators of the form (3.141) which utilize a single parameter k, they put forward the shrinkage estimator:

$$\hat{\alpha}_i^* = c_i z_i$$
$$= c(\lambda_i^{\frac{1}{2}}, k_i) z_i \qquad (5.71)$$

where k_i is any non-negative real number dependent upon the component of Z. Again the shrinkage functions must satisfy the two conditions:

$$c(\lambda_i^{\frac{1}{2}}, 0) = \frac{1}{\lambda_i^{\frac{1}{2}}} \qquad (5.72)$$

and:

$$\frac{1}{\lambda_i^{\frac{1}{2}}} \frac{\delta}{\delta k_i} c(\lambda_i^{\frac{1}{2}}, k_i) < 0 \qquad (5.73)$$

for all non-negative values of k_i. Goldstein and Smith (1974) noted that the generalized ridge estimator may be obtained from (5.71) by choosing c_i's of the form:

$$c_i = c(\lambda_i^{\frac{1}{2}}, k_i)$$
$$= \frac{\lambda_i^{\frac{1}{2}}}{(\lambda_i + k_i)} \qquad (5.74)$$

Further, they noted that the mean squared error function for the resulting estimator is a monotonically decreasing function of k_i from 0 to σ^2/α_i^2 and a monotonically increasing function after that point. As a result, Goldstein and Smith (1974) argued that an appropriate estimator for each k_i could be obtained by utilizing the ordinary least squares estimators for α_i and σ^2 to estimate σ^2/α_i^2.

Adopting results from Chapter 3, it can be seen that the ordinary least squares estimator of α_i is:

$$\hat{\alpha}_i = \lambda_i^{-\frac{1}{2}} z_i \qquad \text{for } i = i,2,3, \ldots p \qquad (5.75)$$

Goldstein and Smith (1974) recommended that the (n - p) observations of z_i not utilized to estimate the α_i's be employed to estimate σ^2. The resulting estimator of k_i becomes:

$$\hat{k}_i = \frac{\lambda_i}{z_i^2} \frac{1}{(n-p)} \sum_{j=p+1}^{n} z_j^2 \qquad (5.76)$$

Hocking, Speed and Lynn (1976) noted that each of the biased estimators outlined in Chapter 2 and 3 may be constructed by placing constraints upon the ordinary least squares estimators. For example, the ridge estimator $\hat{\beta}_k^*$ and Mayer and Willke's (1973) deterministically shrunken estimator minimize the residual sum of squares for a fixed parameter length. The ridge estimator utilizes the Euclidean norm to measure parameter lengths while the deterministically shrunken estimator employs a design dependent norm. Hocking, Speed and Lynn (1976) argued that since the generalized ridge estimator is not formed by placing constraints upon the ordinary least sqyares estimator, the generalized ridge estimator should be potentially superior to the other biased estimators. However, it should be noted that this increase in the flexibility of the estimator is achieved at the expense of a loss in the ability to geometrically interprete the estimator. Unlike the ridge estimator and Mayer and Willke's (1973) estimator, there is no geometrical interpretation for the generalized ridge estimator.

A couple of numerical examples have been provided in the ridge regression literature to demonstrate the use of the generalized ridge estimator. Hocking, Speed and Lynn (1976) reanalysed the pit-props and air pollution data whose ridge solutions were originally presented by Hoerl and Kennard (1970b) and McDonald and Schwing (1973) respectively. They employed the limiting solution proposed by Hemmerle (1975). In addition, Hemmerle and Brandle (1976) utilized their alternative procedure for determining a limiting solution for the generalized ridge estimator to estimate parameters for McDonald and Schwing's (1973) air pollution data. They found that their solution was more conservative than Hemmerle's (1975) in terms of it's departure from the ordinary least squares solution. Hemmerle and Brantle (1976) argued that this is due to a constraint upon the residual sum of squares which is implicit in their solution. The quantity $(\hat{\alpha}^*(K) - \hat{\alpha})'(\hat{\alpha}^*(K) - \hat{\alpha})$ in the mean squared error function (5.63) discourages wide departures of the components of $\hat{\alpha}^*(K)$ from the components of $\hat{\alpha}$. Besides the numerical examples mentioned above, Hemmerle (1975) constructed an example to show the convergence of his solution may be very slow if one of the values of $e_{i(0)}$ is close to 0.25.

As part of a simulation study to test the efficiency of the generalized ridge estimator, Guilkey and Murphy (1975) introduced the concept of the directed ridge estimator. Guilkey and Murphy (1975) noted that the ordinary least squares estimator $\hat{\alpha}$ may produce a relatively precise estimate of α_i if λ_i is a large eigenvalue. As a result, they argued that the iterative generalized ridge estimator

defined by (5.24) and (5.25) should only be used to alter the diagonal elements of Λ corresponding to relatively small eigenvalues. To be more specific, they defined an eigenvalue λ_i to be relatively small if:

$$\lambda_i < 10^{-c} \lambda_{max} \tag{5.77}$$

where c is an arbitrary constant. Suppose that I_i is an indicator variable which takes on the value one if condition (5.77) is satisfied and zero otherwise. Incorporating this indicator variable, Guilkey and Murphy's (1975) directed ridge estimator may be expressed as:

$$\hat{\alpha}^*(K^*_{(j)}) = (\Lambda + K^*_{(j)})^{-1} X^* Y \tag{5.78}$$

where $K^*_{(j)}$ is a diagonal matrix whose i'th diagonal element is given by:

$$k^*_{i(j)} = \begin{cases} I_i \dfrac{\hat{\sigma}^2}{\hat{\alpha}_i^2} & \text{for } j = 0 \\ I_i \dfrac{\hat{\sigma}^2}{\hat{\alpha}_i^*(k^*_{i(j-1)})} & \text{for } j = 1,2,3, \ldots \end{cases} \tag{5.79}$$

In addition to (5.78) Guilkey and Murphy (1975) proposed a simplified version of their directed ridge estimator which utilizes a single value k for all nonzero biasing parameters. They suggested that the parameter k be choosen by allowing k to increase until the residual sum of squares has increase from $(n - p - 1)\hat{\sigma}^2$ to $q(n - p - 1)\hat{\sigma}^2$.

Guilkey and Murphy (1975) developed a number of simulations to compare the efficiencies of their directed ridge estimator with the ordinary least squares estimator and the generalized ridge estimator obtained by substituting ordinary least squares estimates of α_i and

σ^2 into (5.23). Provided that the degree of multicollinearity in the simulated explanatory variables was large enough, they found that the directed ridge and generalized ridge estimators were generally better than the ordinary least squares estimators in terms of their associated mean squared errors. The most dramatic reductions in the mean squared errors were reported for the directed ridge estimator defined by (5.78). Other simulation studies utilizing the generalized ridge estimator were reported by Lawless and Wang (1976) and Hemmerle and Brantle (1976).

Sommers (1964) introduced a generalization of the ridge estimator based upon powers of the $X'X$ matrix. He recommended an estimator of the form:

$$\hat{\beta}^*(k,q) = ((X'X)^q + kI_p)^{-1}(X'X)^{q-1}X'Y \tag{5.80}$$

where $k \geq 0$ and q is any non-negative integer. Dwivedi (1973) and Goldstein and Smith (1974) independently contemplated the same generalization. In a later paper, Hoerl and Kennard (1975) demonstrated that $\hat{\beta}^*(k,q)$ is one of the classes of generalized ridge estimators defined by (5.2) and (5.4). In particular, they showed that:

$$\begin{aligned} P'\hat{\beta}^*(k,q) &= P'((X'X)^q + kI_p)^{-1}(X'X)^{q-1}X'Y \\ &= P'(P\Lambda^q P' + kI_p)^{-1}P\Lambda^{q-1}P'X'Y \\ &= (\Lambda^q + kI_p)^{-1}\Lambda^{q-1}P'X'Y \\ &= (\Lambda + k\Lambda^{1-q})^{-1}P'X'Y \\ &= (\Lambda + K)^{-1}X^* Y \end{aligned} \tag{5.81}$$

where K is a diagonal matrix with it's i'th diagonal element equal

to k/λ_i^{q-1}. By comparing (5.4) and (5.81), it can be seen that the power generalization of the ridge estimator is a particular case of the generalized ridge estimator.

Goldstein and Smith (1974) noted that the power generalization defined by (5.80) makes the data analysis more sensitive to the eigenvalue spectrum of the $X'X$ matrix. However, it can be seen from (5.23) that the optimum k_i's for the generalized ridge estimator are inversely proportional to the squares of the ordinary least squares estimates $\hat{\alpha}$ not the eigenvalues λ_i. As a result, Hoerl and Kennard (1975) argued that an estimator of β based upon the powers of the $X'X$ matrix may not be too helpful in practice.

Later, Dwivedi (1973) analysed the properties of $\hat{\beta}^*(k,q)$ in more detail. Since:

$$\begin{aligned}\hat{\beta}^*(k,q) &= ((X'X)^q + kI_p)^{-1}(X'X)^{q-1}X'Y \\ &= C(k,q)\,\hat{\beta} \qquad (5.82)\end{aligned}$$

where:

$$\begin{aligned}C(k,q) &= ((X'X)^q + kI_p)^{-1}(X'X)^q \\ &= (I_p + k(X'X)^{-q})^{-1}, \qquad (5.83)\end{aligned}$$

it can be seen that $\hat{\beta}^*(k,q)$ is a linear transformation of the ordinary least squares estimator. The expected value of $\hat{\beta}^*(k,q)$ is given by:

$$\begin{aligned}E(\hat{\beta}^*(k,q)) &= ((X'X)^q + kI_p)^{-1}E(\hat{\beta}) \\ &= ((X'X)^q + kI_p)^{-1}\,\beta \quad . \qquad (5.84)\end{aligned}$$

It can be seen from (5.84) that $\hat{\beta}^*(k,q)$ will be a biased estimator

of β for all non-zero values of k. Assuming that the $X'X$ matrix is different from the identity matrix, $\hat{\beta}^*(k,q)$ will be a biased estimator whenever q is non-zero. The variance-covariance matrix for $\hat{\beta}^*(k,q)$ is given by:

$$\begin{aligned} Var(\hat{\beta}^*(k,q)) &= Var(C(k,q)\hat{\beta}) \\ &= \sigma^2(I_p + k(X'X)^{-q})^{-1}(X'X)^{-1}(I_p + k(X'X)^{-q})^{-1} \\ &= \sigma^2(X'X + k(X'X)^{1-q})^{-1}(X'X) \quad . \end{aligned} \tag{5.85}$$

Using (5.82), Dwivedi (1973) developed the following sufficient condition for the mean squared error admissibility of $\hat{\beta}^*(k,q)$:

<u>Theorem 5.2</u>: $\hat{\beta}^*(k,q)$ is a mean squared error admissible estimator of β in (1.1) if:

$$k < \frac{\sigma^2 \lambda_i^{q-1}}{\alpha_i^2} \tag{5.86}$$

for each component i.

<u>Proof</u>: By definition of the mean squared error function:

$$\begin{aligned} MSE(\hat{\beta}^*(k,q)) &= E((\hat{\beta}^*(k,q) - \beta)'(\hat{\beta}^*(k,q) - \beta)) \\ &= E((C\hat{\beta} - C\beta + C\beta - \beta)'(C\hat{\beta} - C\beta + C\beta - \beta)) \\ &= E((\hat{\beta} - \beta)'C'C(\hat{\beta} - \beta)) + \beta'(C - I_p)'(C - I_p)\beta \\ &= \sigma^2 tr(C^2(X'X)^{-1}) + \beta'(C - I_p)^2 \beta \end{aligned} \tag{5.87}$$

where:

$$C = C(k,q) \tag{5.88}$$

Applying the binomial inverse theorem,

$$\begin{aligned} C &= (I_p + k(X'X)^{-q})^{-1} \\ &= I_p - k(kI_p + (X'X)^{-q})^{-1} \quad . \end{aligned} \tag{5.89}$$

Substituting (5.89) into (5.87) the mean squared error function for $\hat{\beta}^*(k,q)$ becomes:

$$\begin{aligned} MSE(\hat{\beta}^*(k,q)) &= \sigma^2 tr(C^2(X'X)^{-1}) + \beta'(C - I_p)^2\beta \\ &= \sigma^2 tr((X'X)^{-1}(I_p + k(X'X)^{-q})^{-2}) \\ &\quad + k^2\beta'(kI_p + (X'X)^{-q})^{-2}\beta \\ &= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i^{q-1}}{(\lambda_i^q + k)} + k^2 \sum_{i=1}^{p} \frac{(k\alpha_i^2 - \sigma^2\lambda_i^{q-1})}{(\lambda_i^q + k)^2} \end{aligned} \quad \ldots (5.90)$$

Dwivedi (1973) observed that:

$$\frac{1}{\lambda_i} - \frac{\lambda_i^{q-1}}{(\lambda_i^q + k)} = \frac{k}{(\lambda_i^q + k)} > 0 \quad (5.91)$$

for all positive values of k. Combining (5.90) and (5.91), it follows that:

$$MSE(\hat{\beta}^*(k,q)) < MSE(\hat{\beta}) + k^2 \sum_{i=1}^{p} \frac{(k\alpha_i^2 - \sigma^2\lambda_i^{q-1})}{(\lambda_i^q + k)^2} . \quad (5.92)$$

It can be seen from (5.92) that $\hat{\beta}^*(k,q)$ is a mean squared error admissible estimator if:

$$k^2 \sum_{i=1}^{p} \frac{(k\alpha_i^2 - \sigma^2\lambda_i^{q-1})}{(\lambda_i^q + k)^2} < 0 \quad (5.93)$$

Inequality (5.93) will be satisfied and $\hat{\beta}^*(k,q)$ will be a mean squared error admissible estimator if:

$$k < \sigma^2 \frac{\lambda_i^{q-1}}{\alpha_i^2} \quad (5.94)$$

for each component i. Notice that this condition reduces to the standard admissibility condition for the ridge estimator when q equals 1. In addition, it can be seen that as q increases, the admissibility interval for k shrinks.

It can be seen from (1.24) and (5.90) that the mean squared error functions for $\hat{\beta}$ and $\hat{\beta}^*(k,q)$ are influenced to the greatest extent by the smallest eigenvalues of the $X'X$ matrix. As a result, Dwivedi (1973) suggested that it would be reasonable to concentrate on that part of the mean squared error function containing the smallest eigenvalues. To this end, they introduced the concept of the 'primary component' of an estimator. By definition, the 'primary component' of an estimator is that portion of the mean squared error function which involves terms containing eigenvalues less than or equal to one. Dwivedi (1973) suggested that attention could be restricted to 'principal components' when comparing two estimators. They realized that the choice of all eigenvalues less than or equal to one was arbitrary. However, Dwivedi (1973) argued that by considering all eigenvalues less than or equal to one, all eigenvalues which contribute significantly to the overall mean squared error would be included.

Suppose that the p eigenvalues for the $X'X$ matrix have been ordered so that:

$$0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_t \leq 1 < \lambda_{t-1} \leq \ldots \leq \lambda_p \quad . \tag{5.95}$$

As a result of (5.90) the 'primary component' of $\hat{\beta}^*(k,q)$ is given by:

$$PC(\hat{\beta}^*(K)) = \sigma^2 \sum_{i=1}^{t} \frac{\lambda_i^{q-1}}{(\lambda_i^q + k)} + k^2 \sum_{i=1}^{t} \frac{(k\alpha_i^2 - \sigma^2\lambda_i^{q-1})}{(\lambda_i^q + k)^2} \quad . \qquad \ldots (5.96)$$

Dwivedi (1973) demonstrated that the 'primary component' of $\hat{\beta}^*(k,q-1)$ will be less than the 'primary component' of $\hat{\beta}^*(k,q)$ provided that:

$$0 < k < g_i \qquad \text{for } i = 1,2,3 \ldots t \quad , \qquad (5.97)$$

where:

$$g_i = -\left\{\frac{\lambda_i^{q-1}(1+\lambda_i)(\lambda_i\alpha_i^2 - \sigma^2)}{4\alpha_i^2}\right\} + \left\{\frac{4\lambda_i^{2q}\alpha_i^2\sigma^2 + \lambda_i^{2q-2}(1+\lambda_i)^2(\lambda_i\alpha_i^2-\sigma^2)^2}{16\alpha_i^4}\right\}^{\frac{1}{2}} \quad . \qquad (5.98)$$

As a consequence of this last condition, they noted that the contribution of the 'primary component' to the mean squared error function decreases as q increases. In fact, it can be readily seen that g_i tends to zero as q tends to infinity. Therefore, the mean squared error admissible estimators of the form $\hat{\beta}^*(k,q)$ tend to the least squares solution as q increases. Dwivedi (1973) illustrated his thesis with some Monte Carlo simulations of the estimator $\hat{\beta}^*(k,q)$.

In some further remarks, Dwivedi (1973) described some of the difficulties inerent in the choice of values for k and

q. If q is at all large, the resultant estimator may be subject to significant round-off errors. It was suggested that k and q be choosen to minimize the sum of squares of the differences between $\hat{\beta}^*(k,q)$ and $\hat{\beta}$.

Chapter 6

Tests Of Hypotheses And Confidence Intervals For Ridge Regression

In the previous chapters, ridge regression was presented as a technique for improving the estimated coefficients for the general linear model when the $X'X$ matrix is ill-conditioned. One rational for using a biased estimator such as the ridge estimator in these situations is that the introduction of a small amount of bias into the the estimator can often result in a large reduction in the variance of the estimator. As a result, the mean squared error for the biased estimator can be smaller than that of the ordinary least squares estimator. Obenchain (1977) suggested that one might be tempted to use this logic to argue that the confidence intervals for the ridge estimator which are centered at the ridge estimates can be shorter than the centered confidence intervals for the ordinary least squares estimator. He pointed to Marquardt and Snee's (1975) paper in which they seemed to infer this property. Obenchain (1977) considered this particular question when he developed the confidence intervals for the ridge estimator. This section outlines the procedure for constructing confidence intervals or tests of hypotheses developed by Obenchain (1977) as well as an interesting test of hypothesis which he gave in an earlier paper.

The general linear model defined by (1.1) may be reformulated to explicitly include a constant term. In this case, the model becomes:

$$Y = \mu \underline{1} + X\beta + \varepsilon \tag{6.1}$$

where $\underline{1}$ represents a unit vector. The parameter vector β will again be assumed to be a p x 1 vector. In the remainder of this chapter, it will be assumed that the mean observation has been subtracted from each observation of the dependent and independent variables so that the location parameter μ can be ignored without loss of generality. It will also be assumed that Y is normally distributed conditional upon X.

In order to simplify the notation in this chapter, it is convenient to rewrite the generalized ridge estimator as:

$$\begin{aligned} \hat{\beta}^*(K) &= (X'X + K)^{-1}X'Y \\ &= P(\Lambda + K)^{-1}\Lambda\hat{\alpha} \\ &= P\,\Delta\hat{\alpha} \end{aligned} \tag{6.2}$$

where:

$$\Delta = (\Lambda + K)^{-1}\Lambda \tag{6.3}$$

The i'th diagonal element of Δ is denoted by δ_i. δ_i corresponds to the shrinkage factor which is applied to the i'th component of the ordinary least squares estimator of α by the generalized ridge estimator. It is necessary to assume that the δ_i's are known quantities for the derivations contained in this section.

The general linear hypothesis can be denoted by:

$$H: \quad A\beta = \rho \tag{6.4}$$

where: A is a known r x p matrix; ρ is a known r x 1 vector and r is less than or equal to p. It is assumed that the rank of the

matrix A is r.

An estimator of β which satisfies the general linear hypothesis can be constructed by minimizing the residual sum of squares function $\Phi(b)$ subject to the constraints implied by (6.4). The required Lagrangian equation is of the form:

$$L(b,\eta) = \Phi(b) - 2\eta'(Ab - \rho)$$

$$= (Y - Xb)'(Y - Xb) - 2\eta'(Ab - \rho) \quad . \qquad (6.5)$$

The partial derivatives of (6.5) with respect to b and η must satisfy:

$$\frac{\delta}{\delta b} L(b,\eta) = -2X'Y + 2X'Xb - 2A'\eta$$

$$= 0 \qquad (6.6)$$

and:

$$\frac{\delta}{\delta \eta} L(b,\eta) = -2Ab + 2\rho$$

$$= 0 \quad . \qquad (6.7)$$

Premultiplying by $A(X'X)^{-1}$ and substituting (6.7) into (6.6) gives:

$$A(X'X)^{-1}A'\eta = \rho - A\hat{\beta} \quad . \qquad (6.8)$$

Since A is a r x p matrix of rank r, the inverse of the $A(X'X)^{-1}A'$ matrix exists and:

$$\eta = (A(X'X)^{-1}A')^{-1}(\rho - A\hat{\beta}) \quad . \qquad (6.9)$$

Substituting (6.9) into (6.6) gives:

$$b = (X'X)^{-1}(X'Y + A'\eta)$$

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

$$= \hat{\beta} - A^{*}(A\hat{\beta} - \rho) \quad , \tag{6.10}$$

where:

$$A^{*} = (X'X)^{-1}A'(A(X'X)^{-1}A')^{-1} \quad . \tag{6.11}$$

Equations (6.10) and (6.11) define the restricted least squares estimator of β. In the remainder of this section, the restricted least squares estimator will be denoted by $\hat{\beta}^{H}$. It can be shown that $\hat{\beta}^{H}$ is the linear, minimum variance, unbiased estimator of β if the hypothesis defined by (6.4) is true.

Before developing a test of the general linear hypothesis, a simpler test for the significance of a single ridge regression coefficient is considered. In this case, the hypothesis becomes:

$$H_1: \quad \beta_i = 0 \quad . \tag{6.12}$$

The hypothesis defined by (6.12) is equivalent to (6.4) if A is the 1 x p matrix:

$$A = (0 \ldots 0 \ 1 \ 0 \ldots 0) \tag{6.13}$$

whose elements are all zero except the i'th and:

$$\rho = 0 \quad . \tag{6.14}$$

It should be noted that since:

$$\alpha = P'\beta \quad , \tag{6.15}$$

H_1 implies that:

$$T_i'\alpha = 0 \tag{6.16}$$

where T_i' represents the i'th row of the orthogonal matrix P which diagonalizes the $X'X$ matrix.

Obenchain (1977) derived a significance test for the i'th

component of the generalized ridge estimator in a similar fashion as the significance test for the ordinary least squares estimator is constructed. The t-statistic for the ordinary least squares estimate of β_i is given by:

$$t_i = (\hat{\beta}_i - 0)/(s^2(X'X)^{-1}_{ii})^{\frac{1}{2}} \tag{6.17}$$

where s^2 is an unbiased estimator of σ^2 and $(X'X)^{-1}_{ii}$ represents the (i,i)'th component of the inverse of the $X'X$ matrix. s^2 in (6.17) is defined by:

$$s^2 = \frac{Y'(I_n - X(X'X)^{-1}X')Y}{n - p - 1} \quad . \tag{6.18}$$

It should be noted that the numerator of (6.17) represents difference between the ordinary least squares estimator of β_i and it's expected value under H_1 while the denominator is an unbiased estimator of the standard deviation of the numerator. The test of the hypothesis H_1 is carried out by comparing the computed value of (6.17) with a Student's-t variate having (n - p - 1) degrees of freedom.

Suppose that e_i represents the expected value of $\hat{\beta}^*_i(K)$ assumming that the hypothesis defined by (6.12) is true. Obenchain (1977) proposed a statistic of the form:

$$t^*_i = (\hat{\beta}^*_i(K) - e_i)/(s^2W)^{\frac{1}{2}} \tag{6.19}$$

to test for the significance of the i'th component of the ridge estimator. W in (6.19) is a scaling factor which makes s^2W an unbiased estimator of the variance of the numerator. Obenchain (1977) observed that the expected value of $\hat{\beta}^*_i(K)$ under the hypothesis

defined by (6.12) is usually unknown. e_i is a known quantity only in special cases such as when all the δ_i's are equal. If the value of e_i is a known quantity, W becomes the normalizing constant which makes s^2W an unbiased estimator of the variance of $\hat{\beta}_i^*(K)$.

In order to evaluate (6.19) whenever the expected value of $\hat{\beta}_i^*(K)$ under H_1 is unknown, Obenchain (1977) suggested that an estimator of e_i be substituted into (6.19). It was noted above that $\hat{\beta}^H$ is the linear, minimum variance, unbiased estimator of β which satisfies the general linear hypothesis defined by (6.4). Obenchain (1977) suggested estimating e_i by means of:

$$\hat{e}_i = T_i'\Delta P'\hat{\beta}^H \tag{6.20}$$

where $\hat{\beta}^H$ is the restricted least squares estimator which satisfies conditions (6.13) and (6.14).

Obenchain (1977) utilized the definitions of $\hat{e}_i$ and $\hat{t}_i^*$ provided above to prove the following theorem:

<u>Theorem 6.1</u>: Assume that $\delta_1, \delta_2, \ldots, \delta_p$ are all positive. The exact t-statistic for the i'th component of the generalized ridge estimator is identical to the ordinary least squares t-statistic defined by (6.17) whenever e_i is estimated by $\hat{e}_i$.

<u>Proof</u>: Substituting $\hat{e}_i$ into the numerator of t_i^* gives:

$$b_i^* - \hat{e}_i = T_i'\Delta\hat{\alpha} - T_i'\Delta P'\hat{\beta}^H$$

$$= T_i'\Delta(\hat{\alpha} - P'\hat{\beta}^H) \quad . \tag{6.21}$$

Utilizing the defintions of A, $\hat{\beta}^H$ and ρ, equation (6.21) may be expressed as:

$$\begin{aligned} b_i^* - \hat{e}_i &= T_i'\Delta P'A^*(A\hat{\beta} - \rho) \\ &= T_i'\Delta P'A^*A(\hat{\beta} - \beta) \\ &= T_i'\Delta P'(X'X)^{-1}A'(A(X'X)^{-1}A')^{-1}(\hat{\beta} - \beta) \\ &= d_i^2(\hat{\beta}_i - \beta_i) \\ &= d_i^2\hat{\beta}_i \end{aligned} \tag{6.22}$$

where:

$$d_i^2 = \frac{T_i'\Delta\Lambda^{-1}T_i}{(X'X)_{ii}^{-1}} \quad . \tag{6.23}$$

The variance of (6.22) is given by:

$$\begin{aligned} \mathrm{Var}(d_i^2\hat{\beta}_i) &= d_i^4\ \mathrm{Var}(\hat{\beta}_i) \\ &= \sigma^2 d_i^4 (X'X)_{ii}^{-1} \quad . \end{aligned} \tag{6.24}$$

Therefore, the exact t-statistic for $\hat{\beta}_i^*(K)$ is:

$$\begin{aligned} t_i^* &= \frac{d_i^2\hat{\beta}_i}{(s^2 d_i^4 (X'X)_{ii}^{-1})^{-\frac{1}{2}}} \\ &= t_i \end{aligned} \tag{6.25}$$

provided that $\delta_i > 0$. Since it was assumed that all of the δ_i's are positive, it follows that (6.25) is defined and that the generalized ridge estimator has the same t-statistic as the ordinary least squares estimator.

Several comments regarding the exact t-statistic for the ridge estimator were provided by Obenchain (1977). The t-statistic for

the ordinary least squares estimate of β_i has the same sign as the estimate. Obenchain (1977) observed that this is not necessarily true for the t-statistic defined by (6.19). Obenchain (1977) also noted that a t-statistic corresponding to (6.17) could have been constructed according to:

$$t_i^* = (\hat{\beta}_i^*(K) - 0)/(\tilde{s}^2 T_i' \Lambda^{-1} T_i)^{\frac{1}{2}} \qquad (6.26)$$

where $\tilde{s}^2$ is the residual mean sum of squares using $\hat{\beta}^*(K)$ as an estimator of β instead of $\hat{\beta}$. Obenchain (1977) observed that the expected value of $\hat{\beta}^*(K)$ is generally unknown and that $\tilde{s}^2$ tends to over-estimate σ^2. Therefore, Obenchain (1977) argued that (6.26) would not be a very useful formulation of the t-statistic for $\hat{\beta}_i^*(K)$.

Obenchain (1977) developed a test of the general linear hypothesis defined by (6.4) in the same manner as he derived the significance test for a single component of the generalized least squares estimator. It is a common statistical result that under the general linear hypothesis, the quadratic form:

$$F = \frac{(\hat{\beta} - \beta)'A'(A(X'X)^{-1}A')^{-1}A(\hat{\beta} - \beta)}{rs^2}$$

$$= \frac{(A\hat{\beta} - \rho)'(A(X'X)^{-1}A')^{-1}(A\hat{\beta} - \rho)}{rs^2} \qquad (6.27)$$

has a F distribution with r and $(n - p - 1)$ degrees of freedom. Equation (6.27) may be rewritten as:

$$F = \frac{u'Q^{-1}u}{r} \qquad (6.28)$$

where:

$$u = A\hat{\beta} - \rho \tag{6.29}$$
$$Q = s^2 A(X'X)^{-1}A'$$

The vector u represents the difference between $A\hat{\beta}$ and the expected value of $A\hat{\beta}$ under the general linear hypothesis. Q is an unbiased estimator of the variance of u under the hypothesis.

It was observed by Obenchain (1977) that an F-test for the generalized ridge estimator can be formed by replacing u with:

$$u^* = A\hat{\beta}^*(K) - AE(\hat{\beta}^*(K)) \tag{6.30}$$

and Q with an unbiased estimator of the variance of u^* in (6.28). An estimator of the expected value of $\hat{\beta}^*(K)$ under the general linear hypothesis can be constructed in the same manner as $\hat{e}_i$ was formed. The resultant estimator of the expected value of $A\hat{\beta}^*(K)$ becomes:

$$\begin{aligned} AP\Delta P'\hat{\beta}^H &= AP\Delta P'(\hat{\beta} - A^*(A\hat{\beta} - \rho)) \\ &= AP\Delta P'(\hat{\beta} - A^*A\hat{\beta} + A^*\rho) \end{aligned} \tag{6.31}$$

Substituting (6.31) into (6.30) gives:

$$\begin{aligned} u^* &= AP\Delta P'\hat{\beta} - AP\Delta P'(\hat{\beta} - A^*A\hat{\beta} + A^*\rho) \\ &= AP\Delta P'A^*(A\hat{\beta} - \rho) \\ &= Ru \end{aligned} \tag{6.32}$$

where:

$$R = AP\Delta P'A^* \tag{6.33}$$

Since the variance of u^* satisfies:

$$Var(u^*) = R\ Var(u)\ R' \tag{6.34}$$

Obenchain (1977) pointed out that an unbiased estimator of the variance of u^* may be formed according to:

$$Q^* = RQR'$$

$$= s^2 R(A(X'X)^{-1}A')^{-1}R' \quad . \tag{6.35}$$

Obenchain (1977) utilized (6.28), (6.32) and (6.35) to prove the following theorem:

Theorem 6.2: If the diagonal elements of Δ are all positive, the F-statistic under the general linear hypothesis defined by (6.4) for the generalized least squares estimator $\hat{\beta}^*(K)$:

$$F^* = \frac{u^{*\prime}Q^{-1}u^*}{r} \tag{6.36}$$

is identical to the ordinary least squares statistic.

Proof: A sufficient condition for F^* to equal the F-statistic defined by (6.28) is that:

$$R = AP\Delta P'A^* \tag{6.37}$$

be invertible. R will be invertible if each of the diagonal elements of Δ are positive. In this case,

$$\begin{aligned} F^* &= \frac{u^{*\prime}Q^{-1}u^*}{r} \\ &= \frac{u'R'(RQR')^{-1}Ru}{r} \\ &= \frac{u'Q^{-1}u}{r} \\ &= F \quad . \end{aligned} \tag{6.38}$$

It follows from (6.38) that the F-statistic for the generalized ridge estimator is equivalent to the F-statistic for the least squares estimator.

It was mentioned earlier that Marquardt and Snee (1975) seemed

to suggest that it is possible to construct confidence intervals centered at the ridge estimates that are shorter than the corresponding confidence intervals which are centered at the ordinary least squares estimates. Obenchain (1977) utilized Theorem 6.2 to argue that this is not true. He noted that Theorem 6.2 implies that the test of the hypothesis:

$$H_1: \ \beta_i = b_i \tag{6.39}$$

using the t-statistic for the ridge estimator is equivalent to the corresponding test for the ordinary least squares estimator. Since the $100(1-\alpha)\%$ confidence interval centered at $\hat{\beta}_i$ includes all the values of β_i for which H_1 can be accepted at level α, it follows that the t-statistics for $\hat{\beta}_i$ and $\hat{\beta}_i^*(K)$ define the same confidence intervals centered at $\hat{\beta}_i$. These confidence intervals have the property that the likelihood function for β_i is equal at both ends of any interval. Obenchain (1977) argued that it would be feasible to construct confidence intervals for $\hat{\beta}_i^*(K)$ which are centered at points other than $\hat{\beta}_i$. However, he pointed out that the likelihood function would have different values at the ends of any such interval. Obenchain (1977) noted that these intervals must be longer than the symmetric confidence intervals. Therefore, Obenchain (1977) agrued that no shifted confidence interval can be shorter than the corresponding interval centered at the ordinary least squares estimate.

McCabe (1978) and Obenchain (1977) introduced a criterion for comparing the ridge estimator with the ordinary least squares esti-

mator. The criterion is based upon the central F-test which was described above. Obenchain (1977) defined the associated probability $AP(\hat{\beta}^*(K))$ of a ridge estimator to be the percentage point of the central F distribution having p and (n - p - 1) degrees of freedom which satisfies:

$$(\hat{\beta}^*(K) - \hat{\beta})'X'X(\hat{\beta}^*(K) - \hat{\beta}) = ps^2F(p,n-p-1,\alpha) \quad . \quad (6.40)$$

$AP(\hat{\beta}(K))$ corresponds to the probability that $\hat{\beta}$ is further away from β than $\hat{\beta}^*(K)$. It follows from the definition of the associated probability that $AP(\hat{\beta})$ equals one and $AP(0)$ is the observed significance level of $\hat{\beta}$.

Obenchain (1977) observed that the associated probability of the ridge estimator is a random variable. $AP(\hat{\beta}^*(K))$ depends upon $\hat{\beta}, \hat{\beta}^*(K)$ and s^2 which are all stochastic. Obenchain (1977) formulated the distribution of the associated probability of $\hat{\beta}^*(K)$ according to:

$$\begin{aligned}
\tilde{F} &= \frac{(\hat{\beta}^*(K) - \hat{\beta})'(X'X)(\hat{\beta}^*(K) - \hat{\beta})}{ps^2} \\
&= \frac{(\hat{\alpha}^*(K) - \hat{\alpha})'\Lambda(\hat{\alpha}^*(K) - \hat{\alpha})}{ps^2} \\
&= \frac{\hat{\alpha}'(\Delta - I_p)'\Lambda(\Delta - I_p)\hat{\alpha}}{ps^2} \\
&= \frac{1}{ps^2} z'(\Delta - I_p)'(\Delta - I_p)z \\
&= \frac{1}{ps^2} \sum_{i=1}^{p} (\delta_i - 1)^2 z_i^2 \qquad (6.41)
\end{aligned}$$

where:

$$z = \Lambda^{\frac{1}{2}}\hat{\alpha} \quad . \qquad (6.42)$$

Obenchain (1977) observed that z_i^2/s^2 is a noncentral F-variate since Z and s^2 are independent. The noncentrality parameter for z_i^2/s^2 is equal to $(\alpha_i^2\lambda_i)/\sigma^2$. Therefore, Obenchain (1977) concluded that F is a weighted sum of noncentral F-variates.

Obenchain (1977) recommended the use of the associated probability criterion in conjunction with the ridge trace. He suggested that estimates of the coefficients be choosen by plotting different ridge estimates and their associated probabilities against Vinod's (1976b) m scale. In this way, the associated probability criterion can be used to control the amount of shrinkage which results from the use of the ridge estimator. Obenchain (1977) illustrated the associated probability criterion with a six factor problem. McCabe (1978) provided an extensive review of the properties of the associated probability criterion. He investigated it's use with other biased estimators besides the ridge estimator.

Obenchain (1975a) derived a test of the composite hypothesis:

$$H_1: \lambda_1\alpha_1^2 = \lambda_2\alpha_2^2 = \ldots = \lambda_p\alpha_p^2 \quad . \qquad (6.43)$$

He called the hypothesis defined by (6.43) the shrunken hypothesis. Obenchain (1975) argued that this hypothesis would be useful in determining when to apply ridge regression.

Consider the mean squared error function for the i'th component

of $\hat{\alpha}^*(K)$:

$$\begin{aligned} MSE(\hat{\alpha}_i^*(K)) &= E(\delta_i\hat{\alpha}_i - \alpha_i)^2 \\ &= Var(\delta_i\hat{\alpha}_i) + (\delta_i - 1)^2\alpha_i^2 \\ &= \sigma^2\delta_i\lambda_i^{-1} + (\delta_i - 1)^2\alpha_i^2 \quad . \end{aligned} \tag{6.44}$$

(6.44) is minimized when:

$$\delta_i = \frac{\lambda_i^2}{\sigma^2\lambda_i^{-1} + \alpha_i^2} \quad . \tag{6.45}$$

Under the shrunken hypothesis, each of the p quantities defined by (6.45) are equal to a constant δ^*. In this case, the estimator $\delta^*\hat{\alpha}$ minimizes the mean squared error function for $\hat{\alpha}^*(K)$. Therefore, the estimator:

$$\tilde{\beta}_s^* = \delta^*\hat{\beta} \tag{6.46}$$

would be the most appropriate estimator of β provided that the hypothesis defined by H_1 is true.

The likelihood ratio statistic was utilized by Obenchain (1975a) to develop a test of the hypothesis defined by (6.43). This statistic is defined according to:

$$\lambda = \frac{L(\tilde{\beta},\tilde{\sigma}|X,Y)}{L(\hat{\beta},\hat{\sigma}|X,Y)} \tag{6.47}$$

where the numerator of (6.47) represents the restricted maximum of the likelihood function for β and σ^2 under H_1. In order to evaluate (6.47), Obenchain (1975a) reformulated the ordinary least squares estimator of α according to:

$$\begin{aligned} \hat{\alpha} &= P'\hat{\beta} \\ &= P'(X'X)^{-1}X'Y \end{aligned}$$

$$= \Lambda^{-\frac{1}{2}} H\, Y \tag{6.48}$$

where H is the n x p orthogonal matrix which satisfies:

$$X = H\Lambda^{\frac{1}{2}} P \quad . \tag{6.49}$$

As a result of the standardization assumed for the observations, (6.48) may be rewritten as:

$$\hat{\alpha} = \Lambda^{-\frac{1}{2}} Y'Yr \tag{6.50}$$

where r is the vector of correlations between the dependent variable and each of the explanatory variables. The ordinary least squares estimator of σ^2 may be expressed in terms of r as:

$$\hat{\sigma}^2 = \frac{Y'Y(1 - r'r)}{n} \quad . \tag{6.51}$$

In addition, the restricted least squares estimates of α and σ^2 which satisfy (6.43) are given by:

$$\tilde{\alpha}_i = \text{sign}(r_i)\ \lambda_i^{-1} Y'Y|\bar{r}| \tag{6.52}$$

and:

$$\tilde{\sigma}_i^{\,2} = \frac{Y'Y(1 - p|\bar{r}|^2)}{n} \quad . \tag{6.53}$$

where $|\bar{r}|$ denotes the average of the absolute values of the p correlations r_i.

Substituting (6.50), (6.51), (6.52) and (6.53) into (6.47) gives:

$$\begin{aligned} -2\ln &= -2\ln\left((2\pi e\tilde{\sigma})^{-n/2} / (2\pi e\hat{\sigma})^{-n/2} \right) \\ &= -n\ln(\tilde{\sigma}/\hat{\sigma}) \\ &= -n\ln\left\{\frac{1 - p|\bar{r}|}{1 - r'r}\right\} \end{aligned}$$

$$= - n \ln \left\{ 1 + \frac{r'r - p|\bar{r}|}{1 - r'r} \right\} . \qquad (6.54)$$

It is a well known result that the statistic defined by (6.54) has an asymptotic χ^2 distribution with one degree of freedom. Therefore, the shrunken hypothesis is rejected if the value of (6.54) computed from the sampled data is larger than an appropriate point from the χ^2 distribution. Obenchain (1975a) observed that the computed value of (6.54) will be small if all the r_i's are approximately equal or $r'r$ close to zero. The latter case occurs when the multiple R^2 statistic is small.

Obenchain (1975a) suggested that it is probably not necessary to estimate δ^* if the shrunken hypothesis is not rejected. In this case, he pointed out that one should have some confidence in the relative magnitudes and signs of the ordinary least squares estimates. As a result, Obenchain (1975a) argued that it may not be wise to introduce any bias into the estimation process. Obenchain (1975a) suggested that problems for which the explanatory variables exhibit serious multicollinearity and the shrunken hypothesis is rejected should be prime candidates for ridge regression.

Obenchain (1974a) developed a small sample version of the test of the shrunken hypothesis. As a result of a large number of simulations, he concluded that the small sample test was rather conservative. Obenchain (1974a) utilized the test of hypothesis to develop several criteria for choosing generalized ridge estimates. He illustrated his procedures using one of the examples originally

studied by Hoerl and Kennard (1970a). McDonald (1975) commented upon Obenchain's (1975a) results.

Chapter 7

The Bayesian Interpretation Of Ridge Regression

Ridge estimators were developed in the previous chapters to estimate the parameter vector β for the general linear model in situations where the explanatory variables exhibit serious multicollinearity. It was assumed in the development of the estimators that the components of β were fixed unknown constants. Baysian inference provides an alternative approach to constructing estimators of β. Instead of assuming that β consists of unknown constants, β is regarded as being a random vector. A prior distribution is assumed for β. The prior distribution represents whatever knowledge of β is available before observing the dependent variable Y. A posterior distribution for β is obtained by combining the prior density function for β and the likelihood function of β given Y. The Bayes estimator of β corresponds to the expected value of the posterior distribution. Lindley and Smith (1972), Barnard (1974), and Hsiang (1975) observed that the ridge estimators can be developed as Baysian estimators.

Suppose that the conditional distribution of Y given β is normal with:

$$E(Y|\beta) = X\beta \tag{7.1}$$

and:

$$Var(Y|\beta) = \sigma^2 I_n \quad . \tag{7.2}$$

In addition, assume that β has a natural conjugate prior distribution with:

$$E(\beta) = A \tag{7.3}$$

and:

$$\mathrm{Var}(\beta) = Z \tag{7.4}$$

According to Bayes' theorem, the posterior density function for β is given by:

$$f(\beta|Y) = \frac{f(Y|\beta)f(\beta)}{f(Y)} \tag{7.5}$$

provided that $f(Y)>0$. It can be seen from (7.5) that:

$$f(\beta|Y) \propto \exp(-\tfrac{1}{2}Q) \tag{7.6}$$

where:

$$Q = (Y - X\beta)'(\sigma^2 I_n)^{-1}(Y - X\beta) + (\beta - A)'Z^{-1}(\beta - A) \quad . \tag{7.7}$$

Letting:

$$B^{-1} = \sigma^{-2}X'X + Z^{-1} \tag{7.8}$$

and:

$$b = \sigma^{-2}X'Y + A'Z^{-1} \tag{7.9}$$

Q may be rewritten as:

$$\begin{aligned} Q &= \beta'B\beta - 2b'\beta + (Y'(\sigma^2 I_n)^{-1}Y - A'Z^{-1}A) \\ &= (\beta - Bb)'B^{-1}(\beta - Bb) - b'B'b \\ &\quad + (Y'(\sigma^2 I_n)^{-1}Y + A'Z^{-1}A) \quad . \end{aligned} \tag{7.10}$$

Substituting (7.10) into (7.6) and integrating out the nuisance parameters, it can be seen that β has a normal posterior distribution with mean Bb and variance B. Therefore, the Baysian estimator of β is:

$$\begin{aligned} E(\beta|Y) &= Bb \\ &= (\sigma^{-2}X'X + Z^{-1})^{-1}(\sigma^2 X'Y + A'Z^{-1}) \quad . \end{aligned} \tag{7.11}$$

Hsiang (1975) assumed that β has a natural conjugate prior

distribution with:

$$E(\beta) = 0 \tag{7.12}$$

and:

$$Var(\beta) = \sigma_\beta^2 I_p \quad . \tag{7.13}$$

Substituting (7.12) and (7.13) into (7.11), the Baysian estimator of β becomes:

$$\begin{aligned} E(\beta|Y) &= (\sigma^{-2}X'X + \sigma_\beta^{-2}I_p)^{-1}(\sigma^{-2}X'Y) \\ &= (X'X + kI_p)^{-1}X'Y \end{aligned} \tag{7.14}$$

where:

$$k = \sigma^2/\sigma_\beta^2 \quad . \tag{7.15}$$

The estimator defined by (7.14) is equivalent to the ordinary ridge estimator of β for the the value of k which satisfies (7.15). Using this formulation of the estimator, Hsiang (1975) argued that ridge regression should only be applied if there is no prior knowledge of β which contradicts the assumption that the β_i's are identically, independently and normally distributed with mean zero and a common variance.

It should be noted that there are two extreme cases for the Baysian estimator defined by (7.14). Hsiang (1975) observed that:

$$\lim_{\sigma_\beta^2 \to +\infty} E(\beta|Y) = \hat{\sigma} \quad , \tag{7.16}$$

so that the ordinary least squares estimator of β is equivalent to the Baysian estimator when an infinite, uniform prior distribution is assumed for β. The other extreme case occures when each β_i is known to be equal to zero. In this case,

$$\lim_{\sigma_\beta^2 \to 0^+} E(\beta|Y) = 0 \qquad (7.17)$$

so that the Baysian estimator is always equal to the true value of β. In additon, Hsiang (1975) noted that k will be small if the variance of the unknown disturbances is small compared to the common variance of the parameters.

Hsiang (1975) observed that the generalized ridge estimator $\hat{\beta}^*(K)$ can also be derived as a Baysian estimator. Suppose that the prior distribution for β is the same as above except that the β_i's are no longer assumed to have a common variance σ_β^2. Instead, a separate variance $\sigma_{\beta_i}^2$ is assumed for each component of β. In this case, Z is a diagonal matrix whose i'th diagonal element is $\sigma_{\beta_i}^2$. Under these assumptions, the Baysian estimator of β becomes:

$$E(\beta|Y) = (X'X + K)^{-1}X'Y \qquad (7.18)$$

where K is a diagonal matrix whose i'th diagonal element is equal to:

$$k_i = \sigma^2/\sigma_{\beta_i}^2 \quad . \qquad (7.19)$$

Lindley and Smith (1972) provided an extensive discussion of the estimation procedures for β within a Baysian framework. They contrasted the ordinary least squares or ridge regression approaches to estimating β with the Baysian procedures. Landrum (1975) developed a Baysian approach to ridge regression which involved placing realistic bounds upon the ridge estimators and then analysing the matrix of eigenvalues for the ridge regression. Swamy, Mehta and Rappoport (1975) proposed an estimator of β along the lines of the Baysian

estimators.

Holland (1973) investigated the empirical Bayes procedures for computing values of k for the ordinary ridge estimator. Lawless and Wang (1976) developed a mechanical rule for choosing k based upon the Baysian formulation of the ridge estimator. They noted that the unconditional expectation of $\lambda_i\hat{\alpha}_i^2$ is given by:

$$\begin{aligned} E(\lambda_i\hat{\alpha}_i^2) &= E(E(\lambda_i\hat{\alpha}_i^2|\alpha_i)) \\ &= E(Var(\lambda_i^{\frac{1}{2}}\hat{\alpha}_i|\alpha_i)) + E(E(\lambda_i^{\frac{1}{2}}\hat{\alpha}_i|\alpha_i))^2 \\ &= \sigma^2 + E(\lambda_i^{\frac{1}{2}}\alpha_i)^2 \\ &= \sigma^2 + Var(\lambda_i^{\frac{1}{2}}\alpha_i) + (E(\lambda_i^{\frac{1}{2}}\alpha_i))^2 \\ &= \sigma^2 + \lambda_i\sigma_\alpha^2 \quad . \end{aligned} \tag{7.20}$$

If it is assumed that the X´X matrix is a correlation matrix, the sum of the p expectations defined by (7.20) satisfies:

$$\sum_{i=1}^{p} E(\lambda_i\hat{\alpha}_i^2) = p(\sigma^2 + \sigma_\alpha^2) \quad . \tag{7.21}$$

It follows from (7.21) that the ratio defined by (7.15) equals:

$$\begin{aligned} k &= \sigma^2/\sigma_\alpha^2 \\ &= \frac{1}{p\sigma^2}\sum_{i=1}^{p} E(\lambda_i\hat{\alpha}_i^2) - 1 \quad . \end{aligned} \tag{7.22}$$

Lawless and Wang (1976) observed that a value for k could be choosen by substituting the ordinary least squares estimates of of α and σ^2 into (7.22). In fact, they choose to use the estimator:

$$\hat{k} = \frac{1}{p\sigma^2}\sum_{i=1}^{p} \lambda_i\hat{\alpha}_i \quad . \tag{7.23}$$

Wichern and Churchill (1978) considered the estimator defined by (7.23) in their simulation study of mechanical rules for choosing the biasing parameters for ordinary ridge estimators.

In their conversations, Dwivedi and Zellner adopted a more general framework for the Baysian formulation of the ridge estimator. They assumed that the variances of both β and Y are dependent upon a random variable σ. In particular, Dwivedi and Zellner defined the variances of β and Y to be:

$$\mathrm{Var}(\beta) = \sigma^2 I_p \tag{7.24}$$

and:

$$\mathrm{Var}(Y) = \sigma^2 Z \tag{7.25}$$

where σ is a random variable. The density function for the prior distribution of σ was taken to be an inverted gamma function of the form:

$$f(\sigma) = \sigma^{-(\nu_0 + 1)} \exp(-\nu_0 c_0^2/2\sigma^2) \tag{7.26}$$

where $\nu_0 > 0$.

The joint posterior function for β and σ under Dwivedi and Zellner's assumptions can be obtained in the same manner as (7.5). In particular,

$$f(\beta,\sigma|Y) = \frac{f(Y|\beta,\sigma)f(\beta|\sigma)f(\sigma)}{f(Y)} \quad . \tag{7.27}$$

Substituting the density functions assumed above into (7.27), it follows that:

$$f(\beta,\sigma|Y) \propto \sigma^{-(n+p+\nu_0+1)} \exp\Big(-\frac{1}{2\sigma^2}(\nu_0 c_0^2 + (\beta-A)'Z^{-1}(\beta-A) + (Y-X\beta)'(Y-X\beta))\Big)$$

$$= \sigma^{-(n'+k+1)} \exp\left(-\frac{1}{2\sigma^2}\left(n'c^2 + (\beta-\bar{\beta})'(Z^{-1}-X'X)(\beta-\bar{\beta})\right)\right) \quad \ldots (7.28)$$

where:

$$\bar{\beta} = (X'X + Z^{-1})^{-1}(X'Y + A'Z^{-1})$$

$$n' = n + \nu_0 \qquad (7.29)$$

$$n'c^2 = \nu_0 c_0^2 + Y'Y + A'Z^{-1}A - \bar{\beta}(Z^{-1} + X'X)\bar{\beta} \quad .$$

Integrating the nuisance parameters out of (7.28), it can be seen that the Baysian estimator of β under Dwivedi and Zellner's assumptions is $\bar{\beta}$. Dwivedi and Zellner argued that the prior parameters A, Z, ν_0 and c_0^2 should be assigned values which are representative of the available prior knowledge of the problem. They pointed out that if:

$$A = 0 \qquad (7.30)$$

and:

$$Z^{-1} = kI_p \quad , \qquad (7.31)$$

the Baysian estimator $\bar{\beta}$ reduces to the ordinary ridge estimator.

The Baysian formulation of ridge regression described above incorporates prior information regarding β into the estimation procedures by assuming a prior distribution for the random vector β. Swindel (1976) and Fromby and Johnson (1977) considered an alternative approach to incorporating prior information into the estimation process. Instead of assuming that β is a random vector, they regarded β as a fixed vector and assumed that the prior information for β is random. In particular, Fromby and Johnson (1977) assumed that the prior information for β is represented by:

$$B = \beta + \nu \tag{7.32}$$

where B is a random vector and ν is an error term. Suppose that the error term in (7.32) satisfies:

$$E(\nu) = 0 \tag{7.33}$$

and:

$$Var(\nu) = \sigma^2 k^{-1} I_p \tag{7.34}$$

The constant k in (7.34) represents the confidence with which the prior information defined by (7.32) is held.

Fromby and Johnson (1977) utilized mixed estimation procedures due to Theil (1963) to construct the estimator:

$$\tilde{\beta}^* = (X'X + kI_p)^{-1}(X'Y + kB) \tag{7.35}$$

for β. They argued that the estimator defined by (7.35) is a compromise between the ordinary least squares estimator and the prior information defined by (7.32). In particular, Fromby and Johnson (1977) noted that:

$$\lim_{k \to 0^+} \tilde{\beta}^* = \hat{\beta} \tag{7.36}$$

and:

$$\lim_{k \to +\infty} \tilde{\beta}^* = B \tag{7.37}$$

Fromby and Johnson (1977) described a number of mean squared error properties for $\tilde{\beta}^*$. They also provided a numerical example to illustrate the estimator.

Chapter 8
A Combined Estimator

A number of biased estimators for the parameter vector β in the general linear model were put forward in the previous chapters to cope with the effects of severe ill-conditioning in the $X'X$ matrix. These biased estimators included the generalized least squares estimators, ridge estimators and stochastically shrunken estimators. Conditions were developed under which each estimator is mean squared error admissible when compared with the ordinary least-squares estimator. In practice, it can be difficult to justify the use of many of the biased estimators because of their arbitrariness. Two notable exceptions are the ordinary ridge estimator proposed by Hoerl and Kennard (1970a) and Marquardt's (1970) generalized least squares estimator.

There are probably two main reasons for preferring the ordinary ridge and generalized least squares estimators over the other biased estimators. First, these two estimators have geometrical interpretations which may be invoked to justify their use. The ordinary ridge estimator minimizes the residual sum of squares for a fixed parameter length. In comparison, the generalized least squares estimator $\hat{\beta}_r^+$ is choosen to minimize the residual sum of squares within a r-dimensional subspace of R^p which accounts for most of the variation in the $X'X$ matrix. Secondly, the ordinary ridge and generalized least squares estimators are two of the simplest biased estimators. Both depend upon a single biasing parameter which must be estimated by the analyst. In comparison, the generalized ridge estimator $\hat{\beta}^*(K)$ requires a separate biasing parameter

k_i for each component of α.

Marquardt (1970) noted that the generalized least squares and ordinary ridge estimators share many properties in common. He observed that both estimators can be superior to the ordinary least squares estimator when the $X'X$ matrix is ill-conditioned. However, Marquardt (1970) pointed out that the two estimators are most efficient in different types of situations. The generalized least squares estimator is most appropriate when some of the eigenvalues of the $X'X$ matrix are equal to zero. On the other hand, the ridge estimator is better suited to problems where all the eigenvalues of the $X'X$ matrix are nonzero but some are very small. As a result, Marquardt (1970) suggested that it might be useful to combine the properties of both estimators. He proposed the combined estimator:

$$\hat{\beta}_1^*(k,r) = (P_r \Lambda_r P_r' + kI_p)^{-1} X'Y \qquad (8.1)$$

where Λ_r and P_r are defined by (3.17) and (3.18) respectively. Marquardt (1970) suggested that the rank r be choosen for $\hat{\beta}_1^*(k,r)$ to remove the zero eigenvalues of the $X'X$ matrix and k to deflate the effects of the remaining small eigenvalues.

Some insight can be gained into the combined estimator $\hat{\beta}_1^*(k,r)$ by considering the canonical form of the estimator:

$$\begin{aligned}\hat{\alpha}_1^*(k,r) &= P'\hat{\beta}_1^*(k,r) \\ &= P'(P_r \Lambda_r P_r' + kI_p)^{-1} X'Y \\ &= (P'P_r \Lambda_r P_r' P + kI_p)^{-1} P'X'Y\end{aligned}$$

$$= \begin{bmatrix} kI_{(p-r)} & 0 \\ 0 & \Lambda_r + kI_r \end{bmatrix}^{-1} \Lambda\hat{\alpha} \quad . \tag{8.2}$$

It follows from (8.2) that the individual components of $\hat{\alpha}_1^*(k,r)$ are given by:

$$\hat{\alpha}_{i1}^*(k,r) = \begin{cases} \dfrac{\lambda_i}{k}\hat{\alpha}_i & \text{for } i = 1,2, \ldots , p-r \\ \dfrac{\lambda_i}{\lambda_i + k}\hat{\alpha}_i & \text{for } i = p-r+1, p-r+2, \ldots , p \end{cases} \quad . \tag{8.3}$$

It can be seen from (8.3) that the last r components of $\hat{\alpha}_1^*(k,r)$ correspond to the components of the canonical form of the ridge estimator.

It can be seen from the canonical form of the combined estimator that $\hat{\beta}_1^*(k,r)$ may be an inappropriate estimator for use in situations where some of the eigenvalues for the $X'X$ matrix are assumed to be zero. In contrast to the generalized least squares estimator, the combined estimator defined by (8.1) forces the estimates of all the components of α to be nonzero even when the rank of the $X'X$ matrix is assumed to be less than p. Further, for a fixed value of k, the ordinary ridge estimator shrinks the estimates of the first $(p - r)$ components of α more than the combined estimator $\hat{\alpha}_1^*(k,r)$. In fact, if k is less than any of the first $(p - r)$ λ_i's, the combined estimator inflates the corresponding ordinary least squares estimates.

In order to circumvent some of the problems mentioned above,

an alternative formulation of a combined generalized least squares and ridge estimator is considered. It is proposed that β be estimated by:

$$\hat{\beta}_2^*(k,r) = P_r(\Lambda_r + kI_r)^{-1}P_r'X'Y \quad . \tag{8.4}$$

As was the case for $\hat{\beta}_1^*(k,r)$, the two biasing parameters k and r are choosen according to Marquardt's specifications. The assigned rank r of the $X'X$ matrix is choosen to remove the effects of the eigenvalues assumed to be equal to zero from the estimation procedure and k to deflate the effects of the remaining small eigenvalues.

The combined estimator defined by (8.4) was first considered by Farebrother (1975) in the context of a comparison of the relative efficiency of the ridge estimator for estimating estimable functions of β when the rank of the $X'X$ matrix is assumed to be less than p. He compared linear transformations of the combined estimator $\hat{\beta}_2^*(k,r)$ with the corresponding linear transformations of the generalized least squares estimator of rank r. However, Farebrother (1975) did not consider $\hat{\beta}_2^*(k,r)$ as a simple estimator of the coefficient vector β. In the remainder of this chapter, the properties of the combined estimator defined by (8.4) are detailed. Conditions are developed under which $\hat{\beta}_2^*(k,r)$ is mean squared error admissible. Finally, a series of simulation experiments are presented to illustrate some of the benifits of the estimator.

The combined estimator $\hat{\beta}_2^*(k,r)$ may be expressed in it's canon-

ical form as:

$$\hat{\alpha}_2^*(k,r) = P'\hat{\beta}_2^*(k,r)$$
$$= P'P_r(\Lambda_r + kI_r)^{-1}P_r'X'Y$$
$$= P'P_r(\Lambda_r + kI_r)^{-1}P_r'P\Lambda\hat{\alpha}$$
$$= \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_r + kI_r \end{bmatrix}^{-1} \Lambda\hat{\alpha} \quad . \tag{8.5}$$

The individual components of $\hat{\alpha}_2^*(k,r)$ are given by:

$$\hat{\alpha}_{i2}^*(k,r) = \begin{cases} 0 & \text{for } i = 1,2, \ldots , p-r \\ \dfrac{\lambda_i}{\lambda_i + k}\hat{\alpha}_i & \text{for } i = p-r+1, p-r+2, \ldots , p \end{cases} \quad . \tag{8.6}$$

It can be seen from (8.6) that the estimates of the individual components of α produced by $\hat{\alpha}_2^*(k,r)$ are consistent with Marquardt's (1970) suggestions regarding the formation of a combined estimator. The first $(p - r)$ components of $\hat{\alpha}_2^*(k,r)$ which correspond to the zero eigenvalues of the $X'X$ matrix are constrained to be equal to zero. The remaining components are deflated to remove the effects of eigenvalues which are small but not assumed to be equal to zero.

The combined estimator $\hat{\beta}_2^*(k,r)$ may be rewritten as:

$$\hat{\beta}_2^*(k,r) = P_r(\Lambda_r + kI_r)^{-1}P_r'X'Y$$
$$= P_r(\Lambda_r + kI_r)^{-1}P_r'(X'X)$$
$$= V(k,r) \tag{8.7}$$

where:

$$V(k,r) = P_r(\Lambda_r + kI_r)^{-1}P_r'(X'X)$$
$$= P_r(\Lambda_r + kI_r)^{-1}P_r'P\Lambda P'$$

$$= P_r(\Lambda_r + kI_r)^{-1}P_r'(P_{(p-r)}\Lambda_{(p-r)}P_{(p-r)}' + P_r\Lambda_r P_r') \quad . \tag{8.8}$$

It follows from (8.8) that $\hat{\beta}_2^*(k,r)$ is a linear transformation of the ordinary least squares estimator with the transformation dependent upon k and r through V(k,r). The expected value of the combined estimator defined by (8.4) is given by:

$$\begin{aligned} E(\hat{\beta}_2^*(k,r)) &= V(k,r)\ E(\hat{\beta}) \\ &= P_r(\Lambda_r + kI_r)^{-1}P_r'(P_{(p-r)}\Lambda_{(p-r)}P_{(p-r)}' \\ &\qquad + P_r\Lambda_r P_r')\beta \quad . \end{aligned} \tag{8.9}$$

Therefore, $\hat{\beta}_2^*(k,r)$ is a biased estimator of β if k is not equal to zero or $\Lambda_{(p-r)}$ is a non-null matrix. In addition, $\hat{\beta}_2^*(k,r)$ is said to be conditionally unbiased relative to the constraints implied by the columns of $P_{(p-r)}$ if $\Lambda_{(p-r)}$ is a null matrix and k equals zero.

The variance-covariance matrix for $\hat{\beta}_2^*(k,r)$ is given by:

$$\begin{aligned} Var(\hat{\beta}_2^*(k,r)) &= E((\hat{\beta}_2^*(k,r) - E(\hat{\beta}_2^*(k,r)))(\hat{\beta}_2^*(k,r) \\ &\qquad - E(\hat{\beta}_2^*(k,r)))') \\ &= V(k,r)\ E((\hat{\beta} - \beta)(\hat{\beta} - \beta)')\ V(k,r) \\ &= \sigma^2\ P_r(\Lambda_r + kI_r)^{-1}P_r'(X'X)P_r(\Lambda_r + kI_r)^{-1}P_r' \\ &= \sigma^2\ P_r(I_r + k\Lambda_r^{-1})^{-2}\ \Lambda_r^{-1}P_r' \quad . \end{aligned} \tag{8.10}$$

The canonical form of the combined estimator defined by (8.4) may be utilized to develop the mean squared error function for $\hat{\beta}_2^*(k,r)$. Suppose that the individual components of the canonical form of the estimator are represented by:

$$\hat{\alpha}^*_{i2}(k,r) = c_i \hat{\alpha}_i \tag{8.11}$$

where:

$$c_i = \begin{cases} 0 & \text{for } i = 1,2, \ldots , p-r \\ \dfrac{\lambda_i}{\lambda_i + k} & \text{for } i = p-r+1, p-r+2, \ldots , p \end{cases} \tag{8.12}$$

The mean squared error function for $\hat{\beta}^*_2(k,r)$ is given by:

$$\begin{aligned} MSE(\hat{\beta}^*_2(k,r)) &= E((\hat{\beta}^*_2(k,r) - \beta)'(\hat{\beta}^*_2(k,r) - \beta)) \\ &= E((\hat{\alpha}^*_2(k,r) - \alpha)'(\hat{\alpha}^*_2(k,r) - \alpha)) \\ &= \sum_{i=1}^{p} Var(\hat{\alpha}^*_{i2}(k,r)) + \sum_{i=1}^{p} (E(\hat{\alpha}^*_{i2}(k,r)) - \alpha_i)^2 \\ &= \sum_{i=1}^{p} c_i^2 Var(\hat{\alpha}_i) + \sum_{i-1}^{p} (c_i - 1)^2 \alpha_i^2 \\ &= \sigma^2 \sum_{i=p-r+1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + \sum_{i=1}^{p-r} \alpha_i^2 \\ &\quad + k^2 \sum_{i=p-r-1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2} \quad . \end{aligned} \tag{8.13}$$

A sufficient mean squared error admissibility condition for $\hat{\beta}^*_2(k,r)$ may be derived by fixing a rank for the $X'X$ matrix and determining an appropriate condition for k. Suppose that the $X'X$ matrix is assumed to be of rank r. Consider an arbitrary sequence of r constants μ_i which satisfy:

$$\begin{aligned} \mu &= \sum_{i=p-r+1}^{p} \mu_i \\ &= \sum_{i=1}^{p-r} \alpha_i^2 - \sigma^2 \sum_{i=1}^{p-r} \frac{1}{\lambda_i} \quad . \end{aligned} \tag{8.14}$$

For example, the r constants may be assumed to be equal so that:

$$\mu_i = \frac{1}{r}\mu$$

$$= \frac{1}{r}\left\{\sum_{i=1}^{p-r} \alpha_i^2 - \sigma^2 \sum_{i=1}^{p-r} \frac{1}{\lambda_i}\right\} \quad . \tag{8.15}$$

The r constants defined by (8.14) can be utilized to prove the following theorem:

<u>Theorem 8.1</u>: The combined estimator $\hat{\beta}_2^*(k,r)$ is mean squared error admissible for a positive value k and fixed rank r if the inequality:

$$a_i k^2 + 2b_i k + \lambda_i^2 \mu_i < 0 \tag{8.16}$$

where:

$$a_i = \alpha_i^2 - \lambda_i^{-1}\sigma^2 + \mu_i \tag{8.17}$$

$$b_i = \mu_i \lambda_i - \sigma^2$$

is satisfied for $i = p-r+1, p-r+2, \ldots, p$.

<u>Proof</u>: By definition $\hat{\beta}_2^*(k,r)$ is mean squared error admissible if:

$$MSE(\hat{\beta}_2^*(k,r)) = \sum_{i=1}^{p-r} \alpha_i^2 + \sum_{i=p-r+1}^{p} (c_i - 1)^2 \alpha_i^2 + \sigma^2 \sum_{i=p-r+1}^{p} \frac{c_i^2}{\lambda_i}$$

$$< \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}$$

$$= MSE(\hat{\beta}) \quad . \tag{8.18}$$

Substituting the constants μ_i into (8.18), $\hat{\beta}_2^*(k,r)$ is mean squared error admissible if:

$$\sum_{i=p-r+1}^{p} (c_i - 1)^2 \alpha_i^2 + \sigma^2 \sum_{i=p-r+1}^{p} \frac{c_i^2}{\lambda_i} + \sum_{i=p-r+1}^{p} \mu_i$$

$$< \sigma^2 \sum_{i=p-r+1}^{p} \frac{1}{\lambda_i} \tag{8.19}$$

Inequality (8.19) is true if each combination of components in the summations satisfies:

$$(c_i - 1)^2 \alpha_i^2 + \sigma^2 \lambda_i^{-1} c_i^2 + \mu_i < \sigma^2 \lambda_i^{-1} \tag{8.20}$$

or:

$$(c_i - 1)\alpha_i^2 + \sigma^2 \lambda_i^{-1}(c_i + 1) + \frac{\mu_i}{c_i - 1} > 0 \quad . \tag{8.21}$$

Substituting (8.12) into (8.21) gives:

$$k^2 \alpha_i^2 - \sigma^2 \lambda_i^{-1}(2\lambda_i k + k^2) + (\lambda_i + k)^2 \mu_i < 0 \quad . \tag{8.22}$$

It can be seen from (8.22) that the combined estimator $\hat{\beta}_2^*(k,r)$ is mean squared error admissible if the biasing parameter k satisfies:

$$(\alpha_i^2 - \sigma^2 \lambda_i^{-1} + \mu_i)k^2 + 2(\mu_i \lambda_i - \sigma^2)k + \lambda_i^2 \mu_i$$

$$< 0 \tag{8.23}$$

for $i = p-r+1, p-r+2, \ldots, p$.

Theorem 8.1 illustrates that it is feasible to construct mean squared error admissible combined estimators of the form (8.4). Although inequality (8.16) provides a sufficient condition under which $\hat{\beta}_2^*(k,r)$ is admissible, it does not define a practical criterion for choosing k given a fixed rank r. A more useful condition may be obtained by considering the context in which it is proposed that the combined estimator be used. Marquardt (1970) recommended that the assigned rank r be choosen to remove the effects of the eigenvalues

for the $X'X$ matrix which are assumed to be equal to zero from the estimation process. The biasing parameter k is then choosen to deflate the effects of the remaining small eigenvalues. In accordance with this recommendation, a mean squared error admissible combined estimator $\hat{\beta}_2^*(k,r)$ may be obtained by choosing the rank r such that $\hat{\beta}_r^+$ is admissible and the biasing parameter k so that the mean squared error of $\hat{\beta}_2^*(k,r)$ is less than or equal to that of the generalized least squares estimator $\hat{\beta}_r^+$.

A sufficient mean squared error admissibility condition for $\hat{\beta}_2^*(k,r)$ based upon a comparison of the mean squared error functions for the combined and generalized least squares estimators is given by the theorem:

<u>Theorem 8.2:</u> Suppose that the assigned rank r for the $X'X$ matrix is choosen so that inequality (3.48) is satisfied. The combined estimator $\hat{\beta}_2^*(k,r)$ is mean squared error admissible if:

$$0 < k < \frac{\sigma^2}{\alpha_i^2} \tag{8.24}$$

for $i = p-r+1, p-r+2, \ldots, p$.

<u>Proof:</u> The mean squared error functions for $\hat{\beta}_r^+$ and $\hat{\beta}_2^*(k,r)$ are given by (3.49) and (8.13) respectively. The combined estimator will have a smaller mean squared error than the generalized least squares estimator if:

$$\mathrm{MSE}(\hat{\beta}_2^*(k,r)) = \sigma^2 \sum_{i=p-r+1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + \sum_{i=1}^{p-r} \alpha_i^2$$

$$+ k^2 \sum_{i=p-r+1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2}$$

$$< \sigma^2 \sum_{i=p-r+1}^{p} \frac{1}{\lambda_i} + \sum_{i=1}^{p-r} \alpha_i^2$$

$$= \mathrm{MSE}(\hat{\beta}_r^+) \qquad (8.25)$$

Inequality (8.25) is satisfied if:

$$\sigma^2 \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \frac{\alpha_i^2}{(\lambda_i + k)^2} < \sigma^2 \frac{1}{\lambda_i} \qquad (8.26)$$

for $i = p-r+1, p-r+2, \ldots, p$. In the proof of Theorem 2.4, it was shown that inequality (8.26) is satisfied if condition (8.24) holds Therefore, the combined estimator will be mean squared error admissible since the mean squared error function for $\hat{\beta}_2^*(k,r)$ is less than the mean squared error function for the admissible generalized least squares estimator $\hat{\beta}_r^+$.

It was mentioned earlier that both the generalized least squares and ordinary ridge estimators have geometrical interpretations. These interpretations may be invoked to justify the use of the estimators. A geometrical interpretation of the combined estimator $\hat{\beta}_2^*(k,r)$ is provided by:

<u>Theorem 8.3</u>: The combined estimator $\hat{\beta}_2^*(k,r)$ minimizes the residual sum of squares function $\Phi(b)$ within the

r-dimensional subspace spanned by P_r for a fixed value of the Euclidean norm.

Proof: Adopting the notation from Chapter 3, let:

$$\xi = XP_r \tag{8.27}$$

denote the projection of the points of X onto the eigenvectors which form the last r columns of P. Suppose that β^* denotes the projection of β onto the eigenvector coordinates. The sum of squares of the residuals functions for any estimator b^* of β^* is given by:

$$\Phi(b^*) = (Y - \xi b^*)'(Y - \xi b^*) \tag{8.28}$$

Suppose that the estimator b^* is choosen to minimize (8.28) subject to the constraint that the Euclidean norm of b^* equals a constant d^2. The resultant estimator corresponds to the vector b^* which minimizes the Lagrangian equation:

$$\begin{aligned}\tilde{F}(b^*) &= \Phi(b^*) - k(b^{*\prime}b^* - d^2) \\ &= (Y - \xi b^*)'(Y - \xi b^*) - k(b^{*\prime}b^* - d^2)\end{aligned} \tag{8.29}$$

$\tilde{F}(b^*)$ is minimized when:

$$\begin{aligned}\frac{d}{db^*}\tilde{F}(b^*) &= -2\xi'Y + 2\xi'\xi b^* + 2kb^* \\ &= 2(\xi'\xi + kI_r)b^* - 2\xi'Y \\ &= 0\end{aligned} \tag{8.30}$$

or:

$$\begin{aligned}b^* &= (\xi'\xi + kI_r)^{-1}\xi'Y \\ &= (P_r'X'XP_r + kI_r)^{-1}P_r'X'Y \\ &= (\Lambda_r + kI_r)^{-1}P_r'X'Y\end{aligned} \tag{8.31}$$

The solution for (8.31) may be expressed in terms of the original coordinates as:

$$b = P_r b^*$$
$$= P_r(\Lambda_r + kI_r)^{-1} P_r X Y \qquad (8.32)$$

It follows from equations (8.29) through (8.32) that the combined estimator $\hat{\beta}_2^*(k,r)$ minimizes the residual sum of squares for a fixed Euclidean norm when the lengths of the estimates are measured using the projections of the estimators on the eigenvector coordinates.

The results of a series of Monte Carlo simulations are presented in order to illustrate the potential usefulness of the combined estimator $\hat{\beta}_2^*(k,r)$. For the purposes of the simulations, it was assumed that the design matrices were 6 x 4 matrices. The method of singular value decomposition was employed to construct the design matrices. In particular, the design matrices were formed by:

$$X = U\Lambda D \qquad (8.33)$$

where: U was a 6 x matrix whose columns were orthogonal; Λ was a 4 x 4 diagonal matrix; and D was a 4 x 4 orthogonal matrix. It follows from (8.33) that the $X'X$ matrices were given by:

$$X'X = (U\Lambda D)'(U\Lambda D)$$
$$= D'\Lambda^2 D \qquad (8.34)$$

It can be seen from (8.34) that the eigenvalues for the $X'X$ matrices were equal to the squares of the diagonal elements of the Λ matrices.

The methodology of Dempster, Schatzoff and Wermuth (1977) was utilized to transform the $X'X$ matrices into correlation matrices.

Suppose that T is a diagonal matrix whose i'th diagonal element is the square root of the (i,i)'th element of the $X'X$ matrix defined by (8.34). In this case, the adjusted design matrix:

$$X = U\Lambda DT \quad , \tag{8.35}$$

leads to the $X'X$ matrix:

$$\begin{aligned} X'X &= (U\Lambda DT)'(U\Lambda DT) \\ &= T'D'\Lambda^2 DT \end{aligned} \tag{8.36}$$

which is a correlation matrix.

The two design matrices:

$$X_1 = \begin{bmatrix} 0.61557 & 0.11557 & 0.52425 & -0.16602 \\ 0.16557 & 0.66557 & -0.19478 & -0.40224 \\ 0.48395 & 0.04719 & 0.63604 & -0.26546 \\ -0.01862 & 0.44976 & -0.07947 & -0.64779 \\ 0.59719 & 0.03395 & 0.51518 & -0.24375 \\ 0.04976 & 0.58138 & -0.10464 & -0.51101 \end{bmatrix} \tag{8.37}$$

and:

$$X_2 = \begin{bmatrix} 0.63182 & 0.13182 & 0.49214 & -0.11720 \\ 0.21088 & 0.71088 & -0.24164 & -0.32016 \\ 0.40300 & 0.04442 & 0.69583 & -0.26295 \\ -0.06999 & 0.35930 & -0.07469 & -0.71661 \\ 0.62348 & -0.01795 & 0.44973 & -0.25304 \\ 0.01741 & 0.58812 & -0.08604 & -0.48690 \end{bmatrix} \tag{8.38}$$

were constructed using (8.35). The corresponding $X'X$ matrices were given by:

$$X_1'X_1 = \begin{bmatrix} 1.00000 & 0.24500 & 0.90221 & -0.45619 \\ 0.24500 & 1.00000 & -0.11813 & -0.89614 \\ 0.90221 & -0.11813 & 1.00000 & -0.19815 \\ -0.45619 & -0.89614 & -0.19815 & 1.00000 \end{bmatrix} \tag{8.39}$$

and:

$$X_2'X_2 = \begin{bmatrix} 1.00000 & 0.22500 & 0.82452 & -0.36361 \\ 0.22500 & 1.00000 & -0.16150 & -0.79401 \\ 0.82452 & -0.16150 & 1.00000 & -0.18166 \\ -0.36361 & -0.79401 & -0.18166 & 1.00000 \end{bmatrix} \tag{8.40}$$

The eigenvalues for the $X_1'X_1$ and $X_2'X_2$ matrices were (2.32624, 1.56719, 0.09612, 0.01046) and (2.14750, 1.55946, 0.24099, 0.05206) respectively.

Two simulation models were constructed using the design matrices defined by (8.37) and (8.38). The parameter vectors for the simulation models were taken to be:

$$\beta_1' = (0.60119, 1.10539, 0.26553, -1.91277) \quad (8.41)$$

and:

$$\beta_2' = (1.15789, 1.54238, -0.02996, -1.26172) \quad . \quad (8.42)$$

The parameters for the canonical forms of both models were given by:

$$\alpha' = (2.00, 1.00, 0.50, -0.25) \quad . \quad (8.43)$$

The four levels of σ^2: 0.005, 0.011, 0.02 and 0.08 were assumed for each model. The simulation experiments consisted of 1,000 simulations of each model at the four different levels of σ^2.

Eight different estimators were utilized to simulate the estimation of the parameter vectors β_1 and β_2. Besides the ordinary least squares estimator, three generalized least squares and four combined estimators were considered. The assigned ranks for the combined estimators were taken to be 3 and 4. Two different algorithms were utilized to choose appropriate values of the biasing parameter k_i. As a result of Theorem 8.2, the methodology of Chapter 2 may be utilized to develop rules for estimating k. Based upon the procedures described in that chapter, the algorithms:

$$k_1 = \frac{\hat{\sigma}^2}{\max(\alpha_1^2, \alpha_2^2, \ldots, \alpha_r^2)} \tag{8.44}$$

and:

$$k_2 = r\hat{\sigma}^2 / (\sum_{i=1}^{r} \alpha_i^2) \tag{8.45}$$

were employed to calculate a value of k for the combined estimator $\hat{\beta}_2^*(k,r)$. Tables 5 and 6 summarize the average values of the k_i's which were calculated in the simulation experiments.

The average simulated mean squared errors for the various estimators of β_1 and β_2 are summarized in Tables 7 and 8. It can be seen from these tables that both the generalized least squares and combined estimators provided improvements in their average simulated mean squared errors when compared with the ordinary least squares estimator. In fact, the generalized least squares estimator of rank 3 always had a smaller average mean squared error than $\hat{\beta}$. $\hat{\beta}_3^+$ produced smaller average mean squared errors than all of the other generalized least squares estimators except when σ^2 was set to 0.08. In these cases, the generalized least squares estimator of rank 2 had the smallest mean squared errors.

It should be noted that the combined estimator $\hat{\beta}_4^*(4,k_i)$ corresponds to the ordinary ridge estimator with the biasing parameter k_i. It can be seen from Tables 7 and 8 that the ordinary ridge estimators provided substantial improvements over $\hat{\beta}$ in terms of the average mean squared errors. In all cases, $\hat{\beta}_2^*(3,k_i)$ produced

Table 5 - A Comparison Of The Average Values Of The Biasing Parameters k_i For The Estimates Of β_1

Estimator	Level Of σ^2			
	0.005	0.011	0.020	0.080
$\hat{\beta}_2^*(3,k_1)$	0.00126	0.00277	0.00505	0.02033
$\hat{\beta}_2^*(3,k_2)$	0.00284	0.00618	0.01110	0.04134
$\hat{\beta}_2^*(4,k_1)$	0.00126	0.00273	0.00479	0.01502
$\hat{\beta}_2^*(4,k_2)$	0.00350	0.00717	0.01198	0.03453

Table 6 - A Comparison Of The Average Values Of The Biasing Parameters k_i For The Estimates Of β_2

Estimator	Level Of σ^2			
	0.005	0.011	0.020	0.080
$\hat{\beta}_2^*(3,k_1)$	0.00126	0.00277	0.00506	0.02067
$\hat{\beta}_2^*(3,k_2)$	0.00286	0.00629	0.01140	0.04452
$\hat{\beta}_2^*(4,k_1)$	0.00126	0.00277	0.00506	0.01992
$\hat{\beta}_2^*(4,k_2)$	0.00372	0.00804	0.01423	0.04939

Table 7 - A Comparison Of The Average Mean Squared Errors For Various Estimators Of The Parameter Vector β_1

Estimator	Level Of σ^2			
	0.005	0.011	0.020	0.080
$\hat{\beta}_1^+$	1.31467	1.31728	1.32119	1.34724
$\hat{\beta}_2^+$	0.31771	0.32396	0.33334	0.39585
$\hat{\beta}_3^+$	0.11979	0.18855	0.29167	0.97919
$\hat{\beta}$	0.52830	1.16225	2.11319	8.45277
$\hat{\beta}_2^*(3,k_1)$	0.11841	0.18242	0.27300	0.77903
$\hat{\beta}_2^*(3,k_2)$	0.11730	0.17809	0.26181	0.70834
$\hat{\beta}_2^*(4,k_1)$	0.44745	0.86470	1.42747	5.11493
$\hat{\beta}_2^*(4,k_2)$	0.37987	0.70022	1.10858	3.44913

Table 8 - A Comparison Of The Average Mean Squared Errors For Various Estimators Of The Parameter Vector 2

Estimator	Level Of σ^2			
	0.005	0.011	0.020	0.080
$\hat{\beta}_1^+$	1.31485	1.31768	1.32191	1.35015
$\hat{\beta}_2^+$	0.31791	0.32439	0.33412	0.39899
$\hat{\beta}_3^+$	0.08867	0.12007	0.16717	0.48119
$\hat{\beta}^*$	0.12080	0.26577	0.48321	1.93[illegible]6
$\hat{\beta}_2^*(3,k_1)$	0.08852	0.11930	0.16453	0.44[illegible]2
$\hat{\beta}_2^*(3,k_2)$	0.08843	0.11877	0.16276	0.42268
$\hat{\beta}_2^*(4,k_1)$	0.11681	0.24732	0.42678	1.39340
$\hat{\beta}_2^*(4,k_2)$	0.11102	0.22553	0.37490	1.13675

substantially smaller average mean squared errors than $\hat{\beta}_2^*(4,k_i)$ or the ordinary least squares estimator. In fact, the combined estimator $\hat{b}_2^*(3,k_i)$ produced the smallest average simulated mean squared errors of the eight estimators considered for both parameter vectors and all levels of σ^2 except 0.08. In the simulation runs where σ^2 was set to 0.080, the use of the generalized least squares estimator of rank 2 resulted in the smallest average mean squared errors.

It can be seen from the results of the simulation experiments that the combined estimator $\hat{\beta}_2^*(k,r)$ offers potential for significant improvements in the mean squared errors when compared to the generalized least squares and ordinary ridge estimators. In addition, it was pointed out earlier that the combined estimator provides greater flexibility than either $\hat{\beta}_r^+$ or $\hat{\beta}_k^*$. At the same time, the combined estimator retains a geometrical interpretation. As a result, it is recommended that the combined estimator $\hat{\beta}_2^*(k,r)$ be considered as an alternative to the other biased estimators mentioned in this thesis when some of the eigenvalues for the $X'X$ matrix are assumed to be equal to zero and other eigenvalues close to zero.

Appendix

A Convergence Theorem

The following convergence theorem is required in the derivation of Hocking, Speed and Lynn's (1976) iterative estimators:

Theorem A.1: The sequence defined by:

$$c_{i+1} = c_i^2/(c_i^2 + L) \qquad (A.1)$$

has three points of accumulation depending upon L and the initial value of the sequence. Suppose that:

$$c_1^* = \tfrac{1}{2} + (\tfrac{1}{4} + L)^{\frac{1}{2}} \qquad (A.2)$$

and:

$$c_2^* = \tfrac{1}{2} - (\tfrac{1}{4} - L)^{\frac{1}{2}} \quad . \qquad (A.3)$$

Assume that c^* denotes the limiting value of the sequence defined by (A.1). The possible values of c^* are given by:

1) If $L > \frac{1}{4}$, $c^* = 0$.

2) If $L \leq \frac{1}{4}$ and:

   i) $c_0 > c_2^*$, then $c^* = c_1^*$.

   ii) $c_0 < c_2^*$, then $c^* = 0$.

   iii) $c_0 = c_2^*$, then $c^* = c_2^*$.

Proof: If the sequence defined by (A.1) has any accumulation points, they must satisfy:

$$c^* = c^{*2}/(c^{*2} + L) \quad . \qquad (A.4)$$

Solving (A.4) leads to the possible accumulation points:

$$c^* = 0 \qquad (A.5)$$

or:

$$c^* = \tfrac{1}{2} \pm (\tfrac{1}{4} - L)^{\frac{1}{2}} \quad . \qquad (A.6)$$

First, consider the case where $L > \frac{1}{4}$ so that:

$$(c_i - \tfrac{1}{2})^2 - \tfrac{1}{4} + L < 0 \quad . \tag{A.7}$$

Inequality (A.7) is equivalent to:

$$c_i^2 - c_i + L > 0 \tag{A.8}$$

or:

$$c_i/(c_i^2 + L) < 1 \quad . \tag{A.9}$$

As a result of (A.1) and (A.9),

$$\begin{aligned} c_{i+1}/c_i &= c_i/(c_i^2 + L) \\ &< 1 \quad . \end{aligned} \tag{A.10}$$

It follows from (A.10) that the c_i's form a decreasing sequence. Since the c_i's are bounded from below by 0, it follows that c^* equals 0 if $L > \frac{1}{4}$.

Next, consider the case where $L \leq \frac{1}{4}$. It should be noted that if $c_0 > c_1^*$, then $c_i > c_1^*$ for all positive integers i. This can be seen by observing that if:

$$c_i > c_1^* \quad , \tag{A.11}$$

then:

$$\begin{aligned} c_{i+1}^* &= 1/(1 + L/c_i^2) \\ &> 1/(1 + L/c_1^{*2}) \\ &= c_1^* \quad . \end{aligned} \tag{A.12}$$

A similar result holds if $c_0 < c_1^*$. The ratio test provides that the sequence of c_i's defined by (A.1) is monotonically increasing if:

$$\begin{aligned} c_{i+1}/c_i &= c_i/(c_i^2 + L) \\ &> 1 \end{aligned} \tag{A.13}$$

or:

$$(c_i - \tfrac{1}{2})^2 - \tfrac{1}{4} + L < 0 \qquad (A.14)$$

In the same manner, the sequence is monotonically decreasing if:

$$(c_i - \tfrac{1}{2})^2 - \tfrac{1}{4} + L > 0 \qquad (A.15)$$

for each i. The accumulation points for the sequence when $L \leq \frac{1}{4}$ can be obtained by comparing different combinations of (A.11), (A.14) and (A.15).

REFERENCES

Allen, D.M.: The Relationship Between Variable Selection And Data Augmentation And A Method For Prediction. Technometrics 16, 125-127 (1974).

Bacon, R.W., Hausman, J.A.: The Relationship Between Ridge Regression And The Minimum Mean Square Error Estimator Of Chipman. Oxford Bulletin Of Economic And Statistics 36, 115-124 (1974).

Banerjee, K.S., Carr, R.N.: A Comment On Ridge Regression, Biased Estimation For Nonorthogonal Problems. Techometrics 13, 895-898 (1971).

Barnard, G.A.: On Ridge Regression And General Principles Of Estimation. (memeo), (1974).

Bhattacharya, P.K.: Estimating The Mean Of A Multivariable Normal Population With General Quadratic Loss Function. Annals Of The Mathematical Statistics 37, 1819-1824 (1966).

Bibby, J.: Minimum Mean Square Error Estimation, Ridge Regressionm And Some Unanswered Questions. Progress In Statistics 1, 107-121 (1972).

Bolding, J.T., Houston, S.R.: A Fortran Computer Program For Computation Of Ridge Regression Coefficients. Educational And Psychological Measurement 34, 151-152 (1974).

Brown, P.J.: Centering And Scaling In Ridge Regression. Technometrics 19, 35-36 (1977).

Brown, M.W., Rock, R.A.: The Choice Of Additive Constants In Ridge Regression. (Abstract), South African Statistical Journal 9, 83 (1975).

Brown, W.G., Beattie, B.R.: Improving Estimates Of Economic Parameters By Use Of Ridge Regression With Production Function Applications. American Journal Of Agricultural Economics, 21-32 (1975).

Chipman, J.S.: On Least Squares With Insufficient Observations. Journal Of The American Statistical Association 59, 1078-1111 (1964).

Coniffe, D., Stone, J.: A Critical View Of Ridge Regression. The Statistician 22, 181-187 (1973).

Coniffe, D., Stone, J.: A Reply To Smith And Goldstein. The Statistician 24, 67-68 (1975).

Dempster, A.P., Schatzoff, M., Wermuth, N.: A Simulation Study Of Alternatives To Ordinary Least Squares. Journal Of The American Statistical Association 72, 77-91 (1977).

Dwivedi, T.D.: Properties Of The Family Of Biased Estimators. Ph.D. Thesis, Department Of Mathematics, Clarkson College Of Technology, New York, (1973).

Dwivedi, T.D., Srivastava, V.K.: On The Minimum Mean Squared Error Estimators In A Regression Model. Communications Statistics A7(5), 487-494 (1978).

Dwivedi, T.D., Srivastava, V.K., Hall, R.L.: Finite Properties Of Ridge Estimators In Linear Regression Model. (Submitted To Technometrics), (1976).

Farebrother, R.W.: The Minimum Mean Square Error Linear Estimator And Ridge Regression. Technometrics 17, 127-128 (1975).

Farebrother, R.W.: Further Results On The Mean Square Error Of Ridge Regression. Journal Royal Statistical Society, Series B 38, 248-250 (1976).

Farebrother, R.W.: Partitioned Ridge Regression. Technometrics 20, 121-122 (1978).

Fromby, T.B., Johnson, S.R.: MSE Evaluation Of Ridge Estimators Based On Stochastic Prior Information. Communications Statistics A6(13), 1245-1258 (1977).

Furnival, G.M., Wilson, R.W.: Regression By Leaps And Bounds. Technometrics 16, 499-511 (1974).

Goldstein, M., Smith, A.F.M.: Ridge-Type Estimators For Regression Analysis, Journal Of The Royal Statistical Society, Series B 36, 284-291 (1974).

Goode, B.: Ridge Regression And Multiple Regression. Presented At The Joint National Meeting Of The Operations Research Society Of America And The Institute Of Management Sciences, (1975).

Guikey, D.K., Murphy, J.L.: Directed Ridge Regression Techniques In Cases Of Multicollinearity. Journal Of The American Statistical Association 70, 769-775 (1975).

Gunst, R.F., Mason, R.L.: Biased Estimation In Regression: An Evaluation Using Mean Squared Error. Journal Of The American Statistical Association 72, 616-627 (1977).

Hawkins, D.M.: On The Investigation Of Alternative Regressions By Principal Component Analysis. Applied Statistics 22,

275-286 (1973).

Hawkins, D.M.: Relations Between Ridge Regression And Eigenvalues Of The Augmented Correlation Matrix. Technometrics 17, 477-480 (1975).

Hemmerle, W.J.: An Explicit Solution For Generalized Ridge Regression. Technometrics 17, 309-314 (1975).

Hemerle, W.J., Brantle, T.F.: Explicit And Constrained Generalized Ridge Estimation. Technometrics 20, 109-119 (1978).

Hodges, J.L., Lehmann, E.L.: Some Applications Of The Cramér-Rao Inequality. Proceedings Of The Second Berkeley Symposium On Mathematical Statistics And Probability, 13-22 (1951).

Hoerl, A.E.: Application Of Ridge Analysis To Regression Problems. Chemical Engineering Progress 58, 54-59 (1962).

Hoerl, A.E.: Ridge Analysis. Chemical Engineering Progress Symposium, Series 60, 67-77 (1964).

Hoerl, A.E.: On Regression Analysis And Biased Estimation. (Abstract), Technometrics 10, 422-423 (1968).

Hoerl, A.E., Kennard, R.W.: Ridge Regression: Biased Estimation For Nonorthogonal Problems. Technometrics 12, 55-67 (1970a).

Hoerl, A.E., Kennard, R.W.: Ridge Regression: Applications To Nonorthogonal Problems. Technometrics 12, 69-82 (1970b).

Hoerl, A.E., Kennard, R.W.: A Note On A Power Generalization Of Ridge Regression. Technometrics 17, 269 (1975).

Hoerl, A.E., Kennard, R.W.: Ridge Regression: Iterative Estimation Of The Biasing Parameter. Communications Statistics A5(1), 77-88 (1976).

Hoerl, A.E., Kennard, R.W., Baldwin, K.F.: Ridge Regression: Some Simulations. Communications Statistics A4(2), 105-123 (1975).

Hocking, R.R.: Criteria For Selection Of A Subset Regression: Which One Should Be Used?. Technometrics 14, 967-970 (1972).

Hocking, R.R.: The Analysis And Selection Of Variables In Linear Regression. Biometrics 32, 1-49 (1976).

Hocking, R.R.: Speed, F.M., Lynn, M.J.: A Class Of Biased Estimators

In Linear Regression. Technometrics 18, 425-437 (1976).

Hollard, P.: Weighted Ridge Regression: Combining Ridge And Robust Regression Methods. NBER Working Paper No. 11, NBER Computer Research Center, Cambridge, (1973).

Hsiang, T.C.: A Baysian View On Ridge Regression. The Statistician 24, 267-268 (1975).

James, W., Stein, C.: Estimation With Quadratic Loss. Proceedings Of The Fourth Berkeley Symposium On Mathematical Statistics And Probability, 361-379 (1961).

Jeffers, J.N.R.: Two Case Studies In The Application Of Principal Component Analysis. Applied Statistics 3, 225-236 (1967).

Landrum, F.G.: A Baysian Approach To Ridge Regression. Presented at the Joint National Meeting of The Operations Research Society of America and The Institute of Management Sciences, (1975).

Lawless, J.F., Wang, P.: A Simulation Study Of Ridge And Other Regression Estimators. Communications In Statistics A5(4), 307-323 (1976).

Lindley, D.V., Smith, A.F.M.: Bayes Estimates For The Linear Model. Journal Of The Royal Statistical Society, Series B 34, 1-18 (1972).

Lowerre, J.M.: On The Mean Square Error Of Parameter Estimates For Some Biased Estimators. Technometrics 16, 461-464 (1974).

Malinvaud, E.: Statistical Methods Of Econometrics. Chicago: Rand, McNally And Company 1966.

Marquardt, D.W.: An Algorithm For Least Squares Estimation Of Nonlinear Parameters. Journal Of The Society Of Industrial Applied Mathematics 2, 431-441 (1963).

Marquardt, D.W.: Generalized Inverses, Ridge Regression, Biased Linear Estimation And Nonlinear Estimation, Technometrics 12, 591-611 (1970).

Marquardt, D.W., Snee, R.N.: Ridge Regression In Practice. The American Statistician 29, 3-20 (1975).

Mayer, L.S.: On Equivalence Classes Of Biased Estimators. Unpublished Manuscript, (*).

Mayer, L.S., Willke, T.A.: On Biased Estimation In Linear Models. Technometrics 15, 497-508 (1973).

McCabe, G.P.: Evaluation Of Regression Coefficient Estimates Using α-Acceptability. Technometrics 20, 131-140 (1978).

McDonald, G.C.: Discussion. Technometrics 17, 443-445 (1975).

McDonald, G.C., Galarneau, D.I.: A Monte Carlo Evaluation Of Some Ridge-Type Estimators. Journal Of The American Statistical Association 70, 407-416 (1975).

McDonald, G.C., Schwing, R.C.: Instabilities Of Regression Estimates Relating Air Pollution To Mortality. Technometrics 15, 463-481 (1973).

Meeter, D.A.: On A Theorem Used In Nonlinear Least Squares. SIAM Journal Of Applied Mathematics 14, 1176-1179 (1966).

National Bureau Of Economic Research: Troll Experimental Programs: Robust And Ridge Regression. Computer Research Center For Economics And Management Sciences, (1975).

Needler, J.A.: Discussion On The Paper By Professor Lindley And Dr. Smith. Journal Of The Royal Statistical Society, Series B 34, 18-20 (1972).

Newhouse, J.P., Oman, S.D.: An Evaluation Of Ridge Estimators. Report No. R-716-RR, Rand Corporation, Santa Monica, California.

Obenchain, R.L.: Ridge Analysis Following A Preliminary Test Of The Shrunken Hypothesis. Technometrics 17, 431-442 (1975a).

Obenchain, R.L.: Residual Optimality: Ordinary Vs. Weighted Vs. Biased Least Squares. Journal Of The American Statistical Association 70, 375-379 (1975b).

Obenchain, R.L.: Classical F-Tests And Confidence Regions For Ridge Regression. Technometrics 19, 429-439 (1977).

Obenchain, R.L.: Data Analytic Displays For Ridge Regression. (Abstract). Proceedings Of The Eleventh Annual Symposium On The Interface (Computer Science And Statistics), (1978).

Obenchain, R.L., Vinod, H.D.: Estimates Of Partial Derivatives From Ridge Regression On Ill-Conditioned Data. Presented at NBER-NSF Seminar on Baysian Inference in Econometrics, Michigan, (1974).

Press, S.J.: Applied Multivariate Analysis. New York: Holt, Rinehart And Winston, Inc. 1972.

Pukelsheim, F.: Equality Of Two Blues And Ridge-Type Estimates. Communications Statistics A6(7), 603-610 (1977).

Rao, C.R.: Linear Statistical Inference And Its Applications (Second Edition). New York: John Wiley And Sons, Inc. 1965.

Royden, H.L.: Real Analysis (Second Edition). London: Collier-Macmillan Ltd. 1968.

Sclove, S.L.: Improved Estimators For Coefficients In Linear Regression. Journal Of The American Statistical Association 60, 234-246 (1968).

Searle, S.R.: Linear Models. New York: John Wiley And Sons Inc. (1971).

Smith, A.F.M, Goldstein, M.: Ridge Regression: Some Comments On A Paper By Coniffe And Stone. The Statistician 24, 61-66 (1975).

Sommers, R.W.: Sound Application Of Regression Analysis In Chemical Engineering. Presented at the A.I.Ch.E. Symposium on Avoiding Pitfalls in Engineering Applications Of Statistical Methods. Memphis, Tenn. (1964).

Srivastava, V.K.: Estimation Of Large Econometric Models. Journal Of Statistical Reasearch, (1975).

Stein, C.: Inadmissibility Of The Usual Estimator For The Mean Of A Multivariate Normal Distribution. Proceedings Of The Third Berkeley Symposium On Mathematical Statistics And Probability, 197-206 (1956).

Stein, C.: Multiple Regression. Contributions To Probability And Statistics, Essays In Honour Of Harold Hotelling, Stanford University Press, 424-443 (1960).

Strawderman, W.E.: Minimax Adaptive Generalized Ridge Regression Estimators, Journal Of The American Statistical Association 73, 77-91 (1978).

Swamy, P.A.V.B.: Criteria, Constraints And Multicollinearity In Random Coefficient Regression Models. Annals Of Economic And

Social Measurement 2, 429-450 (1973).

Swamy, P.A.V.B., Mehta, J.S., Rappoport, P.N.: Relative Efficiencies Of A Competitor Of Hoerl And Kennard's Ridge Regression Estimator. Special Studies Paper, Division of Research and Statistics, Federal Reserve Board, Washington D.C., (1975).

Swindel, B.F.: Instability Of Regression Coefficients Illustrated. The American Statistician 28, 63-65 (1974).

Swindel, B.F.: Good Ridge Estimators Based Upon Prior Information. Communications Statistics A5(11), 1245-1258 (1976).

Swindel, B.F., Chapman, D.D.: Good Ridge Estimators. (Abstract), Biometrics 30, 385-386.

Theobald, C.M.: Generalizations Of Mean Square Error Applied To Ridge Regression. Journal Of The Royal Statistical Society, Series B 36, 103-106 (1974).

Vinod, H.D.: Ridge Estimation Of A Trans-Log Production Function. Proceedings of the 1974 Annual Meetings of the Business And Economics Section of the American Statistical Association, Washington, D.C., (1974).

Vinod, H.D.: A Ridge Estimator Whose MSE Dominates OLS. (Abstract), The Institute Of Mathematical Statistics Bulletin 5, 189 (1976a).

Vinod, H.D.: Application Of New Ridge Regression Methods To A Study Of Bell System Scale Economics. Journal of the American Statistical Association 76, 835-841 (1976b).

Vinod, H.D.: Canonical Ridge And Econometrics Of Joint Production. Journal Of Econometrics 4, 147-166 (1976c).

Whichern, D.W., Churchill, G.A.: A Comparison Of Ridge Estimators. Technometrics 20, 301-310 (1978).