WEIGHTED RIDGE REGRESSION

by

JOHN WARREN

A thesis submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

DEPARTMENT OF STATISTICS

RALEIGH

1981

APPROVED BY:

[Signatures]

Co-chairman of Advisory Committee
ABSTRACT

WARREN, JOHN. Weighted Ridge Regression. (Under the direction of F. G. GIESBRECHT and A. R. MANSON.)

The biased estimation technique known as ridge regression assumes that all the data points have the same importance or weighting. The technique of weighted ridge regression allows the points to carry different weights and such an action has a profound effect on the variance and bias-squared properties of the resulting estimators. Many of the properties of this weighted estimator are compared and contrasted with those of ridge regression and ordinary least squares. The weightings for which weighted ridge estimation is an improvement over ridge regression are derived, and the extension of the weighting technique to specific cases of practical interest are examined.
BIOGRAPHY


He spent a year as a research technician with the Cape Asbestos Group before continuing his studies at Sir John Cass College of the University of London. He was awarded a Bachelor of Science degree in Physics and Mathematics in 1968.

He entered North Carolina State University in 1968 and received the last Master of Experimental Statistics degree awarded by the university in 1970. He was appointed Visiting Instructor of Statistics in 1975 in the Department of Statistics, at North Carolina State University. In June 1980 he accepted an appointment as Statistician in the Environmental Protection agency, Washington, D.C. He married the former Miss Katherine Mary Brooks of Garner, North Carolina, in 1974.
ACKNOWLEDGEMENTS

To give appreciative thanks to all persons who have inspired, counseled, and assisted in this study, would be impossible. The author does, however, wish to single out for special mention Dr. F. G. Giesbrecht and Dr. A. R. Manson for their invaluable advice and encouragement in the completion of this study. Appreciation also is extended to the other members of the committee, Drs. R. G. D. Steel, A. H. E. Grandage, R. A. Schrimper and T. W. Reiland. The author also wishes to express his sincere appreciation to Drs. D. J. Drummond and P. M. Burrows for their encouragement and friendship through the past years.

A special note of gratitude should be sounded for the author's secretary, Judith B. Adcock, for her careful editing and typing of the dissertation.

Finally, the author wishes to thank his wife, Kathy, for her patience during the course of this study.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>v</td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td></td>
</tr>
<tr>
<td>1.1 Preliminary Notation</td>
<td>3</td>
</tr>
<tr>
<td>2. LITERATURE REVIEW</td>
<td></td>
</tr>
<tr>
<td>2.1 The Ordinary Least Squares Estimator</td>
<td>3</td>
</tr>
<tr>
<td>2.2 The Minimum Euclidean Squared-distance Estimator</td>
<td>6</td>
</tr>
<tr>
<td>2.3 The Artificially Restricted Estimator</td>
<td>8</td>
</tr>
<tr>
<td>2.4 The James-Stein Estimator</td>
<td>9</td>
</tr>
<tr>
<td>2.5 The Marquardt Estimator</td>
<td>11</td>
</tr>
<tr>
<td>2.6 The Ridge Regression Estimator of Hoerl and Kennard</td>
<td>14</td>
</tr>
<tr>
<td>3. WEIGHTED ESTIMATION</td>
<td></td>
</tr>
<tr>
<td>3.1 Basic Properties</td>
<td>20</td>
</tr>
<tr>
<td>3.2 Weighted Estimation and the Design of Experiments</td>
<td>26</td>
</tr>
<tr>
<td>3.3 Restricted O.L.S., Weighted Estimation, Infinite-n</td>
<td>30</td>
</tr>
<tr>
<td>3.4 Changing the Sign by Use of Weighted Estimation</td>
<td>34</td>
</tr>
<tr>
<td>4. WEIGHTED RIDGE REGRESSION</td>
<td></td>
</tr>
<tr>
<td>4.1 Basic Properties of Weighted Ridge Regression</td>
<td>37</td>
</tr>
<tr>
<td>4.2 Relationship of Weighted Ridge to Other Estimators</td>
<td>42</td>
</tr>
<tr>
<td>4.3 The Changing of the k-value for Fixed n-Weighting</td>
<td>47</td>
</tr>
<tr>
<td>4.4 Bias-Squared Comparisons for Different Estimators</td>
<td>51</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------------------------</td>
</tr>
<tr>
<td>4.5</td>
<td>Comparison of Bias-Squareds by a Geometric Argument</td>
</tr>
<tr>
<td>4.6</td>
<td>The Sequential Weighted Ridge Estimator</td>
</tr>
<tr>
<td>5.</td>
<td>THE CALCULATION OF n-WEIGHTS</td>
</tr>
<tr>
<td>5.1</td>
<td>A-optimality of the Variance-covariance Matrix</td>
</tr>
<tr>
<td>5.2</td>
<td>A-optimality of the Bias-squared Matrix</td>
</tr>
<tr>
<td>5.3</td>
<td>D-optimality of the Variance-covariance Matrix</td>
</tr>
<tr>
<td>5.4</td>
<td>Minimum Residual Sum of Squares</td>
</tr>
<tr>
<td>5.5</td>
<td>Stability of Estimates of Parameters</td>
</tr>
<tr>
<td>5.5.1</td>
<td>The $Q^*$ Standard</td>
</tr>
<tr>
<td>5.5.2</td>
<td>The $Q^{**}$ Standard</td>
</tr>
<tr>
<td>5.5.3</td>
<td>The $S^*$ Standard</td>
</tr>
<tr>
<td>5.5.4</td>
<td>The $S^{**}$ Standard</td>
</tr>
<tr>
<td>5.6</td>
<td>Sign Change of an Estimated Parameter</td>
</tr>
<tr>
<td>5.7</td>
<td>Iterative Sequential n-weighting</td>
</tr>
<tr>
<td>6.</td>
<td>SPECIFIC WEIGHTED RIDGE REGRESSIONS</td>
</tr>
<tr>
<td>6.1</td>
<td>Functional Weighted Ridge Regression</td>
</tr>
<tr>
<td>6.1.1</td>
<td>The single eigenvector case</td>
</tr>
<tr>
<td>6.1.2</td>
<td>The p-eigenvectors case</td>
</tr>
<tr>
<td>6.2</td>
<td>Prior Ridge Regression</td>
</tr>
<tr>
<td>6.3</td>
<td>Variance Weighted Ridge Regression</td>
</tr>
<tr>
<td>7.</td>
<td>AN EXAMPLE</td>
</tr>
<tr>
<td>8.</td>
<td>SUMMARY</td>
</tr>
<tr>
<td></td>
<td>REFERENCES</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Commonly used techniques to estimate the k-value</td>
<td>19</td>
</tr>
<tr>
<td>3.1</td>
<td>An interpretation of n-weighting</td>
<td>24</td>
</tr>
<tr>
<td>7.1</td>
<td>Initial values for the example (p=5, \ n_0=9)</td>
<td>112</td>
</tr>
<tr>
<td>7.2</td>
<td>Transformed values for the example (p=5, \ n_0=9)</td>
<td>113</td>
</tr>
<tr>
<td>7.3</td>
<td>Ridge regression values</td>
<td>115</td>
</tr>
<tr>
<td>7.4</td>
<td>Iterative n-weightings for A-optimality of the variance-covariance matrix</td>
<td>116</td>
</tr>
<tr>
<td>7.5</td>
<td>Iterative n-weightings for D-optimality of the variance-covariance matrix</td>
<td>122</td>
</tr>
<tr>
<td>7.6</td>
<td>n-weightings for 1st iteration, A-optimality of the variance-covariance matrix</td>
<td>123</td>
</tr>
<tr>
<td>7.7</td>
<td>n-weightings for 1st iteration, D-optimality of the variance-covariance matrix</td>
<td>124</td>
</tr>
<tr>
<td>7.8</td>
<td>One point infinitely weighted, A-optimality of the bias matrix, then A-optimality of the variance-covariance matrix</td>
<td>126</td>
</tr>
<tr>
<td>7.9</td>
<td>Two points infinitely weighted, A-optimality of the bias matrix, then A-optimality of the variance-covariance matrix</td>
<td>127</td>
</tr>
<tr>
<td>7.10</td>
<td>The points infinitely weighted, A-optimality of the bias matrix, the A-optimality of the variance-covariance matrix</td>
<td>128</td>
</tr>
</tbody>
</table>
7.11 Four points infinitely weighted, A-optimality of the bias matrix, the A-optimality of the variance-covariance matrix .............. 129
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Trace mean square error for the ridge estimator</td>
<td>16</td>
</tr>
<tr>
<td>2.2</td>
<td>The ridge-plot</td>
<td>18</td>
</tr>
<tr>
<td>4.1</td>
<td>The difference in bias-squareds resulting in all positive eigenvalues</td>
<td>56</td>
</tr>
<tr>
<td>4.2</td>
<td>The difference in bias-squareds resulting in indeterminant eigenvalues due to different orientation of axes</td>
<td>56</td>
</tr>
<tr>
<td>4.3</td>
<td>The differences in bias-squareds resulting in indeterminant eigenvalues due to the interaction of ellipses</td>
<td>59</td>
</tr>
<tr>
<td>4.4</td>
<td>The differences in bias-squareds resulting in positive equivalent eigenvalues</td>
<td>59</td>
</tr>
<tr>
<td>5.1</td>
<td>An augmented design for $p = 2$, $n_o = 6$</td>
<td>82</td>
</tr>
<tr>
<td>7.1</td>
<td>The ridge-plot for estimates in standard ridge regression</td>
<td>117</td>
</tr>
<tr>
<td>7.2</td>
<td>Iterative n-weightings for A-optimality of the variance-covariance matrix</td>
<td>118</td>
</tr>
<tr>
<td>7.3</td>
<td>Individual weighting of points for A-optimality of the variance-covariance matrix</td>
<td>119</td>
</tr>
<tr>
<td>7.4</td>
<td>Beta-coefficients as a function of n-weights chosen, $k = 0.01$</td>
<td>120</td>
</tr>
</tbody>
</table>
1. INTRODUCTION

There are many situations when the possible linear relationship of a response variable to a set of known explanatory variables may be investigated. If the ith observation of the response variable is denoted by $Y_i$, and the corresponding values of the set of known variables by $X_{i1}, X_{i2}, \ldots, X_{ip}$, then the functional relationship between the Y-variable and the X-variables may be expressed as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i,$$

(1.1)

where

$\beta_0$ is a constant,

$\beta_j$ is a regression coefficient $j = 1, 2, \ldots, p$,

and

$\epsilon_i$ is an unobservable error $i = 1, 2, \ldots, n_o$.

The estimation of the regression coefficients of equation (1.1) is usually done by the technique known as ordinary least squares estimation (O.L.S.). When the model describing the linear relationship (equation (1.1)) is correct, then O.L.S. estimation provides unbiased estimates of the parameters $\beta_0, \beta_1, \ldots, \beta_p$.

There are, however, situations when these unbiased estimates do not give physically meaningful values when put into the context of the engineering, biological or economic situation under investigation.
It is under such circumstances that alternatives to O.L.S. estimation are sought. There are situations where the cause of meaningless estimates may be an inadequate choice of model, choice of an incorrect model, misspecification of the distribution of error terms in the model, etc. These problems are, to a certain degree, correctable when appropriate statistical techniques are used. However, situations arise where the choice of model may be correct, the choice of distribution of error terms appropriate, but O.L.S. estimation may fail to yield satisfactory estimates. If the initial set of known variables is called the design matrix (of dimension $n_o \times p$), then the set of responses will be called the observed vector ($n_o \times 1$). When the design matrix is of full column rank but at least one column can nearly be expressed as a linear function of the remaining columns, the matrix is said to be suffering from multicollinearity or ill conditioning. Exact multicollinearity would imply a perfect linear function existed between the columns, and the resulting design matrix would be of less than full column rank.

The key disadvantage associated with O.L.S. estimation, in the presence of multicollinearity, is that the estimates have unacceptably large variances and consequently may be of unreasonable magnitude or sign. It will be assumed that only the initial set of variables and responses are available to the analyst. If further controlled experimentation or data collection were possible, then the "correlation bonds" between the variables in the design matrix may be weakened by the judicious choice of levels of the known variables. This approach has been investigated by Silvey (1969), Dykstra (1971), and others.
1.1 Preliminary Notation

The severity of multicollinearity may be reduced by transforming
the initial design matrix into the standardized form. In this
standardized form, the "X'X" matrix is in the form of a correlation
matrix. The linear model in the original variables may be written as

\[ Y^o = \beta_o \mathbf{1} + X^o \beta^o + \xi \]  \hspace{1cm} (1.1.1)

where

- \( Y^o \) is an \( n_o \times 1 \) vector of observable random variables and
  represents the response or dependent variable,
- \( \mathbf{1} \) is the \( n_o \times 1 \) unit vector,
- \( X^o \) is the \( n_o \times p \) initial design matrix and represents the
  regressor or independent variables,
- \( X^o = [x^o_1, x^o_2, \ldots, x^o_p] \),
- \( \beta^o \) is the \( p \times 1 \) vector of unknown parameters,
- \( \beta_o \) is an unknown constant,
- \( \xi \) is the \( n_o \times 1 \) vector of unobservable independent random
  errors, assumed to be normally distributed with mean zero
  and variance \( \sigma^2 \).

The standardized model is

\[ Y = X_\beta + \xi \]  \hspace{1cm} (1.1.2)
where

\[ Y = Y^o - \bar{Y}_1, \]

\[ \bar{y} = \frac{1}{n_0} Y_1, \]

\[ X = [x_1, x_2, \ldots, x_p], \]

\[ x_j = \frac{(x_j^o - \bar{x}_j)}{(x_j^o - \bar{x}_j)^\top (x_j^o - \bar{x}_j)}^{1/2}, \]

\[ x_j = \bar{x}_j 1, \]

and

\[ \bar{x}_j = \frac{1}{n_0} x_j 1. \]

The standardized coefficients may be related to the original coefficients by

\[ \beta = S\beta^o, \quad (1.1.3) \]

where \( S \) is a diagonal matrix with a jth element \( \{(x_j^o - \bar{x}_j)^\top (x_j^o - \bar{x}_j)\}^{1/2} \).

This standardization is sometimes referred to as 'the removal of unnecessary multicollinearity' by analysts.
2. LITERATURE REVIEW

The estimation of the unknown parameters, \( \beta \), in the standard model

\[
\mathbf{y} = \mathbf{X}\beta + \varepsilon,
\]

occupies a considerable portion of modern statistical literature. The principal estimators and the paragraph numbers discussing them are:

2.1 The Ordinary Least Squares Estimator,
2.2 The Minimum Euclidean Squared-distance Estimator,
2.3 The Artificially Restricted Estimator,
2.4 The James-Stein Estimator,
2.5 The Marquardt Estimator, and
2.6 The Ridge Regression Estimator of Hoerl and Kennard.

Other techniques of estimation, such as nonparametric estimation, will not be considered in this dissertation.

2.1. The Ordinary Least Squares Estimator.

The estimation technique and discussion of the properties of the O.L.S. estimator

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

(2.1.1)

may be found in standard texts such as Goldberger (1964) or Searle (1971).

If the Euclidean squared-distance of the O.L.S. estimator from the true parameter vector \( \beta \) is defined as

\[
L^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta),
\]

(2.1.2)
then

\[ E(L^2) = \sigma^2 \text{tr}(X'X)^{-1}, \quad (2.1.3) \]

with

\[ V(L^2) = 2\sigma^4 \text{tr}(X'X)^{-2}. \quad (2.1.4) \]

Equations (2.1.3) and (2.1.4) may be written in canonical form as

\[ E(L^2) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \quad (2.1.5) \]

and

\[ V(L^2) = 2\sigma^4 \sum_{i=1}^{p} \frac{1}{\lambda_i^2} \quad (2.1.6) \]

where \( \lambda_i \) is the \( i \)th eigenvalue of \( X'X \). It is clear that when \( \lambda_p \) is small, \( E(L^2) \) of equation (2.1.5) is large, implying that the O.L.S. estimates may be far from the true parameter values. As \( \lambda_p \) is small, it follows that \( V(L^2) \) of equation (2.1.6) will be large.

2.2. The Minimum Euclidean Squared-distance Estimator

The estimator which minimizes the expected Euclidean squared-distance from the estimates to the true parameter values is defined to be \( \tilde{\beta}_E \). If the estimator is a linear function of the observed values then

\[ \tilde{\beta}_E = F'y, \quad (2.2.1) \]

By definition of expected Euclidean squared-distance from estimate to parameter,
\[ E(L_E^2) = E(\beta E - \bar{q})'(\beta E - \bar{q}) \quad (2.2.2) \]

Substitution of equation (2.2.1) into equation (2.2.2) yields

\[ E(L_E^2) = \text{tr}\{(F'X-I)\beta \beta'(F'X-I)' + \sigma^2 FF'\} \quad (2.2.3) \]

Minimization of the expression for \( E(L_E^2) \) given in equation (2.2.3) with respect to \( F \) yields,

\[ F' = \frac{\beta \beta'X'(X \beta \beta'X' + \sigma^2 I)^{-1}}{\sigma^2} \quad (2.2.4) \]

By use of Corollary A.1.3 of the Appendix, Equation (2.2.4) may be expressed as

\[ F' = \frac{1}{\sigma^2} \frac{\beta \beta'X'}{\sigma^2 + \beta'X'X \beta} \frac{\beta \beta'X'X \beta \beta'X'}{\sigma^2} \quad (2.2.5) \]

and substituted into equation (2.2.1) to give

\[ \beta_E = \left[ \frac{\beta \beta'X'}{\sigma^2 + \beta'X'X \beta} \right] Y \quad (2.2.6) \]

The expected value of \( \beta_E \) of Equation (2.2.6) is

\[ E(\beta_E) = \left[ \frac{\beta'X'X \beta}{\sigma^2 + \beta'X'X \beta} \right] \beta \quad (2.2.7) \]

and consequently, \( \beta_E \) underestimates \( \beta \).

It should be noted that substitution of estimates for parameters in Equation (2.2.6) yields an estimator whose exact properties are unknown.
2.3. The Artificially Restricted Estimator

It is well known that placing restrictions on $\beta$ in the form $R\beta = r$ results in estimators with smaller variance than those obtained by O.L.S. estimation. Formally, this means that $V(\beta) - V(\beta_{\text{restricted}})$ is positive semi-definite. This holds true even when the restrictions placed on $\beta$ are not valid (Goldberger, 1964). Toro-Vizcarrondo and Wallace (1968) investigated this phenomenon with respect to its influence on mean square error estimation.

With the artificial restriction $R\beta = r$ (where $R$ is $q \times p$ and of rank $q$), the restricted estimator is

$$\beta_{\text{restricted}} = \hat{\beta} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r), \quad (2.3.1)$$

and has a variance-covariance matrix,

$$[(X'X)^{-1} - (X'X)^{-1}R'R(X'X)^{-1}R']^{-1}R(X'X)^{-1}\sigma^2. \quad (2.3.2)$$

The difference in mean squared error, $\text{MSE}(\beta) - \text{MSE}(\beta_{\text{restricted}})$, is

$$\begin{align*}
(X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}[R(X'X)^{-1}R']\sigma^2 \\
- (R\hat{\beta} - r)(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}.
\end{align*} \quad (2.3.3)$$

This matrix is positive semi-definite when

$$Q_R = R(X'X)^{-1}R'\sigma^2 - (R\hat{\beta} - r)(R\hat{\beta} - r)' \quad (2.3.4)$$

is positive semi-definite.
The condition for $Q_R$ to be positive semi-definite is

$$\lambda = \frac{(R\mathbf{S}-\mathbf{r})(R(X'X)^{-1}R)^{-1}(R\mathbf{S}-\mathbf{r})}{2\sigma^2} \leq \frac{1}{2}$$ \hspace{1cm} (2.3.5)

The $\lambda$ of Equation (2.3.5) may be related to the non-centrality parameter of a non-central $F$-distribution. Although this technique of artificially restraining an estimator has some merits, it is seldom used by analysts owing to the difficulty of determining the optimum restriction.

2.4. The James-Stein Estimator

James and Stein (1961) defined an estimator that was a function of the O.L.S. estimator of the form

$$\hat{\mathbf{b}}_{JS} = d\hat{\mathbf{b}}$$ \hspace{1cm} (2.4.1)

Their criterion for the determination of $d$ was to minimize the expected Euclidean squared-distance between the estimates to the corresponding values. If "d" of equation (2.4.1) is

$$d = 1 - \frac{c\sigma^2}{\hat{\mathbf{b}}\cdot\hat{\mathbf{b}}}$$ \hspace{1cm} (2.4.2)

where $0 \leq c \leq 2(p-2)$, then

$$E(\hat{\mathbf{b}}_{JS}-\mathbf{b})' (\hat{\mathbf{b}}_{JS}-\mathbf{b}) < E(\hat{\mathbf{b}}-\mathbf{b})' (\hat{\mathbf{b}}-\mathbf{b}).$$ \hspace{1cm} (2.4.3)

If $\sigma^2$ is unknown, but estimated from the data by $s^2$, then the limits of $c$ in equation (2.4.2) are
\[ 0 \leq c \leq \frac{2(p-2)(n_o-p)}{(n_o-p+2)}, \] (2.4.4)

and \( E(\tilde{\beta}_{JS} - \beta)(\tilde{\beta}_{JS} - \beta) \) is minimized when

\[ c = \frac{(p-2)(n_o-p)}{(n_o-p+2)}. \] (2.4.5)

Sclove (1968) points out that when the elements of the vector \( \beta \) can be separated into two distinct parts

a) parameters where unbiased estimators are desired, and

b) parameters where James-Stein estimators are desired,

then O.L.S. estimation may be combined with biased James-Stein estimation.

Let

\[ \hat{\beta}^* = \begin{bmatrix} \hat{\beta}_a^* \\ \hat{\beta}_b^* \end{bmatrix}, \]

with the O.L.S. estimate of \( \beta_a \) denoted by \( \hat{\beta}_a(p_a \times 1) \), and the James-Stein estimate of \( \beta_b \) denoted by \( \hat{\beta}_{bJS}(p_b \times 1) \). Define

\[ \hat{\beta}_{bJS} = \left[ 1 - \frac{c_b^* \hat{\beta}_b^2}{\hat{\beta}_b^2} \right] \hat{\beta}_b, \] (2.4.6)

where \( \hat{\beta}_b \) is the O.L.S. estimate of \( \beta_b \), and if \( c_b^* \) is such that

\[ 0 \leq c_b^* \leq \frac{2(p_b-2)(n_o-p)}{(n_o-p+2)}, \]
then

\[ E(\tilde{\beta}_b - \beta_b \mid \sigma^2)^{\prime}(\tilde{\beta}_b - \beta_b \mid \sigma^2) < E(\tilde{\beta}_b - \beta_b \mid \sigma^2)^{\prime}(\tilde{\beta}_b - \beta_b \mid \sigma^2), \]  

(2.4.7)

with a minimum mean square error when

\[ c_b^* = \frac{(p-2)(n_0 - p)}{(n_0 - p + 2)}. \]  

(2.4.8)

It may be shown that when \( \sigma^2 \) is known, the expected shrinkage in the James-Stein estimator is more than for the case where \( \sigma^2 \) is unknown, and estimated from the data. The greatest disadvantage with James-Stein estimation is that the theoretical properties of the estimator have only been investigated for the case of complete orthogonality of design, \( X'X = I \). When the design used is not orthogonal, James-Stein-type estimation yields an estimator with intractable properties.

2.5 The Marquardt Estimator

The preceding estimators have assumed the design matrix \( X \) to be of full column rank which implies that \( XX' \) has \( p \) non-zero eigenvalues. There are occasions when some of the smallest eigenvalues are very close to zero and severe multicollinearity exists. As Marquardt (1970) reasons,
practical estimation problems give rise to matrices $X'X$ having eigenvalues that may be grouped into three types: substantially greater than zero, slightly greater than zero, precisely zero (except for rounding error). In computations it may sometimes be difficult to distinguish between adjacent types.

The Marquardt, or fractional rank estimator, replaces the $(X'X)^{-1}$ used in O.L.S. estimation by $(X'X)^+,$ a matrix of rank $r,$ $(r < p).$

Marquardt defined $(X'X)^{-1}$ as

$$ (X'X)^{-1} = \sum_{j=1}^{p} \frac{1}{\lambda_j} \xi_j \xi_j', $$

$$ PAP', $$

where $P = [\xi_1, \ldots, \xi_p].$ He then defined $(X'X)^+ r$ as

$$ (X'X)^+ r = \sum_{j=1}^{r} \frac{1}{\lambda_j} \xi_j \xi_j', $$

$$ P_r P_r', $$

where

$$ P = \begin{bmatrix} P_r & P_{p-r} \\ \end{bmatrix}, $$

and

$$ \Lambda = \begin{bmatrix} \Lambda_r & 0 \\ 0 & \Lambda_{p-r} \end{bmatrix}. $$
Suppose the matrix \( X'X \) has an unknown rank, but known to be either \( r \) or \((r+1)\). The fractional rank estimator is defined as

\[
\hat{\beta}_M^+ = (X'X)^+_r X'Y
\]  \hspace{1cm} (2.5.2)

where

\[
(X'X)^+_r = \sum_{j=1}^{r^*} \frac{1}{\lambda_j} \xi_j \xi_j' + \left[ \frac{d_r}{\lambda_{r^*+1}} \right] \xi_{r^*+1} \xi_{r^*+1}'
\]

and

\[
d_r = r - r^*, \quad 0 < d_r < 1.
\]

From equation (2.5.2), note that \( \hat{\beta}_M^+ \) is a biased estimator unless \( A_{p-r} \) is the null matrix. The variance of the fractional rank estimator defined in equation (2.5.2) is

\[
V(\hat{\beta}_M^+) = V[P_r \Lambda_{r}^{-1} P_r \hat{X}'X \hat{\beta}]
\]  \hspace{1cm} (2.5.3)

\[
= P_r \Lambda_{r}^{-1} P_r \sigma^2,
\]

and the bias-matrix is

\[
B(\hat{\beta}_M^+) = [P_r \Lambda_{r}^{-1} P_r \hat{X}'X-I]\hat{\beta}^[d][P_r \Lambda_{r}^{-1} P_r \hat{X}'X-I]'. \hspace{1cm} (2.5.4)
\]

For the mean square error of \( \hat{\beta}_M^+ \) to be less than the variance of ordinary least squares, it is sufficient that

\[
\sigma^2 \sum_{j=r+1}^{p} \frac{1}{\lambda_j} \xi_j > \hat{\beta}_M^+ \hat{\beta}.. \hspace{1cm} (2.5.5)
\]
The major problem with the Marquardt estimator is the determination of "r." In situations where the rank of $X^\top X$ is reasonably large, a clustering of groups of eigenvalues is common and the selection of appropriate $r^*$ (and hence r) difficult.

2.6 The Ridge Regression Estimator of Hoerl and Kennard

Hoerl and Kennard (1970) proposed adding a small positive constant to the main diagonal of $X^\top X$ prior to its inversion in the O.L.S. procedure. The ridge estimator is defined as

$$\hat{\beta}^* = (X^\top X+kI)^{-1}X^\top y,$$

$$= [I+k(X^\top X)^{-1}]^{-1}\hat{\beta}.$$  \hspace*{1cm} (2.6.1)

This estimator has a variance-covariance matrix,

$$V(\hat{\beta}^*) = [I+k(X^\top X)^{-1}]^{-1}(X^\top X)^{-1}[I+k(X^\top X)^{-1}]^{-1}\sigma^2,$$  \hspace*{1cm} (2.6.2)

and bias-matrix,

$$B(\hat{\beta}^*) = k^2(X^\top X+kI)^{-1}\hat{\beta}^*(X^\top X+kI)^{-1}.$$  \hspace*{1cm} (2.6.3)

The expected Euclidean squared-distance $E(L^*^2)$ is defined as

$$E(L^*^2) = E(\hat{\beta}^* - \beta)^\top (\hat{\beta}^* - \beta),$$

and may be written as

$$E(L^*^2) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2} + k^2\hat{\beta}^*(X^\top X+kI)^{-2}\hat{\beta},$$  \hspace*{1cm} (2.6.4)

by application of equations (2.6.2) and (2.6.3). This expected Euclidean squared-distance is equivalent to the trace of the mean square error of $\hat{\beta}^*$. If equation (2.6.4) is written in canonical form, then
\[ E(L^*^2) = \sum_{i=1}^{P} \frac{\lambda_i \sigma_i^2}{\lambda_i + k} + k \sum_{i=1}^{P} \frac{\alpha_i^2}{(\lambda_i + k)^2} \]  \hspace{1cm} (2.6.5)

where \( \alpha = \frac{P}{\lambda_i} \).

The expected Euclidean squared-distance for the O.L.S. estimator may be compared to that for the ridge regression estimator by defining

\[ l^*_{E} = E(L^2) - E(L^*^2) \]

\[ = \sum_{i=1}^{P} \left[ \frac{\sigma_i^2}{\lambda_i} - \frac{(\lambda_i \sigma_i^2 + k \alpha_i^2)}{(\lambda_i + k)^2} \right]. \hspace{1cm} (2.6.6) \]

Ridge regression has a smaller Euclidean squared distance if \( l^*_{E} \) of equation (2.6.6) is positive, which implies that

\[ k \leq \frac{\sigma_i^2}{\alpha_{\text{max}}} \]. \hspace{1cm} (6.2.7) 

It may be shown that the trace of the variance-covariance matrix of the ridge estimator (the first term of equation (2.6.5)), is a monotonically decreasing function for positive \( k \)-value. Similarly, the trace bias-matrix (the second term of equation (2.6.5)) is a monotonically increasing function for positive \( k \)-value. The behavior of the trace of the mean square error matrix of the ridge estimators is illustrated in Figure 2.1. There is no simple method of obtaining the \( k \)-value which minimizes \( E(L^*^2) \) of equation (2.6.5). A commonly used technique is to determine a \( k \)-value which stabilizes the estimates via visual examination of the ridge-plot. In the construction of a ridge-plot, individual estimates are plotted as a
Figure 2.1. Trace mean square error for the ridge estimator.
function of k, and a k-stability chosen visually, Figure 2.2 being a typical example. There are many techniques for the choosing of a k-value to be used in ridge estimation. Table 2.1 gives some of the more commonly used k-values.

It must be noted that a non-stochastic k-value is needed for the mathematical properties of the ridge estimator to be valid. Many of the k-values proposed in the literature are stochastic and their properties discussed with reference to simulation experiments. In this dissertation, it will be assumed the k-value is of a non-stochastic nature, implying that the observed values are not used to determine the appropriate k-value for the ridge estimator.
Estimated $\beta$-coefficients

All estimates tend to zero as $k$ tends to $\infty$.

O.L.S. estimates denoted by $\ast$.

Figure 2.2. The ridge-plot.
Table 2.1. Commonly used techniques to estimate the k-value.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Proponent</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Visual ridge-plot</td>
<td>Hoerl and Kennard (1970)</td>
</tr>
<tr>
<td>(ii) ( k = \frac{\hat{\sigma}^2}{(\hat{\beta}_k^2)^{\frac{1}{2}}} )</td>
<td>Newhouse and Oman (1971)</td>
</tr>
<tr>
<td>(iii) Choose ( k ) such that ( \hat{\beta}_k^2 = c ) if ( c &gt; 0 ), Choose ( k = 0 ) if ( c &lt; 0 ), where ( c = \hat{\beta}<em>k^2 - \sigma^2 \sum</em>{i=1}^{p} \frac{1}{\lambda_i} )</td>
<td>McDonald and Galarneau (1975)</td>
</tr>
<tr>
<td>(iv) ( k = \frac{p\sigma^2}{\hat{\beta}_k^2} )</td>
<td>Hoerl, Kennard and Baldwin (1975)</td>
</tr>
<tr>
<td>(v) Choose ( k ) via k-equality pre-test</td>
<td>Obenchain (1975)</td>
</tr>
<tr>
<td>(vi) ( k = \frac{p\sigma^2}{\hat{\beta}_k^2} )</td>
<td>Hoerl and Kennard (1976)</td>
</tr>
<tr>
<td>(vii) ( k = \frac{p\sigma^2}{\sum_{i=1}^{p} \frac{1}{\lambda_i^2}} )</td>
<td>Lawless and Wang (1976)</td>
</tr>
<tr>
<td>(viii) Choose ( k ) via a stability index</td>
<td>Vinod (1976)</td>
</tr>
<tr>
<td>(ix) Choose ( k ) with reference to the eigenvalues of ( X'X )</td>
<td>Marquina (1978)</td>
</tr>
</tbody>
</table>
3. WEIGHTED ESTIMATION

The estimation procedures discussed in Chapter 2 assumed each observation had the same (or standard) weight. Weighted estimation allows observation to have unequal weights. This weighting, called \( \pi \)-weighting where \( 0 < \pi < \infty \), is a monotone, smooth function. A point with standard weight has a \( \pi \)-weighting equal to unity. With this system of weighting imposed on the linear model the properties of the weighted estimator may be investigated.

3.1. Basic Properties

The observed values, \( Y \), are related to the independent values, \( X \), by the linear model,

\[
Y = X\theta + \varepsilon
\]  \hspace{1cm} (3.1.1)

where the error vector \( \varepsilon \), is normally distributed with mean \( 0 \), variance \( \sigma^2 \), and \( \theta \) is the vector of unknown parameters.

Without loss of generality, consider the weighting of the first observation. The observations have a weighting-matrix \( \Omega_1 \), where, \( \Omega_1 \), is the identity matrix with first element replaced by a weighting, \( \pi \).
\[ \hat{\beta}_{(1)} = (X'\tilde{\Omega}_1^{-1}X)^{-1}X'\tilde{\Omega}_1^{-1}Y. \] (3.1.2)

It may be easily established that,

\[ E(\hat{\beta}_{(1)}) = \beta, \] (3.1.3)

with a variance-covariance matrix,

\[ V(\hat{\beta}_{(1)}) = (X'\tilde{\Omega}_1^{-1}X)^{-1}X'\tilde{\Omega}_1^{-2}X(X'\tilde{\Omega}_1^{-1}X)^{-1}\sigma^2. \] (3.1.4)

From the Gauss-Markov theorem, the difference between variance-covariance matrices of weighted point and O.L.S. estimators is positive semi-definite for any \( \pi \)-weighting other than the trivial case of all \( \pi_i \)'s = 1.

The weighting of more than a single point results in the weighted estimator, defined as,

\[ \hat{\beta} = (X'\tilde{\Omega}_n^{-1}X)^{-1}X'\tilde{\Omega}_n^{-1}Y, \] (3.1.5)

with,

\[ E(\hat{\beta}) = \beta, \] (3.1.6)

and variance-covariance matrix given by,

\[ V(\hat{\beta}) = (X'\tilde{\Omega}_n^{-1}X)^{-1}X'\tilde{\Omega}_n^{-2}X(X'\tilde{\Omega}_n^{-1}X)^{-1}\sigma^2. \] (3.1.7)

The \( \pi \)-weighting for use with weighted estimation is,
\[ \Omega_{n_0} = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \vdots \\ \vdots \\ \pi_{n_0} \end{bmatrix} \]  
(3.1.8)

From equation (3.1.1), the first observation, or data-point, may be written,

\[ y = \underline{w} \beta + \epsilon_1, \]  
(3.1.9)

where \( \underline{w} \) is the first row of \( X \), and \( y \) the first element of \( Y \). Equation (3.1.2) may be written as,

\[ \tilde{\beta}_{(1)} = \left( \frac{1}{n_0} \underline{w} \underline{w}^T + \sum_{i=2}^{n_0} \underline{w}_i \underline{w}_i^T \right)^{-1} \left( \frac{1}{n_0} \underline{w} y + \sum_{i=2}^{n_0} \underline{w}_i y_i \right), \]

where,

\( \underline{w}_i \) is the \( i^{th} \) row of \( X \),

\( y_i \) is the \( i^{th} \) element of \( Y \),

\[ \sum_{i=1}^{n_0} \underline{w}_i \underline{w}_i^T = X^T X \]

and,

\[ \sum_{i=1}^{n_0} \underline{w}_i y_i = X^T y. \]

Therefore equation (3.1.10) may be written as,
\[ \tilde{\beta}(1) = (X'X + nw_1w)'^{-1}(X'Y + nw_1y), \quad (3.1.11) \]

and equation (3.1.7) as,

\[ V(\tilde{\beta}(1)) = (X'X + nw_1w)'^{-1}[X'X + n(n + 2)w_1w'](X'X + nw_1w)'^{-1} \sigma^2, \quad (3.1.12) \]

where,

\[ n = \frac{1}{\pi} - 1. \]

The advantages of using \( n \)-weighting include simplicity in computation, the direct evaluation of the effect of weighting the first point, and a relationship with the design of an experiment (to be discussed in Section 3.2 of Chapter 3). An explanation of \( n \)-weighting and \( \pi \)-weighting may be found in Table 3.1.

The extension to weighted estimation may be easily accomplished by defining,

\[ \tilde{\Omega}_0^{-1} = I + D, \]

where \( D \) is a diagonal matrix with \( i \)-th diagonal element \( n_1 = \frac{1}{\pi} - 1. \)

From equation (3.1.5),

\[ \tilde{\beta} = (X'\tilde{\Omega}_0^{-1}X)_0^{-1}X'\tilde{\Omega}_0^{-1}y_0 \]

\[ = (X'X + X'DX)_0^{-1}(X'Y + X'Dy). \quad (3.1.13) \]

However,

\[ X'DX = \sum_{i=1}^{n_0} n_i w_1 w_i', \]

and,

\[ X'DX = \sum_{i=1}^{n_0} n_i w_1 y_i', \]
Table 3.1. An interpretation of $n$ and $\pi$-weighting.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\pi$</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>$\infty$</td>
<td>This implies the point under consideration has been completely eliminated from the data set.</td>
</tr>
<tr>
<td>$-1 &lt; n &lt; 0$</td>
<td>$1 &lt; \pi &lt; \infty$</td>
<td>Represents the gradual elimination of the point from the data set. In other words, a &quot;discounting&quot; of its &quot;importance.&quot;</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>This is ordinary least squares.</td>
</tr>
<tr>
<td>$0 &lt; n &lt; \infty$</td>
<td>$0 &lt; \pi &lt; 1$</td>
<td>Represents the assigning of greater &quot;importance&quot; to this point. In other words, the residual of the fitted function at this point is diminished with increasing $n$.</td>
</tr>
<tr>
<td>$\infty$</td>
<td>0</td>
<td>The residual at this point is forced to be zero, the fitted function &quot;passes&quot; through this point.</td>
</tr>
</tbody>
</table>
giving,
\[
\hat{\beta} = \left( X'X + \sum_{i=1}^{n_0} n_i \hat{w}_i \hat{w}_i' \right)^{-1} \left( X'\hat{y} + \sum_{i=1}^{n_0} n_i \hat{w}_i \hat{y}_i \right).
\] (3.1.14)

The variance-covariance matrix for \( \hat{\beta} \) is obtained from equation (3.1.12),
\[
V(\hat{\beta}) = (X'\Omega^{-2}X)^{-1} X'\Omega^{-1} X \sigma^2
\]
\[
= (X'X + X'DX)^{-1} (X'X + 2X'DX + X'D^2X)(X'X + X'DX)^{-1} \sigma^2
\]
\[
= \left( X'X + \sum_{i=1}^{n_0} n_i \hat{w}_i \hat{w}_i' \right)^{-1} \left[ X'X + \sum_{i=1}^{n_0} n_i (n_i + 2) \hat{w}_i \hat{w}_i' \right]
\]
\[
\left( X'X + \sum_{i=1}^{n_0} n_i \hat{w}_i \hat{w}_i' \right)^{-1} \sigma^2,
\] (3.1.15)

where,
\[
X'D^2X = \sum_{i=1}^{n_0} n_i \hat{w}_i \hat{w}_i'.
\]

These forms for the weighted estimator and its variance are too cumbersome for efficient algebraic manipulation and so the following nomenclature is adopted. For weighted point estimation let
\[
\tilde{X}'\tilde{X} = X'X + n_w \hat{w}' \hat{w},
\]

with,
\[
\tilde{X}'\hat{y} = X'\hat{y} + n_w \hat{y}.
\]

Equation (3.1.11) may be written,
\[
\hat{\beta}_{(1)} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\hat{y},
\] (3.1.16)
and equation (3.1.12) as,

\[ V(\tilde{\beta}_{(1)}) = (\tilde{X}^T \tilde{X})^{-1} [\tilde{X}^T \tilde{X} + n(n + 1)\tilde{w}_i \tilde{w}_i' \tilde{X}^T \tilde{X}]^{-1} \sigma^2. \]  

(3.1.17)

For weighted estimation,

\[ \tilde{X}^T \tilde{X} = X^T X + \sum_{i=1}^{n_0} n_i \tilde{w}_i \tilde{w}_i', \]

and

\[ \tilde{X}^T \tilde{Y} = X^T Y + \sum_{i=1}^{n_0} n_i \tilde{w}_i y_i'. \]

Equation (3.1.14) may then be written as

\[ \tilde{\beta} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y}, \]  

(3.1.18)

and equation (3.1.15) as,

\[ V(\tilde{\beta}) = (\tilde{X}^T \tilde{X})^{-1} \left[ \tilde{X}^T \tilde{X} + \sum_{i=1}^{n_0} n_i (n_i + 1) \tilde{w}_i \tilde{w}_i' \tilde{X}^T \tilde{X} \right]^{-1} \sigma^2. \]  

(3.1.19)

The potential confusion over dual definition of \( \tilde{\cdot} \tilde{\cdot} \) cannot occur as weighted point estimation, and weighted estimation, will be clearly identified by presence, or absence, of the subscript "(1)" on the estimate vector.

3.2. Weighted Estimation and the Design of Experiments.

One of the advantages of using the \( n \)-weighting system, as opposed to the \( \pi \)-weighting system, is that there exists a simple relationship between \( n \)-weighting and the design of an experiment. The "design of an experiment" aspect solely concerns the structure of the independent, and dependent variables.
Without loss of generality, suppose the first observation-point is repeated "n" times. All these "new" n points are perfectly correlated with the first point. For simplicity of notation, the new (or augmented) matrices will have the same symbol as the standard matrix but with the addition of a tilde. The basic augmented matrices and vectors are:

\[
\tilde{X} = \begin{bmatrix}
X \\
\vdots \\
\tilde{w}' \\
\vdots \\
w'
\end{bmatrix}, \quad (n_o \times p)
\]

\[
\tilde{Y} = \begin{bmatrix}
\tilde{y} \\
\vdots \\
y
\end{bmatrix}, \quad (n_o \times 1)
\]

\[
\tilde{\varepsilon} = \begin{bmatrix}
\varepsilon \\
\vdots \\
\varepsilon
\end{bmatrix}, \quad (n_o \times 1)
\]

where "e" denotes the (unknown) error first point \( (e = \varepsilon_1) \).

The model under consideration is,

\[
\tilde{Y} = \tilde{X} \hat{\beta} + \tilde{\varepsilon}, \quad (3.2.4)
\]
and the repeated point estimator from this model is,

$$
\hat{\beta}_{(1)R} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y}.
$$

(3.2.5)

Obviously,

$$
E[\hat{\beta}_{(1)R}] = \beta,
$$

(3.2.6)

and,

$$
\text{V}[\hat{\beta}_{(1)R}] = E[\hat{\beta}_{(1)R} - \beta][\hat{\beta}_{(1)R} - \beta]' = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\text{VX}(\tilde{X}'\tilde{X})^{-1}\sigma^2,
$$

(3.2.7)

where,

$$
\tilde{V}_o^2 = E(\tilde{e}\tilde{e}').
$$

Now,

$$
E(\tilde{e}\tilde{e}') = \begin{bmatrix}
E(e_1e_1) & E(e_1e_2) \\
E(e_2e_1) & E(e_2e_2)
\end{bmatrix},
$$

and

$$
\begin{bmatrix}
I \\
(n_o \times n_o)
\end{bmatrix}
\begin{bmatrix}
\tilde{H}' \\
(n_o \times n)
\end{bmatrix}
\begin{bmatrix}
H \\
(n \times n_o)
\end{bmatrix}
\begin{bmatrix}
J \\
(n \times n)
\end{bmatrix}^2,
$$

where

I is the identity matrix,

J has all elements equal to unity,

H has the first column with elements equal to unity, all other elements equal to zero.
If the augmented $X$-matrix is written as,

$$
\tilde{X'} = [X' | \mathbf{w}, \mathbf{w}, \ldots, \mathbf{w}],
$$

$$
= [X' | \mathbf{w}],
$$

then,

$$
\tilde{X'}\tilde{X} = [X' | \mathbf{w}][I \begin{array}{c} H' \\ \hline \hline H \end{array}] [X \begin{array}{c} \mathbf{w} \\ \mathbf{w} \end{array}],
$$

i.e.,

$$
\tilde{X'}\tilde{X} = X'X + WHX + X'H'W' + WJW',
$$

where,

$$
WHX = (\mathbf{w}, \ldots, \mathbf{w})(1, 0, \ldots, 0) X,
$$

$$
= nw' w',
$$

and,

$$
WJW' = (\mathbf{w}, \ldots, \mathbf{w})(1, \ldots, 1)\left[\begin{array}{c} w \\ \vdots \\ w' \end{array}\right],
$$

$$
= n^2 w' w'.
$$

It follows that equation (3.2.10) may be simplified to,

$$
V(\hat{\theta}(1)_R) = [X'X + nw' w']^{-1}[X'X + (2n+n^2)w' w'][X'X + nw' w']^{-1}o^2,
$$

(3.2.8)

which is seen to be the same as equation (3.1.12). The implication is that whenever the $\pi$-weighting is such that the $n$-weighting is integer,
the weighted point estimator may be regarded as being the "ordinary least squares" style estimator obtained from a situation where one of the points has been identically repeated "n" times.

Of course, if the W matrix represents new independent observations, then the repeated point estimator reduces to the O.L.S. estimator, because H becomes the null matrix, and the unity matrix J becomes the identity matrix I.

Algebraic manipulations of tedious length show that the weighted estimator is of similar form to the repeated estimator. This link between n-weighting and experimental design enables the data analyst to conceptually view the n-weightings as weakening the correlation bonds between the independent variables for a positive n-weighting, with the reverse being true for a negative weighting.

3.3. Restricted O.L.S. and Weighted Estimation With Infinite n-Weighting

By reference to Table 3.1, it is seen that a \( n \)-weighting of zero implies an n-weighting of \( +\infty \). With this extreme n-weighting, the fitted residual at that point is zero, and the possible relationship between weighted estimation, and O.L.S. estimation restricted to pass through this point must be examined.

Suppose the technique of O.L.S. estimation is used with the restriction that the fitted function has zero residual at the point \((w'y)\). It must be observed that such a restriction is of a stochastic nature as it was not formed independently of the existing data set. Denote the estimator resulting from minimizing the residual sum of squares with a restriction \( w'\hat{\beta} - y = 0 \) by \( \hat{\beta}_{(1)R} \).
Define $S$ as

$$S = (Y - X\beta)'(Y - X\beta) - 2\lambda (w'\beta - y),$$

where $\lambda$ is a Lagrangian multiplier. Differentiating $S$ with respect to $\beta$ gives

$$X'X\beta(1)_R + w\lambda - X'Y = 0,$$

and

$$\hat{\beta}(1)_R = (X'X)^{-1}X'Y - (X'X)^{-1}w\lambda. \quad (3.3.1)$$

However, the restriction is $w'\hat{\beta}(1)_R - y = 0$, and hence,

$$\lambda = \left[ \frac{1}{w'(X'X)^{-1}w} \right] (w'\hat{\beta} - y). \quad (3.3.2)$$

If the results of equation (3.3.2) are substituted into equation (3.3.1), then,

$$\hat{\beta}(1)_R = \hat{\beta} - \left[ \frac{1}{w'(X'X)^{-1}w} \right] (X'X)^{-1}w(\hat{y} - y), \quad (3.3.3)$$

where

$$\hat{y} = w'\hat{\beta}.$$  

Note that $\hat{\beta}(1)_R$ of equation (3.3.3) is unbiased for $\beta$ because the O.L.S. estimate is unbiased for $\beta$.

Equation (3.1.11) may be algebraically rearranged to give

$$\hat{\beta}(1) = \hat{\beta} + \left[ \frac{n}{1 + nw'(X'X)^{-1}w} \right] (X'X)^{-1}w(y - \hat{y}), \quad (3.3.4)$$

by use of Corollary (A.1.3) of the Appendix.
Consider the limit of \( \hat{\beta}_{(1)} \) as \( n \to \infty \):

\[
\lim_{n \to \infty} (\hat{\beta}_{(1)}) = \beta + \left[ \frac{1}{w'(X'X)^{-1}w} \right] (X'X)^{-1} w(y - \hat{y}),
\]

which is seen to be \( \hat{\beta}_{(1)R} \), from equation (3.3.3). Restricted O.L.S.

and weighted regression with infinite weighting yield the same numerical estimate. The variance of the stochastically restricted estimator is

\[
V[\hat{\beta}_{(1)R}] = E[\hat{\beta}_{(1)R} - \beta][\hat{\beta}_{(1)R} - \beta]',
\]

\[
= (X'X)^{-1} X'[E(\varepsilon \varepsilon')] X(X'X)^{-1}
\]

\[
- 2 \left[ \frac{1}{w'(X'X)^{-1}w} \right] (X'X)^{-1} w E[(\varepsilon \varepsilon') - w'(X'X)^{-1} X(\varepsilon \varepsilon')] X(X'X)^{-1}
\]

\[
+ \left[ \frac{1}{w'(X'X)^{-1}w} \right]^2 (X'X)^{-1} w E(\varepsilon e - w'(X'X)^{-1} X(\varepsilon e) - E(\varepsilon e') X(X'X)^{-1} w
\]

\[
+ w'(X'X)^{-1} X e(\varepsilon \varepsilon') X(X'X)^{-1} w w'(X'X)^{-1}.
\]

The following expressions are needed to rewrite \( V[\hat{\beta}_{(1)R}] \) in a more usable form:

\[
E(\varepsilon e) = \sigma^2,
\]

\[
E(\varepsilon \varepsilon') = I\sigma^2,
\]

and,

\[
E(\varepsilon e') = [0, 0, \ldots, 0, \sigma^2, 0, \ldots 0],
\]

\[
= \uparrow \text{ th position}
\]

\[
= h'\sigma^2.
\]
The observation $\mathbf{w}$ is linked to the $X$ matrix by $X\mathbf{h} = \mathbf{w}',$ and so the variance-covariance of $\hat{\mathbf{b}}(1)R$ is,

$$V(\hat{\mathbf{b}}(1)R) = (X'X)^{-1}\sigma^2 + \left\{ \frac{(1-w'(X'X)^{-1}w)}{[w'(X'X)^{-1}w]^2} \right\} (X'X)^{-1}w w'(X'X)^{-1} \sigma^2. \quad (3.3.5)$$

From equation (3.2.8) the variance-covariance matrix of the weighted point estimator is,

$$V(\hat{\mathbf{b}}(1)R) = [X'X + nw w']^{-1} [X'X + n(n+2)w w'] [X'X + nw w']^{-1} \sigma^2.$$

If Corollary (A.1.3) of the Appendix is applied to equation (3.2.8) and the limit of $V(\hat{\mathbf{b}}(1)R)$ considered as $n \to \infty,$ then

$$\lim_{n \to \infty} V(\hat{\mathbf{b}}(1)R) = \left[ \frac{1}{w'(X'X)^{-1}w} \right]^2 (X'X)^{-1}w w'(X'X)^{-1} + w'(X'X)^{-1}, \quad (3.3.6)$$

which, on rearrangement of terms, is seen to be the same as that obtained for the estimator $\hat{\mathbf{b}}(1)R$ in equation (3.3.3). It follows that stochastically restricted O.L.S. and infinite weighted point estimators are the same.

At first glance it would appear that as we have a form of restricted least squares, the difference in variance-covariance matrices,

$$[V(\hat{\mathbf{b}}) - V(\hat{\mathbf{b}}(1)R)],$$

is positive semi-definite. Inspection of equation (3.3.5) indicates that $[V(\hat{\mathbf{b}}) - V(\hat{\mathbf{b}}(1)R)]$ is negative semi-definite if

$$0 < w_i'(X'X)^{-1}w_i \leq 1. \quad (3.3.7)$$

Since $w_i'(X'X)^{-1}w_i$ is the $i$th element on the main diagonal of the idempotent matrix $X(X'X)^{-1}X,$ this condition is always satisfied.
The extension of these concepts to show that restricted least squares estimator for "r" restrictions, is the same estimator as the weighted estimator (for "r" points bearing infinite weighting), follows by straightforward algebra. It should be noted that the number of restrictions cannot exceed the number of independent variables.

3.4. Changing the Sign of an Estimate by Use of Weighted Estimation

When dealing with situations where near multicollinearity exists between the predictor variables, estimates of beta-coefficients with "the wrong sign" are often encountered. The "wrong sign" implies the sign of the estimated coefficient is in contradiction to what prior information, intuition, or physical constraints dictate. Weighted regression may be used to provide estimates with the "correct sign."

The technique may be regarded as having two stages; the first stage consisting of O.L.S. estimation to provide estimates of the regression coefficients, the second stage consisting of reanalyzing the problem with a point weighted such that the resulting estimator has the desired sign. The specific magnitude of the estimate has to be determined by an iterative search technique.

Suppose \( \hat{\beta}_j \) is the coefficient that has the incorrect sign. Let the \( i^{th} \) point be weighted so that the weighted estimator \( \bar{\beta}_j \) is opposite in sign to \( \hat{\beta}_j \). If \( q_j \) is the zero-vector, except for the value 1 in the \( j^{th} \) element, then the weighted estimate of \( \beta_j \) is

\[
\bar{\beta}_j = q_j^* \bar{\beta}^{(1)}.
\] (3.4.1)

If \( \hat{\beta}_j \) has the incorrect sign then
\( \beta_j \tilde{\beta}_j < 0, \) \hspace{1cm} (3.4.2)

which, from equation (3.4.1) may be expressed as,

\[ \tilde{\hat{\beta}} q_j \tilde{q}_j \tilde{\beta}(1)^- < 0. \] \hspace{1cm} (3.4.3)

From equation (3.3.4), with the \( i^{th} \) point carrying \( n^- \)-weighting \( n_i \),

\[ \tilde{\beta}(1) = \hat{\beta} + \left[ \frac{n_i}{1 + n_i w_1' (X'X)^{-1} w_1} \right] (X'X)^{-1} w_1 (y_1 - \hat{y}_1), \]

and substituting for \( \tilde{\beta}(1) \) in equation (3.4.3) yields the condition for change in sign, namely

\[ \frac{\hat{\beta} q_j \tilde{q}_j \hat{\beta} + n_i (\hat{\beta} q_j \tilde{q}_j \hat{\beta}(w_1' (X'X)^{-1} w_1) + \hat{\beta} q_j \tilde{q}_j (X'X)^{-1} w_1 (y_1 - \hat{y}_1))}{(1 + n_i w_1' (X'X)^{-1} w_1)} < 0, \]

or,

\[ n_i < \frac{\hat{\beta}_j^2}{\hat{\beta}_j w_1' (X'X)^{-1} w_1 + \hat{\beta}_j q_j (X'X)^{-1} w_1 (y_1 - \hat{y}_1)}, \] \hspace{1cm} (3.4.4)

because

\[ 1 + n_i w_1' (X'X)^{-1} w_1 > 0. \]

Therefore, weighting the observation "i" by any \( n^- \)-weighting less than the right hand side of equation (3.4.4), will result in the \( j^{th} \) weighted estimate being opposite in sign to the \( j^{th} \) O.L.S. estimate. There is no guarantee that the condition in equation (3.4.4) can be attained, as feasible \( n^- \)-weighting requires all \( n^- \)-weightings larger than -1.
As a specific value is not required for the weighted estimate, it follows that the principal use of this technique will be in conjunction with other criteria. Section 5.6 of Chapter 5 discusses this aspect in greater depth.
4. WEIGHTED RIDGE REGRESSION

The ridge regression technique of Hoerl and Kennard (1970) will be applied to weighted point and weighted estimation. The main properties of the ridge technique have been discussed in Section 2.5 of Chapter 2. First, the basic properties of the weighted point and weighted estimators are derived as these properties will be used repeatedly in comparing weighted ridge estimators with other estimators. These relationships are investigated in Section 4.2 with emphasis given to the comparison of weighted ridge estimators with the ridge estimator. Section 4.3 investigates the properties of estimators derived by fixing the \( n \)-weightings, and allowing the \( k \)-value to vary, this being "the reverse" of the usual ridge-procedure. The problem of comparing bias-squared for two estimators is examined in Section 4.5 where a geometric argument is proposed and examined. Finally, in Section 4.6, the sequential ridge estimator is investigated. This sequential ridge estimator will be used in the calculation of specific \( n \)-weightings in Chapter 5.

4.1. Basic Properties of Weighted Ridge Regression

The weighted point ridge regression estimator is defined as

\[
\hat{\beta}^*_{(1)} = (X'\Omega^{-1}_1X + kI)^{-1}X'\Omega^{-1}_1Y,
\]

which may be expressed as,

\[
\hat{\beta}^*_{(1)} = (X'X + nww' + kI)^{-1}(X'Y + nwy),
\]

by use of equation (3.1.11).
Similarly, the weighted ridge regression estimator is defined as

\[
\hat{\beta}^* = (X'\Omega^{-1}_0 X + kI)^{-1} X'\Omega^{-1}_0 y,
\]  

and may be expressed as,

\[
\hat{\beta}^* = (X'X + \sum_{i=1}^{n_0} n_i w_i w_i' + kI)^{-1} (X'y + \sum_{i=1}^{n_0} n_i w_i y_i),
\]

by use of equation (3.2.21).

Since

\[
(X'\Omega^{-1}_1 X)\tilde{\beta}(1) = X'\Omega^{-1}_1 y,
\]

and

\[
(X'\Omega^{-1}_1 X + kI)^{-1} X'\Omega^{-1}_1 X = I - k(X'\Omega^{-1}_1 X + kI)^{-1},
\]
equation (4.1.1) may be rewritten as,

\[
\tilde{\beta}^*(1) = \tilde{\beta}(1) - k(X'\Omega^{-1}_1 X + kI)^{-1}\tilde{\beta}(1).
\]  

The variance of \(\tilde{\beta}^*(1)\) is given by

\[
V(\tilde{\beta}^*(1)) = \{(X'\Omega^{-1}_1 X + kI)^{-1} X'\Omega^{-1}_1 X\} V(\tilde{\beta}(1)) \{(X'\Omega^{-1}_1 X + kI)^{-1} X'\Omega^{-1}_1 X\}^*,
\]  

and by use of equation (3.1.8) gives,

\[
V(\tilde{\beta}^*(1)) = (X'\Omega^{-1}_1 X + kI)^{-1} X'\Omega^{-1}_1 X (X'\Omega^{-1}_1 X + kI)^{-1}\sigma^2.
\]  

Equations (3.2.8) and (3.1.12), permit (4.1.7) to be rewritten as

\[
V(\tilde{\beta}^*(1)) = [(X'X + kI) + n w w']^{-1} [(X'X + n(2) w w')[(X'X + kI) + n w w']^{-1} \sigma^2.
\]  

The bias-matrix of \(\tilde{\beta}^*(1)\) is defined as
\[ B(\hat{\beta}^*_1) = \left[ E(\hat{\beta}^*_1) - \beta \right] \left[ E(\hat{\beta}^*_1) - \beta \right]^\top. \]  

Substitution for \( \hat{\beta}^*_1 \) from equation (4.1.5) in equation (4.1.9) yields,

\[ B(\hat{\beta}^*_1) = k^2 (X^\top \Omega^{-1}_1 X + kI)^{-1} \hat{\beta}^* \left( X^\top \Omega^{-1}_1 X + kI \right)^{-1}, \]

which may be rewritten as,

\[ B(\hat{\beta}^*_1) = k^2 \left[ (X^\top X + kI) + n_0 \hat{\beta}^* \right] \left[ (X^\top X + kI) + n_0 \hat{\beta}^* \right]^{-1} \]

For the weighted ridge regression estimator, the variance is

\[ V(\tilde{\beta}^*) = \left[ X^\top (X^\top X + kI) X \right]^{-1} \sum_{i=1}^{n_0} \left( \frac{n_0}{n_1} \right) \left( \frac{n_1}{n_1 + n_2} \right) \sigma_i^2 \]

and the bias-matrix

\[ B(\tilde{\beta}^*) = k^2 \left[ X^\top (X^\top X + kI) X \right]^{-1} \hat{\beta}^* \left[ X^\top (X^\top X + kI) + \sum_{i=1}^{n_0} n_i \sigma_i^2 \right] \]

One of the reasons an analyst might be willing to abandon O.L.S. in favor of weighted least squares is that the O.L.S. solution is "too extreme," in the sense that the Euclidean length of the vector is unacceptably large. This expected Euclidean squared-distance from the weighted estimator to the true parameter vector is equivalent to the trace of the variance matrix. By the Gauss-Markov theorem, O.L.S. is the best linear unbiased estimator, and consequently there is no weighting for which the expected Euclidean length of \( \tilde{\beta} \) is less than that.
of \(\hat{\beta}\). Hoerl and Kennard (1970) showed the Euclidean length of the ridge regression estimator is smaller than that for the O.L.S. estimator, and a similar statement may be made for the comparison of the weighted ridge estimator with the weighted estimator. The Euclidean strength length of the estimator is \(\| \tilde{\beta}^* \|_2\), and the positive square root of this quantity is the Euclidean norm \(\| \tilde{\beta}^* \|\). By definition of weighted ridge regression

\[
\tilde{\beta}^* + (\tilde{X}^\top \tilde{X} + kI)^{-1}\tilde{X}^\top \tilde{Y},
\]

and may be expressed as,

\[
\tilde{\beta}^* = (\tilde{X}^\top \tilde{X} + kI)^{-1}\tilde{X}^\top \tilde{\beta},
\]

\[
= \tilde{M}\tilde{\beta}.
\]

The Euclidean norm of \(\tilde{\beta}^*\) in equation (4.1.14) is

\[
\| \tilde{\beta}^* \| = \| \tilde{M}\tilde{\beta} \|,
\]

\[
\leq \| \tilde{M} \| \cdot \| \tilde{\beta} \|. \quad (4.1.15)
\]

The Euclidean norm of the matrix \(\tilde{M}\) is

\[
\| \tilde{M} \| = \sqrt{\frac{\lambda_1^2}{\tilde{M}_1} + \ldots + \frac{\lambda_p^2}{\tilde{M}_p}},
\]

where

\[
\lambda_{\tilde{M}_j} = \frac{\lambda_{\tilde{j}}}{\tilde{\lambda}_j + k}.
\]

It therefore follows that \(\| \tilde{\beta}^* \| < \| \tilde{\beta} \|\) for positive \(k\) value. This indicates that the use of the ridge technique with weighted least squares will always result in an estimator that has smaller Euclidean
length than the initial weighted estimator. The Euclidean length of weighted ridge regression when compared with the Euclidean length of ridge regression, is a function of the specific \( n \)-weightings used. The implication is that it is possible for the weighted ridge estimator to be larger (in Euclidean norm) than ridge regression estimator, which may not be desirable from the data analyst's point of view.

In the development of criteria for selection of \( n \)-weightings it is of importance to know the residual sum of squares. For any estimator (call it \( \hat{\beta}^+ \)), the residual sum of squares is defined as

\[
\phi^+ = (\tilde{Y} - X\hat{\beta}^+)'(\tilde{Y} - X\hat{\beta}^+). \tag{4.1.16}
\]

Equation (4.1.16) may be written as a difference of two parts, one a "total sum of squares," and the other as a "regression sum of squares."

For weighted regression,

\[
\hat{\phi} = \tilde{Y}'\tilde{Y} - \tilde{\hat{\beta}} 'X'X(2\tilde{\hat{\beta}} - \tilde{\hat{\beta}}), \tag{4.1.17}
\]

and for weighted ridge regression,

\[
\hat{\phi}^* = \tilde{Y}'\tilde{Y} - \tilde{\hat{\beta}} 'X'X(2\tilde{\hat{\beta}} - \tilde{\hat{\beta}}^*). \tag{4.1.18}
\]

If the relationship,

\[
\tilde{\hat{\beta}}^* = (I - kF)\tilde{\hat{\beta}},
\]

\[
= \tilde{\hat{\beta}} - kF\tilde{\hat{\beta}},
\]

is used with equation (4.1.18) then,

\[
\hat{\phi}^* = \hat{\phi} + 2k\tilde{\hat{\beta}} 'FX'(\tilde{\hat{\beta}} - \tilde{\hat{\beta}}) + k^2\tilde{\hat{\beta}} 'FX'X\tilde{\hat{\beta}}. \tag{4.1.19}
\]
If all the n-weightings are zero, the ridge regression residual sum of squares is,

$$\phi^* = \phi + k^2 \beta^* \text{FX} \beta^*.$$  \hspace{1cm} (4.1.20)

These expressions for residual sum of squares are used in Section 5.4 of Chapter 5, for the determination of specific n-weightings.

4.2. The Relationship of Weighted Ridge Estimators to Other Estimators.

The relationship between weighted point ridge estimation and ridge estimation can be expressed as

$$\hat{\beta}^* = \hat{\beta}^* + \left[ \frac{n}{\text{FW}} \right] \text{FW} (y - y^*),$$  \hspace{1cm} (4.2.1)

where

$$F = (X^T X + kI)^{-1},$$

$$y^* = W^* \hat{\beta}^*,$$

and

$$\hat{\beta}^* = FX^T Y.$$  

The derivation of equation (4.2.1) follows that of equation (3.3.4) with \((X^T X)^{-1}\) replaced by \(F\), and \(\hat{y}\) replaced by \(y^*\). Weighted ridge regression can also be related to standard ridge regression by writing

$$[\left( X^T X + kI \right) + \sum_{i=1}^{n} n_i W_i W_i^T]^{-1}$$

as

$$F - \left[ \left( X^T X + kI \right) + \sum_{i=1}^{n} n_i W_i W_i^T \right]^{-1} \left[ \sum_{i=1}^{n} n_i W_i W_i^T \right] F,$$
and substitution in equation (4.1.4) to obtain,
\[
\tilde{\hat{\beta}}^* = \left( I + F \left( \sum_{i=1}^{n_0} n_i \tilde{w}_i \tilde{w}_i' \right) \right)^{-1} \hat{\beta}^*
+ \left( \tilde{X}' \tilde{X} + kI \right) + \sum_{i=1}^{n_0} n_i \tilde{w}_i \tilde{w}_i' \right]^{-1} \left[ \sum_{i=1}^{n_0} n_i \tilde{w}_i \tilde{w}_i' \right] \cdot (4.2.2)
\]

Consider the comparison of the variance-covariance matrix of the weighted point ridge estimator, and that of the weighted point estimator.

Rearrangement of terms in equation (4.1.8) yields,
\[
V(\tilde{\hat{\beta}}^*) = [\tilde{X}' \tilde{X} + kI]^{-1} [\tilde{X}' \tilde{X} + n(n+1)ww'] [\tilde{X}' \tilde{X} + kI]^{-1} \sigma^2
= [I + k(\tilde{X}' \tilde{X})^{-1}]^{-1} V(\tilde{\hat{\beta}}) [I + k(\tilde{X}' \tilde{X})^{-1}]^{-1} \cdot (4.2.3)
\]

Let the eigen-values of \(V(\tilde{\hat{\beta}})\) be denoted by \(\tilde{\lambda}_{v1} \sigma^2 \geq \ldots \geq \tilde{\lambda}_{vp} \sigma^2\). Since the matrices \([I + k(\tilde{X}' \tilde{X})^{-1}]^{-1}\) and \(V(\tilde{\hat{\beta}})\) commute, the eigen-values of \(V(\tilde{\hat{\beta}}^*)\), denoted by \(\tilde{\lambda}_{v1}^* \sigma^2\), may be written as,
\[
\tilde{\lambda}^*_{v1} = \frac{\tilde{\lambda}_{v1}}{(1 + k\tilde{\lambda}_i^{-1})^2} \cdot (4.2.4)
\]

where \(\tilde{\lambda}_i\) is the \(i^{th}\) eigen-value of \(\tilde{X}' \tilde{X}\).

The eigen-values \(\tilde{\lambda}_{v1}, \tilde{\lambda}_i\) are both positive and so for positive \(k\) it follows that,
\[
\tilde{\lambda}^*_{v1} < \tilde{\lambda}_{v1} \cdot (4.2.5)
\]

implying that \([V(\tilde{\hat{\beta}}(1)) - V(\tilde{\hat{\beta}}^*)]\) is positive definite. A natural extension of equation (4.2.5) is that the difference \(V(\tilde{\hat{\beta}}) - V(\tilde{\hat{\beta}}^*)\) is positive definite.
The weighted ridge technique will be considered as being an improvement over weighted least squares if there exists a k-value such that the difference between the variance of the weighted estimator and the mean square error of the weighted ridge estimator is positive definite.

Let the canonical forms be defined as,

\[
P\tilde{P} = X'X + \sum_{i=1}^{n_0} n_i w_i w_i',
\]

\[
P_k\tilde{P} = X'X + \sum_{i=1}^{n_0} n_i w_i w_i' + kI,
\]

\[
\tilde{P} = \sum_{i=1}^{n_0} n_i (n_i + 1) w_i w_i',
\]

where,

\(\tilde{P}\) is orthogonal,

\(\tilde{\Lambda}\) is diagonal, bearing the eigenvalues of \(X'X\),

\(\tilde{\Lambda}_k = \tilde{\Lambda} + kI\),

\(\tilde{\omega} = \tilde{P}'\tilde{b}\).

The variance of the weighted estimator may be written as

\[
V(\tilde{b}) = \tilde{P}^{-1}(\tilde{\Lambda} + \tilde{Q})\tilde{P}^{-1}\sigma^2. \tag{4.2.6}
\]

The mean square error of the weighted ridge estimator may be written as

\[
MSE(\tilde{b}^*) = \tilde{P}^{-1}(\tilde{\Lambda}\sigma^2 + \tilde{Q}\sigma^2 + k^2 \tilde{\omega}^{-1})\tilde{P}^{-1}. \tag{4.2.7}
\]

The difference between the variance of \(\tilde{b}\) and mean square error of \(\tilde{b}^*\) may then be written as,
\[
D = \sum_k \sum_{\lambda_k} \frac{1}{\tilde{\lambda}_k} \tilde{\lambda}_k^{-1} (\tilde{\lambda} + \tilde{Q}) \tilde{\lambda}^{-1} \lambda_k \sigma^2 - (\tilde{\lambda} + \tilde{Q}) \sigma^2 - k^2 \tilde{\alpha} \tilde{\alpha} \tilde{\lambda}^{-1} \lambda_k p^2. \tag{4.2.8}
\]

where \(D\) is positive semi-definite as the kernel matrix of equation (4.2.8) is positive semi-definite.

Consider any real, non-null vector \(\xi\) and define \(G(k)\) to be
\[
G(k) = \xi^{\top} \left[ \tilde{\lambda}_k \lambda^{-1} (\tilde{\lambda} + \tilde{Q}) \lambda^{-1} \tilde{\lambda}_k \sigma^2 - (\tilde{\lambda} + \tilde{Q}) \sigma^2 - k^2 \tilde{\alpha} \tilde{\alpha} \right] \xi, \tag{4.2.9}
\]

which may be written as,
\[
G(k) = \sum_{i=1}^{p} \left\{ \frac{k^2 + 2k\tilde{\lambda}_i}{\tilde{\lambda}_i^2} \right\} \xi_i^2 \sigma^2 + \sum_{i=1}^{p} \sum_{i^*=1}^{p} \frac{\xi_i \xi_{i^*} \left[ (2k\lambda_i + k^2)[2k\lambda_{i^*} + k^2] \tilde{\lambda}_i \tilde{\lambda}_{i^*} \tilde{\lambda}_i \tilde{\lambda}_{i^*} \right]}{\tilde{\lambda}_i \tilde{\lambda}_{i^*}} \tag{4.2.10}
\]

For the O.L.S. case when \(k = 0\),
\[
G(0) = \sum_{i=1}^{p} \frac{2\xi_i^2 \sigma^2}{\tilde{\lambda}_i^2}, \tag{4.2.11}
\]

and is positive for all values of \(\tilde{\lambda}_i\) and is continuous in the neighborhood of \(k = 0\). If a positive \(k\)-value is taken such that \(G(k) > 0\) for any \(\xi\), then \(\text{MSE}(\tilde{\xi}^*_k) < \text{MSE}(\tilde{\xi})\), and therefore a positive \(k\)-value does exist such that the mean square error of weighted ridge regression is less than that of weighted regression.

Suppose weighted point ridge regression is compared with standard ridge regression. Equation (4.1.8) may be written in the form,
\[
\text{V}(\tilde{\xi}(1)) = \left[ F - \left( \frac{n}{1+n\tilde{\omega} \tilde{F}_w} \right) F \tilde{F}_w^{-1} \right] \left[ X'X + n(n+2)\tilde{\omega} \tilde{\omega} \right] \left[ F - \left( \frac{n}{1+n\tilde{\omega} \tilde{F}_w} \right) F \tilde{F}_w^{-1} \right] \sigma^2, \tag{4.2.12}
\]

where \(F = (X'X + kI)^{-1}\).
On rearrangement of terms, equation (4.2.13) may be written as,

\[
V(\tilde{\beta}^*(1)) = V(\hat{\beta}^*) + \left\{ \frac{\sigma^2}{(1+nw^T\bar{F}w)^2} \right\} \{kan + En^2\}, \quad (4.2.13)
\]

where,

\[
V(\hat{\beta}^*) = FX^T\bar{X}\bar{F}^2,
\]

\[
A = F_{ww}^TF^2 + F_{ww}^TF,
\]

and

\[
B = k(w^TFw)A + [1 - w^TFw - kw^TF_{ww}^TF]F_{ww}^TF. \quad (4.2.14)
\]

Since \( \bar{F} \) is positive definite, it follows that \( w^TFw \) is positive. As \( F_{ww}^TF \) is positive semi-definite, of rank one, it follows that, \( F_{ww}^TF \), \( F_{ww}^TF^2 \) and \( F_{ww}^TF \) are all positive semi-definite.

From equation (4.2.14), the term

\[
[1 - w^TFw - kw^TF_{ww}^TF] = [1 - w^T(X^TX)^{-1}w] + w^T[(X^TX)^{-1} - F - kF^2]w,
\]

\[
= [1 - w^T(X^TX)^{-1}w] + w^TF[(X^TX)^{-1}F^{-1}F^{-1} - F^{-1} - kI]w,
\]

\[
= [1 - w^T(X^TX)^{-1}w] + k^2w^TF(X^TX)^{-1}Fw,
\]

is always positive.

It follows that \( V(\tilde{\beta}^*(1)) - V(\hat{\beta}^*) \) is positive semi-definite for choice of a positive \( n \)-weighting. This implies that ridge regression cannot be improved upon by positive \( n \)-weighting. If negative \( n \)-weighting is used, it is possible to achieve an improvement.

The bias-matrices of weighted ridge regression and standard ridge regression may be compared in a similar fashion. Equation (4.1.11) may be written as
\[ B(\tilde{\beta}^*) = k^2 F_1 \tilde{\beta} \tilde{\beta}^\top F_1, \quad (4.2.15) \]

where

\[ F_1 = (X'X + kI + nww'w)^{-1}. \]

The bias-squared for ridge regression is,

\[ B(\tilde{\beta}) = k^2 (X'X + kI)^{-1} \tilde{\beta} \tilde{\beta}^\top (X'X + kI)^{-1}, \]

\[ = k^2 F_1 \tilde{\beta} \tilde{\beta}^\top F. \quad (4.2.16) \]

\[ B(\tilde{\beta}^*) = k^2 F_1 F_1^{-1} \tilde{\beta} \tilde{\beta}^\top F_1^{-1} F, \]

\[ = FF_1^{-1} \{ B(\tilde{\beta}^*) \} F_1^{-1} F. \quad (4.2.17) \]

Now,

\[ FF_1^{-1} = I + nww'w', \]

and so equation (4.2.17) may be expressed as,

\[ B(\tilde{\beta}) = B(\tilde{\beta}^*) + nww'w' \{ B(\tilde{\beta}^*) \} + n \{ B(\tilde{\beta}) \} ww'w'F + n^2 Fww'w' \{ B(\tilde{\beta}^*) \} ww'w'. \quad (4.2.18) \]

In equation (4.2.18), \( Fww' \) and \( B(\tilde{\beta}^*) \) are at least positive semi-definite, giving \( B(\tilde{\beta}) - B(\tilde{\beta}^*) \) to be positive semi-definite for \( n > 0. \)

4.3. The Changing of the k-Value for Fixed n-Weighting.

In previous sections of this chapter, it has been assumed the k-value has been fixed and the effect of weighting the observations considered.
The analyst would want to investigate the "reverse" sequence i.e., fixing a set of \( n \)-weightings and then altering the \( k \)-value.

From equation (4.1.12) the variance-covariance matrix of the weighted ridge estimator may be written as,

\[
V[\tilde{\hat{\beta}}^*] = \tilde{F}[X'X + \sum_{i=1}^{n_0} n_i (n_i + 2) w_i w_i'] \tilde{F} \sigma^2,
\]

(4.3.1)

where

\[
\tilde{F} = [(X'X + k_1 I) + \sum_{i=1}^{n_0} n_i w_i w_i']^{-1}.
\]

(4.3.2)

For two different \( k \)-values (\( k_1 \) and \( k_2 \)) equation (4.3.2) may be written as

\[
\tilde{F}(j) = [(X'X + k_j I) + \sum_{i=1}^{n_0} n_i w_i w_i']^{-1}, \quad j=1,2,\ldots
\]

and \([X'X + \sum_{i=1}^{n_0} n_i (n_i + 2) w_i w_i']\) denoted by \( G \).

the variance-covariance matrix of \( \tilde{\hat{\beta}}(k_2) \) by

\[
V[\tilde{\hat{\beta}}(k_2)] = \tilde{F}(1) G \tilde{F}(2) \sigma^2
\]

\[
= \tilde{F}(1) \tilde{F}^{-1}(2) G \tilde{F}(2) \tilde{F}^{-1}(1) \sigma^2,
\]

\[
= \tilde{F}(1) \tilde{F}^{-1}(2) [V[\tilde{\hat{\beta}}(k_2)]] \tilde{F}(2) \tilde{F}^{-1}(1).
\]

(4.3.3)

However,

\[
\tilde{F}(1) \tilde{F}^{-1}(2) = \tilde{F}(1) [X'X + \sum_{i=1}^{n_0} n_i w_i w_i'] + k_1 I + (k_2 - k_1) I,
\]

\[
= I + (k_2 - k_1) \tilde{F}(1),
\]

(4.3.4)

and so,
\[ V \tilde{\beta}^*(k_1) = V[\tilde{\beta}^*(k_2)] + (k_2-k_1)\tilde{F}(1)[V[\tilde{\beta}(k_2)]] + (k_2-k_1)[V[\tilde{\beta}(k_2)]]\tilde{F}(1) \]
\[ + (k_2-k_1)^2\tilde{F}(1)[V[\tilde{\beta}^*(k_2)]]\tilde{F}(1). \] (4.3.5)

If \((k_2 - k_1)\) is positive, then each matrix component of equation (4.3.5) is positive definite. It follows that the increasing of the \(k\)-value from \(k_1\) to \(k_2\) will result in \(V - (k_1) - V - (k_2)\), being positive definite, which may be regarded as an "improvement" in variance properties of the resulting estimator.

From equation (4.1.13) the bias squared is

\[ B(\tilde{\beta}^*) = k^2\tilde{F}\tilde{\beta}\tilde{\beta}^*\tilde{F}. \]

For \(k = k_1\),

\[ B(\tilde{\beta}^*(k_1)) = k_1^2\tilde{F}(1)\tilde{\beta}\tilde{\beta}^*\tilde{F}(1), \]

and for \(k = k_2\),

\[ B(\tilde{\beta}^*(k_2)) = k_2^2\tilde{F}(2)\tilde{\beta}\tilde{\beta}^*\tilde{F}(2). \]

Now,

\[ \tilde{F}(2)[\tilde{X}\tilde{X} + \sum_{i=1}^{n} n_{i-1}w_i w_i - k_2 I] = I, \]

i.e.,

\[ k_2\tilde{F}(2) = I - \tilde{F}(2)\tilde{X}\tilde{X}, \]

and so the bias-squared matrix may be expressed as,

\[ B(\tilde{\beta}^*(k_2)) = (I - \tilde{F}(2)\tilde{X}\tilde{X})\tilde{\beta}\tilde{\beta}^*(I - \tilde{F}(2)\tilde{X}\tilde{X})^*. \]

Use of the relation,
\((I - \tilde{F}_2)(X^\top X) = \tilde{F}_2(1)(X^\top X) + (\tilde{F}_1)(X^\top X) = \tilde{F}_2(2)(X^\top X)\),

results in,

\[
B(\tilde{F}_{k_2}^\ast) = B(\tilde{F}_{k_1}^\ast) + [I - \tilde{F}_2(1)(X^\top X)]BB^{-1}[X^\top X(\tilde{F}_2(1) - \tilde{F}_2(2))]
\]
\[
+ [(\tilde{F}_1 - \tilde{F}_2)(X^\top X)]BB^{-1}[I - X^\top XF(1) - X^\top XF(1)]
\]
\[
+ [(\tilde{F}_1 - \tilde{F}_2)(X^\top X)]BB^{-1}[\tilde{F}_1 - \tilde{F}_2].
\]

(4.3.6)

The difference \(\tilde{F}_1 - \tilde{F}_2\) may be written as,

\[
\tilde{F}_1 - \tilde{F}_2 = \left[X^\top X + \sum_{i=1}^{n_o} n_1 w_1 w_1' + k_1 I\right]^{-1} - \left[X^\top X + \sum_{i=1}^{n_o} n_1 w_1 w_1' + k_2 I\right]^{-1},
\]

i.e.,

\[
\tilde{F}_1 - \tilde{F}_2 = \tilde{F}[\tilde{\Lambda}_{k_1} - \tilde{\Lambda}_{k_2}]
\]

where

\[
\tilde{F}[X^\top X + \sum_{i=1}^{n_o} n_1 w_1 w_1']\tilde{P}^{-1} = \tilde{\Lambda}.
\]

Observe that \(\tilde{\Lambda}_{k_1} - \tilde{\Lambda}_{k_2}\) has \(i^{th}\) diagonal element,

\[
\begin{bmatrix}
\frac{1}{\tilde{\Lambda}_i + k_1} - \frac{1}{\tilde{\Lambda}_i + k_2}
\end{bmatrix},
\]

(4.3.7)

and for \(k_2 > k_1\) is clearly > 0. It follows that \(\tilde{F}_1 - \tilde{F}_2\) is positive definite, implying that all the matrix-terms of equation (4.3.6) are at least positive semi-definite, and therefore that \(B(\tilde{F}_{k_2}^\ast) - B(\tilde{F}_{k_1}^\ast)\) is positive definite.
Equations (4.3.1) to (4.3.7) show that the weighted estimation system acts similarly to the O.L.S. system when the ridge-technique is applied.

4.4. Bias-Squared Comparisons for Different Estimators

As discussed in Section 4.2 of Chapter 4, the use of positive n-weightings reduces bias-squared at the expense of variance. The use of negative n-weightings will reduce variance but at a possible increase in bias-squared. Consider the case where all negative n-weightings have been used and the objective is to minimize the trace of the variance-covariance matrix. The decrease in trace of the variance-covariance matrix matrix could have been obtained if a new $k$-value, $k_*(k = k)$, had been used with standard ridge regression. The comparison of bias-squareds of the two possible estimators that yield equal variance reduction is now of importance.

Let $k_* = k + d$ denote the $k$-variable such that,

$$\left.\text{tr}[V(\hat{\beta}^*)]\right|_{k_*} = \left.\text{tr}[V(\tilde{\beta}^*)]\right|_{k}.$$

From equation (4.1.12) the trace of the variance-covariance matrix may be written,

$$\text{tr}[V(\hat{\beta}^*')] = \sigma^2. \quad (4.4.1)$$

For ridge regression with a $k$-value of $k_* = k + d$,

$$\left.\text{tr}[V(\tilde{\beta}^*)]\right|_{k_*} = \text{tr}[(X'X + kI + dI)^{-1}X'X(X'X + kI + dI)^{-1}]\sigma^2,$$

$$\quad (4.4.2)$$

Equation (4.4.2) may be expressed in canonical form and equated to equation (4.6.1), giving
\[ c = \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k + d)^2}, \quad (4.4.3) \]

which is then solved for \( d \).

For the bias-squared, the difference between bias-matrices may be expressed as

\[
\text{tr}[B(\beta^*)]_k - \text{tr}[B(\hat{\beta})]_k = (k+d)^2 \beta^*[(X^TX + kI) + dI]^{-2} \beta
\]
\[ - k^2 \beta^*[(X^TX + kI) + X^TDX]^{-2} \beta, \quad (4.4.4) \]

or

\[
Z_k = \beta^*[(k+d)^2[(X^TX + kI) + dI]^{-2} - k^2[(X^TX + kI) + X^TDX]^{-2}] \beta. \quad (4.4.5) \]

The kernel of equation (4.4.5) must be positive definite for the bias-squared of weighted ridge regression to be less than that of standard ridge regression. Equation (4.4.5) may be written as

\[
Z_k = \beta^*[B^{-1} - A^{-1}] \beta, \quad (4.4.6) \]

where,

\[
A = k^{-2}[(X^TX + kI) + X^TDX]^2, \]
\[
B = (k+d)^{-2}[(X^TX + kI) + dI]^2, \quad (4.4.7) \]

and \( D \) is diagonal, with \( i^{th} \) element equal to the weighting \( n_i \).

The matrices \( A \) and \( B \) are both symmetric and positive definite, and if \( A - B \) is positive definite, then \( B^{-1} - A^{-1} \) will also be positive definite (Graybill, 1969).

The matrix \( A - B \) is,
\[ k^{-2}(X'X + kI)^2 + k^{-2}(X'X + kI)X'DX + k^{-2}X'DX(X'X + kI) \\
+ k^{-2}(X'DX)(X'DX) - (k+d)^{-2}(X'X + kI)^2 \\
- 2(k+d)^{-2}d(X'X + kI) - (k + d)^{-2}d^2. \]  

(4.4.8)

Rearrangement of terms results in the difference

\[ A - B = \left\{ \frac{2kd + d^2}{k^2(k+d)^2} \right\} \{ (X'X + kI)^2 \} + W_1(X'X+kI) + (X'X+kI)W_1 + W_2, \]

(4.4.9)

where,

\[ W_1 = \frac{1}{k^2} \sum_{i=1}^{n_0} w_i w_i' \left[ -\frac{d}{(k+d)^2} \right] I, \]

(4.4.10)

and

\[ W_2 = \frac{1}{k^2} \left[ \sum_{i=1}^{n_0} w_i w_i' \right]^2 \left[ -\frac{d}{k+d} \right]^2 I. \]

(4.4.11)

If equation (4.4.9) is positive definite, then \( z_k \) of equation (4.4.6) will be positive, implying the bias-squared of weighted ridge regression is less than that for \( k \)-ridge regression for any value of \( k \).

### 4.5. The Comparison of Bias-Squareds by a Geometric Argument

There are many instances where equation (4.4.9) is neither positive definite or negative definite, but of indeterminant form. It is then difficult to determine the behavior of bias-squared for the two competing estimators. A technique of comparing bias-squared via a geometric argument is now proposed.

Consider a biased estimator of \( \hat{\theta} \) that has a bias-matrix \( G\hat{\theta}G' \). The trace of this bias-matrix is
\[
\Delta = \text{tr} G \bar{G} G',
\]
\[
= \bar{G} G' \bar{G}.
\]  \hspace{1cm} (4.5.1)

Let Q be an orthogonal matrix such that
\[
\Delta = \bar{G} Q D g Q' \bar{G},
\]
\[
= \gamma' D \gamma,
\]  \hspace{1cm} (4.5.2)

where \( D \) is diagonal with the diagonal element equal to eigenvalues
of \( GG' \), i.e., \( \lambda_1^2, \lambda_2^2, \ldots, \lambda_p^2 \).

Equation (4.5.2) may be expressed as,
\[
\begin{bmatrix}
\frac{\Delta}{\gamma_1^2} \\
\frac{\lambda_1^2}{g_1}
\end{bmatrix} + \begin{bmatrix}
\frac{\Delta}{\gamma_2^2} \\
\frac{\lambda_2^2}{g_2}
\end{bmatrix} + \ldots + \begin{bmatrix}
\frac{\Delta}{\gamma_p^2} \\
\frac{\lambda_p^2}{g_p}
\end{bmatrix} = 1.
\]  \hspace{1cm} (4.5.3)

This equation (4.5.3) is that of an ellipsoid in \( p \)-dimensions, with
the length of the \( i \)th semi-axis,
\[
a_i = \left[ \frac{\Delta}{\lambda_i^2} \right]^\frac{1}{2}
\]  \hspace{1cm} (4.5.4)

From elementary geometry, the volume of an ellipsoid in \( p \)-dimensions
is given by
\[
\text{volume} = \left[ \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2} + 1\right)} \right] \left[ \sum_{i=1}^{p} a_i \right].
\]  \hspace{1cm} (4.5.5)

Applying equation (4.5.4) to equation (4.5.5) gives
volume = \[ \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2} + 1\right)} \cdot \frac{\Delta^{p/2}}{g_1^\lambda g_2^\lambda \ldots g_p^\lambda} \] \quad (4.5.6)

However,

\[ \frac{\lambda}{g_1^\lambda g_2^\lambda \ldots g_p^\lambda} = |G|, \]

and substitution into equation (4.5.6) yields

volume = \[ \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2} + 1\right)} \cdot \frac{\Delta^{p/2}}{|G|} \] \quad (4.5.7)

From equation (4.5.7) it follows that the bias-squared of an estimator may be related to the volume of an ellipsoid centered on zero. This concept may now be applied to the case where the comparison of two competing estimators resulted in a difference matrix whose eigenvalues were indefinite.

As a geometric argument is being used it is instructive to look at the behavior of bias-squareds for the simple case of two dimensions. One estimator has a bias-squared \( \Delta_\star \) and the other estimator a bias-squared \( \Delta_- \). Figure 4.1 shows the case where \( \Delta_- \) is clearly smaller than \( \Delta_\star \) and so the eigenvalues of the difference in bias-matrices are all positive. (Figure 4.2 shows \( \Delta_- \) is smaller than \( \Delta_\star \) but, due to the intersection of ellipses, the eigenvalues of the difference in bias-matrices being indeterminant and thus obscuring \( \Delta_- < \Delta_\star \). If the ellipse generated by \( \Delta_- \) is rotated such that its axes coincide with the axes of \( \Delta_\star \) the difference in bias-squareds becomes clear. The theoretical extension to the volume of an ellipsoid follows immediately.
Figure 4.1. The difference in bias-squareds resulting in all positive eigenvalues.

Figure 4.2. The difference in bias-squareds resulting in indeterminant eigenvalues due to different orientation of axes.
Define $\Delta_{\text{diff}} = d_i^{-1} - d_i^{*-1}$, and writing equation (4.5.11) in canonical form yields

$$\Delta_{\text{diff}} = (1 + k^{-1}\lambda_1)^2 - (1 + k_*^{-1}\lambda_1)^2,$$

(4.5.12)

then $\Delta_{\text{diff}}$ must be positive for all values of $i$, $(i=1,2,\ldots,p)$. Because $\lambda_1$ and $\tilde{\lambda}_1$ are positive, $\Delta_{\text{diff}}$ of equation (4.5.12) may be expressed as,

$$\Delta_{\text{diff}} = k_*\tilde{\lambda}_1 - k\lambda_1.$$

(4.5.13)

and so the condition for $\Delta_*$ to be larger than $\Delta_-$ is simply that $\Delta_{\text{diff}}$ of equation (4.5.13) must be positive.

There are instances, however, where the rotation to equivalent axes still results in a difference-matrix whose roots have indeterminant form. A further definition for comparison of bias-squareds is necessary. "Equivalent bias-squared" is defined to be a spheroid, centered on zero, of the same volume as the former ellipsoid. This is illustrated in Figure 4.3 where the intersection of ellipses has obscured the fact that $\Delta_* > \Delta_-$. Conversion to equivalent bias-squareds of Figure 4.4 clearly reveals $\Delta_* > \Delta_-$. Equivalent bias-squared, in the form of a spheroid, implies that all the semi-axes have the same length. The equivalent bias-squared of equation (4.4.9) may be written as

$$\frac{\gamma_1^2}{r_*^2} + \frac{\gamma_2^2}{r_*^2} + \ldots \frac{\gamma_p^2}{r_*^2} = 1,$$

which may be rewritten as,

$$r_*^{-2} \tilde{\beta} \cdot \tilde{\beta} = 1,$$
Let the bias-squared of the weighted ridge estimator be written,

\[ \Delta_n = k_2^2 \beta^\prime [(X'X + kI) + \sum_{i=1}^{n_0} n_i \tilde{v}_i \tilde{w}_i \tilde{v}_i]^{-2} \tilde{\beta}, \]  
(4.5.8)

\[ = \tilde{\alpha}' \tilde{D} \tilde{\alpha}. \]

Let the bias-squared of ridge regression, with a \( k_2 \)-value of \( k_2 \) such that there is trace-equality of variances, be written

\[ \Delta_\ast = k_\ast^2 \beta^\prime [X'X + k_\ast I]^{-2} \beta, \]  
(4.5.9)

\[ = \alpha' D_\ast \alpha. \]

In each case the bias-squared has been represented in canonical form with reference to its own axes. These bias-squared equations may be converted to ellipsoids on common axes by replacing both \( \alpha \) and \( \tilde{\alpha} \) by \( \beta \).

The difference in volume due to \( \Delta_\ast \) compared to \( \Delta_n \) is then

\[ \Delta_\ast - \Delta_n = \beta' D_\ast \beta - \tilde{\beta}' \tilde{D} \tilde{\beta}, \]

\[ = \beta' (D_\ast - \tilde{D}) \tilde{\beta}. \]  
(4.5.10)

As \( \tilde{D} \) and \( D_\ast \) are positive definite and symmetric, then for \( D_\ast - \tilde{D} \) to be positive, \( \tilde{D}^{-1} - D_\ast^{-1} \) must be positive definite (Graybill, 1969). The implication is that the \( \ast \)th diagonal elements of \( \tilde{D}^{-1} - D_\ast^{-1} \) satisfy the condition

\[ d_\ast^{-1} - d_i^{-1} > 0, \text{ for all values } i = 1, \ldots, p, \]  
(4.5.11)

where

\( \tilde{d}_i^{-1} \) is the \( \ast \)th eigenvalue of \( [k_\ast X'X + I]^2 \),

and

\( d_i^{-1} \) is the \( i \)th eigenvalue of \( [k_\ast X'X + I]^2 \).
Figure 4.3. The difference in bias-squareds resulting in indeterminant eigenvalues due to the interaction of ellipses.

Figure 4.4. The difference in bias-squareds resulting in positive equivalent eigenvalues.
where
\[ r^2 = \left[ \prod_{i=1}^{p} \left( 1 + k_i^{-2} \lambda_i^{-2} \right) \right]^{1/p} \Delta^{-1}. \]  
(4.5.14)

If equivalent bias-squared is applied to weighted ridge regression then equation (4.5.8) may be expressed in terms of the (rotated) \( \beta \)-axes as
\[ \Delta_\beta = \beta^T \tilde{D} \beta, \]  
(4.5.15)
where \( \tilde{D} \) is a diagonal matrix with \( i^{th} \) diagonal element \( \tilde{d}_i = (1 + k_i^{-1} \lambda_i)^{-2} \).

For ridge regression with \( k \)-value \( k_i \),
\[ \Delta_k = \beta^T D_k \beta, \]  
(4.5.16)
where \( D_k \) is a diagonal matrix, \( i^{th} \) diagonal element \( d^*_i = (1 + k_i^{-1} \lambda_i)^{-2} \).

These may be converted to equivalent bias-squared by use of equation (4.5.14). The condition for the equivalent bias-squared \( k_i \)-ridge regression \( \Delta_k^E \) to be larger than the equivalent bias-squared \( k \)-weighted ridge regression \( \Delta_k^E \) may be expressed as
\[ \Delta_k^E > \Delta_k^E, \]
and is clearly true if \( r_k^2 - \bar{r}^2 > 0 \).  
(4.5.17)

Equation (4.5.17) may be written as
\[ \Delta_k^E = r_k^2 - \bar{r}^2, \]
\[ = k_i^{-p} |\bar{x}^T \bar{x} + k_i I| - k^{-p} |\bar{x}^T \bar{x} + k I|, \]  
(4.5.18)
and when \( \Delta_k^E \) of equation (4.5.18) is positive, the equivalent bias-squared due to \( \Delta_k \) is less than that due to \( \Delta_k^* \).

In summary, the most desirable condition (from the analyst's viewpoint) is that equation (4.4.9) be positive definite. When this
fails, the weaker condition of equation (4.5.13) being positive may be investigated. The weakest of all the conditions is that of equivalent bias-squared of equation (4.5.18). Should an estimator fail to satisfy equation (4.5.18), serious doubts as to the appropriateness of weighted ridge regression should be entertained by the analyst.

4.6. The Sequential Weighted Ridge Estimator

It is not possible to estimate all \( n_o \) \( n \)-weights simultaneously. One technique is the linear or sequential estimation of \( n \)-weights. The specific techniques of estimation of the \( \ell \)th weight will be deferred until Chapter 5. In this section it will be assumed that \((\ell-1)\) weightings have been determined, and the basic properties of the estimator involving the \( \ell \)th \( n \)-weight, \( n_\ell \), will be determined. The following definitions are necessary:

\[
\hat{X}_\ell \hat{Y}_\ell = X'X + \sum_{i=1}^{\ell} n_i \hat{w}_{i-1} \hat{w}_i,
\]

\[
\hat{X}_\ell \hat{Y}_\ell = X'Y + \sum_{i=1}^{\ell} n_i \hat{w}_{i-1} y_i,
\]

\[
F_\ell = (\hat{X}_\ell \hat{Y}_\ell + kI)^{-1},
\]

\[
\hat{w}_\ell = \ell \text{th row of } X,
\]

\[
y_\ell = \ell \text{th observation of } Y,
\]

and

\[
y_\ell^* = \hat{w}_\ell \hat{b}_{\ell-1}^*.
\]

It may be shown that,

\[
\hat{b}^*(\ell) = \hat{b}^*(\ell-1) + \left[\frac{n_\ell}{1 + n_\ell \hat{w}_\ell F_{\ell-1} \hat{w}_\ell} \right] F_{\ell-1} \hat{w}_\ell (y_\ell - y_\ell^*),
\]

(4.6.3)
\[ V(\hat{\theta}^*_k) = F_k [X'X + \sum_{i=1}^{k} n_i(n_i+2)w_iw_i'F_k^{-1}], \] \[ B(\hat{\theta}^*_k) = k^2 F_k \hat{\theta} \hat{\theta}' F_k. \] 

Before investigating the properties of the sequential weighted ridge estimator, a theorem concerning the behavior of the eigenvalues of \( \tilde{X}'\tilde{X} \) is necessary. The behavior of these eigenvalues is the key to understanding the characteristics of the weighted ridge estimator under a sequential estimation scheme.

**Theorem 4.1.** The ith eigenvalue of \( \tilde{X}'\tilde{X} \) is greater than or equal to the ith eigenvalue of \( \tilde{X}'\tilde{X} \) if the n-weighting is positive, but less than or equal to if the n-weighting is negative.

**Proof.**

By definition,

\[ \tilde{X}'\tilde{X} = X'X + nw', \]

where \( w' \) represents the single observation weighted, \( n \) represents the specific weighting for that observation. Both \( X'X \) and \( nw' \) are symmetric, but only \( X'X \) is known to be positive definite. There exists a non-singular matrix \( \tilde{Q} \) such that \( \tilde{Q}'X\tilde{Q} = I \) and \( \tilde{Q}'(nw')\tilde{Q} = n\tilde{D} \), where \( \tilde{D} \) is a diagonal matrix with diagonal elements equal to the eigenvalues of the polynomial equation, \( |w' - \lambda X'X| \) (Graybill, 1969).

Thus,

\[ \tilde{Q}'\tilde{X}'\tilde{X}\tilde{Q} = \tilde{Q}'X\tilde{Q} + n\tilde{Q}'w'\tilde{Q}, \]

\[ = I + n\tilde{D}. \]
Note that when \( n = 0 \),

\[
\bar{q} = \lambda^{-\frac{1}{2}} p,
\]

where \( \lambda = P\text{X}'\text{X}P \), and \( P \) is an orthogonal matrix. The characteristic polynomial is by definition,

\[
|w^\prime w - \lambda x^\prime x| = 0.
\]

This polynomial may be expressed as

\[
|ww^\prime (x^\prime x)^{-1} - \lambda I| |x^\prime x| = 0,
\]

and the equation,

\[
|ww^\prime (x^\prime x)^{-1} - \lambda I| = 0, \tag{4.6.8}
\]

is the polynomial equation for the determination of the eigen-values of \( \overline{ww^\prime (x^\prime x)^{-1}} \). The matrix \( \overline{ww^\prime} \) has a rank of unity and it follows that \( \overline{ww^\prime (x^\prime x)^{-1}} \) has exactly one positive eigen-value. The conclusion is that \( \bar{D} \) of equation (4.6.7) is positive semi-definite.

From equation (4.6.7) it follows that the \( i^{\text{th}} \) root of \( x^\prime x \), denoted by \( \tilde{\lambda}_i \), may be expressed as,

\[
\tilde{\lambda}_i = \lambda_i + n \tilde{d}_i.
\]

where \( \lambda_i \) is the \( i^{\text{th}} \) root of \( x^\prime x \), and \( \tilde{d}_i \) is the \( i^{\text{th}} \) diagonal element of \( \bar{D} \). It follows that \( \tilde{\lambda}_i \geq \lambda_i \) if \( n \) is positive, and \( \tilde{\lambda}_i \leq \lambda_i \) if \( n \) is negative.

Before the properties of the sequential weighted point estimator can be determined, a theorem concerning the maximum and minimum of the quantity \( w_{i-1}^\prime F w_i \) is necessary.

**Theorem 4.2.** If positive \( n \)-weighting is chosen for the \( i^{\text{th}} \) point then \( w_{i-1}^\prime F w_i \) is bounded by zero and unity. If negative \( n \)-weighting is used,
$W_\ell F_\ell W_\ell$ is positive, but unbounded.

**Proof.**

First consider a positive $n$-weighting for the $\ell$th point. By applying Corollary A.1.3 of the Appendix to $F_\ell$ then,

$$F_\ell = F_{\ell-1} - \left[ \frac{n_\ell}{1 + n_\ell W_\ell F_{\ell-1} W_\ell} \right] F_{\ell-1} W_\ell F_{\ell-1},$$

and therefore,

$$W_\ell F_\ell W_\ell = \frac{W_\ell F_{\ell-1} W_\ell}{(1 + n_\ell W_\ell F_{\ell-1} W_\ell)}.$$

However,

$$F_{\ell-1} = \left[ (X'X + kI) + \sum_{i=1}^{\ell-1} n_i W_i W_i \right]^{-1}$$

which is positive definite whenever $n_i > 0$ for $i=1,2,...,(\ell-1)$.

From equation (4.5.10) the conclusions are that,

$$W_\ell F_\ell W_\ell > 0 \text{ for } n_\ell > 1,$$

$$0 < W_\ell F_\ell W_\ell \leq W_\ell F_{\ell-1} W_\ell, \text{ for } n_\ell > 0,$$

and

$$0 \leq W_\ell F_{\ell-1} W_\ell \leq W_\ell F_{\ell} W_\ell, \text{ for } n_\ell < 0.$$

In the determination of the variance-covariance matrix and the bias-matrix of the sequential weighted estimator it is to be assumed that the first $(\ell-1)$ points have been weighted and the $n$-weighting (according to some criterion) for the $\ell$th point is sought.

From equation (4.5.4),
\[ V(\tilde{\beta}^*_{(\ell)} ) = F_{\ell} \left[ X'X + \sum_{i=1}^{\ell} n_i (n_i + 2) \omega_i w_i \omega_i' \right] F_{\ell} \sigma^2, \]
\[ = F_{\ell} G_{\ell}^{-1} F_{\ell} \sigma^2, \]
where
\[ G_{\ell}^{-1} = X'X + \sum_{i=1}^{\ell} n_i (n_i + 2) \omega_i w_i \omega_i'. \]

Applying equation (4.6.9) to equation (4.6.11) yields,
\[ V(\tilde{\beta}^*_{(\ell-1)} ) = V(\tilde{\beta}^*_{(\ell)} ) + \frac{\sigma^2}{(1+n_{\ell} \omega_{\ell} \omega_{\ell-1} w_{\ell-1} w_{\ell})^2} \left\{ A_{\ell} n + B_{\ell} n^2 \right\}, \]
where,
\[ V(\tilde{\beta}^*_{(\ell-1)} ) = F_{\ell-1} G_{\ell-1}^{-1} F_{\ell-1} \sigma^2, \]
\[ A_{\ell} = F_{\ell-1} w_{\ell} \omega_{\ell} \omega_{\ell-1} \left[ I - G_{\ell-1} F_{\ell-1} \right] + (I - F_{\ell-1} G_{\ell-1}^{-1} F_{\ell-1} w_{\ell} \omega_{\ell} \omega_{\ell-1}), \]
\[ B_{\ell} = F_{\ell-1} w_{\ell} \omega_{\ell} \omega_{\ell-1} \left[ I - (w_{\ell} F_{\ell-1} w_{\ell}) G_{\ell-1}^{-1} \omega_{\ell-1} \right] \]
\[ + \left[ (w_{\ell} F_{\ell-1} G_{\ell-1}^{-1} F_{\ell-1} w_{\ell}) I - (w_{\ell} F_{\ell-1} w_{\ell}) F_{\ell-1} G_{\ell-1}^{-1} \right] F_{\ell-1} w_{\ell} \omega_{\ell} \omega_{\ell-1} F_{\ell-1}. \]

It may be demonstrated by methods similar to those used in
Section 4.2 of Chapter 4 that \( V(\tilde{\beta}^*_{(\ell)} ) - V(\tilde{\beta}^*_{(\ell-1)} ) \) is positive semi-definite for a choice of \( n_{\ell} > 0 \), regardless of preceding \( (\ell-1) \) \( n \)-weightings. It follows there exists a negative \( n \)-weighting which makes
\( V(\tilde{\beta}^*_{(\ell)} ) - V(\tilde{\beta}^*_{(\ell-1)} ) \) negative definite.

The bias-matrix of \( \tilde{\beta}^*_{(\ell)} \) is given by,
\[ B(\tilde{\beta}^*_{(\ell)} ) = k^2 F_{\ell} \beta \beta' F_{\ell}, \]
\[ (4.6.13) \]
and by use of Corollary (A.1.3.) of the Appendix may be rewritten as

\[ B(\tilde{\beta}^*_\ell_{(-1)}) = B(\tilde{\beta}^*_\ell) + n_\ell \tilde{\beta}^*_\ell_{-1} \tilde{w}_\ell \tilde{w}_\ell \cdot (B(\tilde{\beta}^*_\ell)) + n_\ell \{ B(\tilde{\beta}^*_\ell) \}_{\tilde{w}_\ell \tilde{w}_\ell} \cdot F_{\ell_{-1}} \]

\[ + n_\ell^2 \tilde{\beta}^*_\ell_{-1} \tilde{w}_\ell \tilde{w}_\ell \cdot (B(\tilde{\beta}^*_\ell)) \}_{\tilde{w}_\ell \tilde{w}_\ell} \cdot F_{\ell_{-1}}. \]  \hspace{1cm} (4.6.14)

The terms in equation (4.6.14) are at least positive semi-definite when \( n_\ell > 0 \), and so the use of a positive \( n_\ell \)-weighting will result in a decrease in bias-squared from the preceding stage. Use of a negative \( n_\ell \)-weighting may possibly cause an increase in bias-squared.

If the experimenter has the reduction of the trace of the variance-covariance matrix of the weighted ridge estimator as the criterion of interest, then the change in trace of the variance-covariance matrix with \( n \)-weighting is of importance. From equation (4.1.11) expressed in canonical form,

\[ \text{tr}[V(\tilde{\beta}^*_1)] = \text{tr}[\tilde{P}\tilde{\lambda}_{-1}^{-1}[\tilde{\lambda} + n(n+1)\tilde{P} \cdot \tilde{w}_k \cdot \tilde{P}] \tilde{\lambda}_{-1}^{-1} \tilde{P} \cdot \sigma^2], \]

\[ = \frac{\tilde{P}}{\tilde{\lambda}_{-1}^2} \sum_{i=1}^{n} \frac{\tilde{\lambda}_{-1}^2}{(\tilde{\lambda}_{+1}+k)^2} + n(n+1) \frac{\tilde{P}}{\tilde{\lambda}_{-1}^2} \sum_{i=1}^{n} \frac{(w^\top P_1)^2 \sigma^2}{(\tilde{\lambda}_{+1}+k)^2}. \]  \hspace{1cm} (4.6.15)

If the Cauchy-Swartz inequality is applied to \( (w^\top P_1)^2 \) of equation (4.6.15), then,

\[ (w^\top P_1)^2 \leq (w^\top w)(P_1^\top P_1), \]

\[ \leq w^\top w, \]

as \( P \) is an orthogonal matrix. An upper bound may therefore be placed on equation (4.6.15) as being,
\[
\text{tr}(V_{\text{max}}(\tilde{\beta}^*_1)) = \sum_{i=1}^{p} \frac{[\lambda_i + n(n+1)\nu^\prime \nu] \sigma^2}{(\lambda_i + k)^2}.
\]

(4.6.16)

From Theorem 4.1,

\[\lambda_i = \lambda_i + n\tilde{d}_i,\]

and as \(n\) becomes very large with respect to the other parameters, then \(\lambda_i \to N\tilde{d}_i\), and it follows that,

\[
\lim_{n \to \infty} \text{tr}(V(\tilde{\beta}^*_1)) = \frac{\sum_{i=1}^{p} \frac{[N(N+1)\nu^\prime \nu + N\tilde{d}_i] \sigma^2}{N \tilde{d}_i^2 + 2N\lambda_i}}{N^2 \tilde{d}_i^2 + 2N\lambda_i}.
\]

(4.6.17)

As \(N \to \infty\), equation (4.6.17) becomes in the limit,

\[
\lim_{N \to \infty} \text{tr}(V(\tilde{\beta}^*_1)) = \sigma^2 \nu^\prime \nu \sum_{i=1}^{p} \frac{\tilde{d}_i^{-2}}{\lambda_i + k}. \]

(4.6.18)

Each term of equation (4.6.18) is positive and monotonically increasing with increasing \(N\) and it follows that equation (4.6.18) is monotonically increasing with increasing \(n\)-weighting. The practical implication is that \(V(\tilde{\beta}^*_1)\) gradually increases as \(n\)-weighting gradually increases.

Consider the trace of the bias matrix given in equation (4.1.15),

\[
\text{tr}[B(\tilde{\beta}^*_1)] = \text{tr}[k^2(\tilde{\tilde{X}}^\prime \tilde{\tilde{X}} + kI)^{-1} \tilde{g}^\prime \tilde{w}^{-1} (\tilde{\tilde{X}}^\prime \tilde{\tilde{X}} + kI)^{-1}],
\]

\[
= k^2 \tilde{g}^\prime \tilde{w}^{-1} (\tilde{\tilde{X}}^\prime \tilde{\tilde{X}} + kI)^{-2} \tilde{g}.
\]

(4.6.19)

In canonical form this becomes,

\[
\text{tr}[B(\tilde{\beta}^*_1)] = k^2 \tilde{g}^\prime \tilde{\alpha}^{-2} \tilde{\alpha} \tilde{g},
\]

\[
= k^2 \nu^\prime \nu \sum_{i=1}^{p} \frac{\tilde{d}_i^{-2}}{\lambda_i + k}. \]

(4.6.20)
From Theorem 4.1,

\[ \tilde{\lambda}_1 = \lambda_1 + n\tilde{d}_1. \]  \hspace{1cm} (4.6.21)

and if equation (4.6.21) is applied to equation (4.6.20) then the bias-squared, \( \text{tr}[B(B^*)] \) is monotonically decreasing for increasing positive n-weighting.

In conclusion, note that for positive n-weighting, the trace variance increases while the trace bias-squared decreases. The reverse is not necessarily true for negative n-weighting, further discussion being left to Chapter 5.
5. THE CALCULATION OF n-WEIGHTS

In Chapter 4 the variance and bias-squared of the weighted ridge estimator were determined assuming the n-weightings were known. The determination of these n-weightings will now be investigated. The advantage in using a sequential technique lies in the ease of computation and in the way an analyst can determine variance and (relative) bias improvements at any stage of the sequence. The disadvantage lies in the fact that, although each stage of the sequence may have an optimal n-weighting, there is no guarantee the final complete set of n-weightings is optimal.

The criteria for calculation of optimal n-weightings to be considered are:

5.1 A-optimality of the Variance-covariance Matrix
5.2 A-optimality of the Bias-squared Matrix
5.3 D-optimality of the Variance-covariance Matrix
5.4 Minimum Residual Sum of Squares
5.5 Stability of Estimated Parameters
  5.5.1 The $Q^*$ standard
  5.5.2 The $Q^{**}$ standard
  5.5.3 The $S^*$ standard
  5.5.4 The $S^{**}$ standard
5.6 Sign Change of an Estimated Parameter
5.7 Iterative Sequential n-weighting.

This listing is not an exhaustive compilation of criteria, but a collection of most commonly used, or literature cited, criteria.
5.1 A-optimality of the Variance-covariance Matrix.

The criterion of A-optimality seeks an n-weighting such that the trace of the variance-covariance matrix is a minimum. It is assumed that the first (ℓ-1) n-weightings have been determined and the A-optimal n-weighting for the ℓth point is desired. From equation (4.6.11) the variance-covariance matrix of \( \tilde{\mathbf{B}}^*_{(\ell)} \) is

\[
V(\tilde{\mathbf{B}}^*_{(\ell)}) = F_{\ell} C_{\ell}^{-1} F_{\ell} \sigma^2
\]

\[
[F_{\ell-1}^{-1} + n_{\ell} w_{\ell} w_{\ell}^{-1}]^{-1} [G_{\ell-1}^{-1} + n_{\ell} (n_{\ell}+2) w_{\ell} w_{\ell}^{-1}] [F_{\ell-1}^{-1} + n_{\ell} w_{\ell} w_{\ell}^{-1}]^{-1} \sigma^2 .
\]

(5.1.1)

If equation (4.6.9) is applied to equation (5.1.1) and the trace of the matrix taken, then

\[
\text{tr}\left[ V(\tilde{\mathbf{B}}^*_{(\ell)}) \right] = \text{tr} \left[ V(\tilde{\mathbf{B}}^*_{(\ell-1)}) \right] + \frac{an_{\ell} + bn_{\ell}^2}{(1 + cn_{\ell})^2}
\]

(5.1.2)

where

\[
a = 2 w_{\ell} F_{\ell-1} w_{\ell} - w_{\ell} F_{\ell-1} G_{\ell-1}^{-1} F_{\ell-1} w_{\ell} - w_{\ell} F_{\ell-1} G_{\ell-1}^{-1} F_{\ell-1} w_{\ell} ,
\]

\[
b = (w_{\ell} F_{\ell-1} w_{\ell}) (w_{\ell} F_{\ell-1} G_{\ell-1}^{-1} F_{\ell-1} w_{\ell}) + w_{\ell} F_{\ell-1} w_{\ell} - (w_{\ell} F_{\ell-1} w_{\ell}) (w_{\ell} F_{\ell-1} G_{\ell-1}^{-1} F_{\ell-1} w_{\ell}) ,
\]

\[
c = w_{\ell} F_{\ell-1} w_{\ell} .
\]
If \( \text{tr}[V(\hat{\beta}^\star_{(\ell)})] \) in equation (5.1.2) is minimized, then

\[
n_{\ell}(A\text{-optimal variance}) = \frac{-a}{2b - ac},
\]

and is non-zero providing \( a \neq 0 \).

The trace of the variance-covariance matrix for this particular A-optimal n-weighting is given by

\[
\text{tr}[V(\hat{\beta}^\star_{(\ell)})] = \text{tr}[V(\hat{\beta}^\star_{(\ell-1)})] - \frac{a^2}{4(b-ac)}. \tag{5.1.4}
\]

Observe that \( \text{tr}[V(\hat{\beta}^\star_{(\ell)})] < \text{tr}[V(\hat{\beta}^\star_{(\ell-1)})] \), provided \((b-ac)\) of equation (5.1.4) is positive. The inference to be made is that the points may be sequentially weighted by use of equation (5.1.3), and the trace of the variance-covariance matrix reduced at each operation. T Tedious but straightforward algebra shows that \( a, g, \) and \((b-ac)\) are all positive. Similar algebra shows that

\[
a - (2b-ac) < 0,
\]

which is sufficient to guarantee the \( n_{\ell} \) weighting of equation (5.1.3) is feasible \((n_{\ell} > -1)\).

5.2 A-optimality of the bias-squared matrix

Here, the n-weighting which minimizes the trace of the bias-matrix is sought. From equation (4.6.13),

\[
B(\hat{\beta}^\star_{(\ell)}) = k^2 F_{\frac{3}{2}} \bar{B} \bar{F}_{\frac{1}{2}}.
\]

If equation (4.6.9) is applied to \( B(\hat{\beta}^\star_{(\ell)}) \) and the trace taken, then
\[ \text{tr}[B(\tilde{\beta}^*_{(L)})] = \text{tr}[B(\tilde{\beta}^*_{(L-1)})] - \frac{1}{(1 + n_L r_1)^2} \left[ \frac{2R_1 n_L + (2R_1 + R_2)n_L^2}{R_2} \right] \] (5.2.1)

where
\[ r_1' = w_L^{-2} f_{L-1} w_L \]
\[ R_1 = \tilde{\beta}^*_{(L-1)} w_L w_L^{-1} f_{L-1} \tilde{\beta} \]

and
\[ R_2 = \tilde{\beta}^*_{(L-1)} w_L w_L^{-1} f_{L-1} \tilde{\beta} \]

If \( \text{tr}[B(\tilde{\beta}^*_{(L)})] \) of equation (5.2.1) is minimized then
\[ n_L (A-optimal bias) = \frac{R_1}{r_2 R_2 - r_1 R_1} \] (5.2.2a)

or
\[ = + \infty, \] (5.2.2b)

where \( r_2 = w_L^{-2} f_{L-1} w_L \).

From Section 4.3 it is known that the trace bias-matrix decreases with increasing \( n \)-weighting. The conclusion to be drawn is that the \( n_L (A-optimal bias) \) given by equation (5.2.2a) is negative. It may be shown, however, that this negative \( n \)-weighting is always less than -1, and consequently infeasible. The only recourse for the analyst is to give the \( L \)th point infinite \( n \)-weighting.

Note that the trace of the bias-squared matrix using \( n (A-optimal bias) = + \infty \) is given by

\[ \text{tr}[B(\tilde{\beta}^*_{(L-1)})] = \text{tr}[B(\tilde{\beta}^*_{(L-1)})] - \frac{k_2}{r_2} \left[ 2r_2 R_1 - r_1 R_2 \right] \] (5.2.3)
5.3 D-optimality of the Variance-Covariance Matrix.

The criterion of D-optimality seeks the n-weighting such that the determinant of the variance-covariance matrix is a minimum. As in the A-optimality criterion, the n-weighting for the kth point is needed. From equation (4.6.11) the variance-covariance matrix of \( \bar{\beta}^* (\lambda) \) is

\[
\mathbf{V}(\bar{\beta}^* (\lambda)) = \mathbf{F}_k \mathbf{G}^{-1}_k \mathbf{F}^\top_k \sigma^2,
\]

and minimizing the determinant of \( F_k G^{-1}_k F^\top_k \) yields

\[
\min |F_k G^{-1}_k F^\top_k \sigma^2| = \sigma^2 \max |F_k G^{-1}_k F^\top_k |.
\]

(5.3.1)

The optimal \( n_k \)-weighting will therefore be that which maximizes

\[
V_D = \frac{|F_k^{-1}|^2}{|G_k^{-1}|}.
\]

(5.3.2)

By repeated use of equation (A.2.2) of the Appendix it follows that maximizing \( V_D \) of equation (5.3.2) is equivalent to minimizing

\[
D_D = \frac{(1+ n_k w_k \bar{\lambda}_k^{-1} G_k^{-1} \bar{\lambda}_k)^2}{1 + n_k (n_k + 2) w_k \bar{G}_k^{-1} G_k^{-1} \bar{\lambda}_k} \frac{\{ \bar{x}_{k-1}^\top \bar{X}_{k-1} + kI \}^2}{|G_k^{-1}|^2}.
\]

(5.3.3)

Minimizing \( D_D \) of equation (5.3.3) yields the n-weighting

\[
n (D-optimal variance) = \frac{w_k \bar{\lambda}_k^{-1} G_k^{-1} \bar{\lambda}_k - w_k \bar{\lambda}_k^{-1} G_k^{-1} \bar{\lambda}_k}{(1-w_k \bar{\lambda}_k^{-1} G_k^{-1} \bar{\lambda}_k) w_k \bar{G}_k^{-1} G_k^{-1} \bar{\lambda}_k}.
\]

(5.3.4)
The $n_{\lambda}$-weighting of equation (5.3.4) may be seen a little clearer by the following argument.

Suppose \( (1 - w_{\lambda} F_{\lambda-1} w_{\lambda}) > 0 \). Observe that both \( w_{\lambda} F_{\lambda-1} w_{\lambda} \) and \( w_{\lambda} G_{\lambda-1} w_{\lambda} \) are positive regardless of the choice of preceding $n$-weights. Consider

\[
h = \frac{w_{\lambda} G_{\lambda-1} w_{\lambda} - w_{\lambda} F_{\lambda-1} w_{\lambda}}{w_{\lambda} G_{\lambda-1} w_{\lambda} - F_{\lambda-1} I_{\lambda}}.\]  

(5.3.5)

If \( G_{\lambda-1} - F_{\lambda-1} \) of equation (5.3.5) is positive definite, then \( h \) must be positive. Requiring \( G_{\lambda-1} - F_{\lambda-1} \) to be positive definite is equivalent to requiring \( F_{\lambda-1}^{-1} - G_{\lambda-1}^{-1} \) is positive definite (Graybill, 1969),

\[
(F_{\lambda-1}^{-1} - G_{\lambda-1}^{-1}) = (\bar{X}_{\lambda-1}^T \bar{X}_{\lambda-1} + kI) - (\bar{X}_{\lambda-1}^T \bar{X}_{\lambda-1} + \sum_{i=1}^{n_{\lambda}} n_{i}(n_{i}+1)w_{i}w_{i}^T),
\]

(5.3.6)

Equation (5.3.6) is certainly positive definite if all the \( n_{i} \)'s are between $-1$ and $0$. With a sequential procedure, the first $n_{\lambda}$ (D-optimal variance) will be negative, and so all the remaining optimal $n$-weightings will be negative. If the quantity \( (1 - w_{\lambda} F_{\lambda-1} w_{\lambda}) \) is negative, the $n$-weighting satisfying equation (5.3.4) may not exist.

As \( BB^T \) of equation (4.6.13) is singular, it follows that the determinant of \( F_{\lambda} BB^T F_{\lambda} \) is zero. This implies it is not possible to search for a D-optimal bias $n$-weighting.
5.4 Minimum Residual Sum of Squares.

The minimum residual sum of squares criterion is of use when ridge regression has caused the residual sum of squares to become unacceptably large. Rather than choose a different k-value it is decided to use weighted ridge regression to reduce the residual sum of squares. It is necessary to make the assumption that the function fitted, up to and including the \((l-1)\)th point, does not pass through the \(l\)th point to be considered. Should this be the case, then any feasible weighting for \(n_\lambda\) would be acceptable as the residual at that point is zero.

The residual sum of squares is defined as

\[
\tilde{\Phi}^* (\lambda) = \left[ \frac{y - X\hat{\beta}^*}{} \right]^T \left[ \frac{y - X\hat{\beta}^*}{} \right].
\]  

From equation (4.3.3),

\[
\tilde{\beta}^* (\lambda) = \tilde{\beta} (\lambda-1) + \left\{ \frac{n_\lambda}{1 + n_\lambda \tilde{w}_\lambda \tilde{F}_\lambda^{-1} \tilde{w}_\lambda} \right\} \tilde{F}_\lambda^{-1} \tilde{w}_\lambda (y_\lambda - y_\lambda^*)
\]

where \(y_\lambda^* = \tilde{w}_\lambda \tilde{\beta}^* (\lambda-1)\). If equation (4.5.3) is used with equation (5.4.1), then

\[
\tilde{\Phi}^* (\lambda) = \tilde{\Phi} (\lambda-1) + \left\{ \frac{2n_\lambda}{1 + n_\lambda \tilde{w}_\lambda \tilde{F}_\lambda^{-1} \tilde{w}_\lambda} \right\} \left( y_\lambda - y_\lambda^* \right) \tilde{w}_\lambda \tilde{F}_\lambda^{-1} \tilde{w}_\lambda \tilde{X}^T \tilde{X} (\tilde{\beta} (\lambda-1) - \tilde{\beta})
\]

\[
+ \left\{ \frac{2n_\lambda}{1 + n_\lambda \tilde{w}_\lambda \tilde{F}_\lambda^{-1} \tilde{w}_\lambda} \right\} \left( y_\lambda - y_\lambda^* \right) \tilde{w}_\lambda \tilde{F}_\lambda^{-1} \tilde{w}_\lambda \tilde{X}^T \tilde{X} \tilde{F}_\lambda^{-1} \tilde{w}_\lambda
\]

Minimization of \(\tilde{\Phi}^* (\lambda)\) results in the optimal \(n\)-weighting,

\[
n_\lambda (\text{Optimal residual S.Sqs.}) = \frac{k \tilde{F}_\lambda^{-1} \tilde{X}^T \tilde{X} \tilde{F}_\lambda^{-1} \tilde{w}_\lambda}{\left( y_\lambda - y_\lambda^* \right) \tilde{w}_\lambda + k \left( \tilde{w}_\lambda \tilde{F}_\lambda^{-1} \tilde{w}_\lambda \right) \tilde{\beta}^T \tilde{F}_\lambda^{-1} \tilde{X}^T \tilde{X} \tilde{F}_\lambda^{-1} \tilde{w}_\lambda}
\]  

(5.4.3)
The residual sum of squares for the n-weighting of equation (5.4.3) is given by

\[ \tilde{\phi}^* (\ell) (\text{optimal residual S.Sqs.}) = \tilde{\phi}^* (\ell-1) - \frac{k^2 (\beta_0 \hat{F}_{\ell-1} X'X F_{\ell-1} \hat{w}_k)^2}{\hat{w}_k \hat{F}_{\ell-1} X'X F_{\ell-1} \hat{w}_k} \]  \[ (5.4.4) \]

The form of equation (5.4.4) suggests use of the secondary criterion, order of selection of weighting points. If the Cauchy-Schwartz inequality is applied to \( (\beta_0 \hat{F}_{\ell-1} X'X F_{\ell-1} \hat{w}_k)^2 \) of equation (5.4.4), then

\[ \tilde{\phi}^* (\ell) \text{[minimum]} = \tilde{\phi}^* (\ell-1) - k^2 \frac{\beta_0}{\hat{w}_k} \hat{F}_{\ell-1} X'X F_{\ell-1} \hat{\beta} \]  \[ (5.4.5) \]

For the \( \tilde{\phi}^* (\ell) \text{[minimum]} \) of equation (5.4.5) to be attained, the weighted point \( \hat{w} \) must be a multiple of \( \hat{\beta} \), and so may be written

\[ \hat{w} = c \hat{\beta} \]  \[ (5.4.6) \]

where \( c \) is a constant.

5.5 Stability of Estimates of Parameters.

Stability of an estimate implies that a small change in the underlying data structure, or data base, should not unduly change the numerical estimate of that parameter. In multicollinear situations, a small change in the data base can often cause dramatic changes in the estimates. When using weighted ridge estimation, it is tempting to use a sufficiently large n-weighting so that small changes in n-weighting will produce very little change in estimate. This large n-weighting is misleading as the inherent instability of the estimates has only been masked by the large n-weighting and not properly alleviated.
It follows that a criterion other than "rate-of-change of estimate" with "n-weighting" is needed. This new criterion must not be influenced by large numerical values of n-weighting. From equation (4.5.4),

\[
\hat{\beta}(\ell) = \hat{\beta}(\ell-1) + \frac{n_\ell}{1 + n_\ell w_\ell F_{\ell-1} w_\ell} F_{\ell-1} w_\ell (y_\ell - \hat{y}_\ell^*) .
\]

The \( \hat{\beta}(\ell) \) of equation (4.3.4) may be minimized with respect to \( n_\ell \) to obtain

\[
\frac{d\hat{\beta}(\ell)}{dn_\ell} = \frac{F_{\ell-1} w_\ell (y_\ell - \hat{y}_\ell^*)}{(1 + n_\ell w_\ell F_{\ell-1} w_\ell)^2} ,
\]

(5.5.1)

and this rate-of-change is clearly dependent on \( n_\ell \). Large values of \( n \) will cause \( d\hat{\beta}(\ell)/dn_\ell \) to be small and so indicate an "apparent" stability.

Define a new variable,

\[
q_\ell = \frac{n_\ell w_\ell F_{\ell-1} w_\ell}{1 + n_\ell w_\ell F_{\ell-1} w_\ell} ,
\]

(5.5.2)

and consider the rate-of-change of \( q_\ell \) with respect to \( n \),

\[
\frac{dq_\ell}{dn_\ell} = \frac{w_\ell F_{\ell-1} w_\ell}{(1 + n_\ell w_\ell F_{\ell-1} w_\ell)^2} .
\]

(5.5.3)

By the chain-rule of differentiation,
\[ \frac{d\hat{\beta}^*}{dq_k} = \begin{bmatrix} \frac{d\hat{\beta}^*}{dn_k} \\ \frac{dn_k}{dq_k} \end{bmatrix}, \]

\[ = \frac{(y_k^* - y_{k-1}^*)}{W_k F_{k-1} W_k} W_k, \]

and is independent of \( n_k \). Notice that

\[ W_k F_k = \frac{W_k F_{k-1}}{1 + n_k W_k F_{k-1} W_k}, \]

and post-multiplication by \( W_k \) yields

\[ W_k F_k W_k = \frac{W_k F_{k-1} W_k}{1 + n_k W_k F_{k-1} W_k}, \]

\[ = \frac{1}{n_k} q_k. \]

Define

\[ Q_k = \sum_{i=1}^{k} q_i, \]

\[ = \sum_{i=1}^{k} n_i W_i^* F_i W_i. \]

This suggests that, instead of a visual plot of "estimate" against "n-weighting", one should use a plot of "estimate" against "Q_k". Stability could then be estimated visually in the same spirit as the "m-plot" of Vinod (1976). If a visual estimate of stability is rejected as being too subjective, the following standards of stability may be applied.
5.5.1 The $Q^*$ Standard. The $Q^*$ standard compares the rate-of-change in the specific problem (with respect to $n_\ell$), with the rate-of-change of an unspecified orthogonal problem (with respect to $n_\ell$). For the unspecified orthogonal problem, the following definitions are the analogs of equations (4.6.4) and (5.5.1) to (5.5.6).

\[
\tilde{\varphi}_{\ell}^{*o} = [\frac{1}{1+k} \mathbf{I} + \sum_{i=1}^{\ell} n_i^\ell w_i w_i]^{-1} [X^\ell Y + \sum_{i=1}^{\ell} n_i^\ell w_i y_i]
\]  

(5.1.1.1)

\[
\frac{d\tilde{\varphi}_{\ell}^{*o}}{dn_{\ell}} = \frac{\mathbf{F}_{\ell-1}^o \mathbf{w}_{\ell}^o (y_{\ell} - y_{\ell-1}^*)}{(1 + n_{\ell}^o \mathbf{w}_{\ell}^o \mathbf{F}_{\ell-1}^o \mathbf{w}_{\ell})^2},
\]

(5.5.1.2)

where

\[
\mathbf{F}_{\ell-1}^o = [\frac{1}{1+k} \mathbf{I} + \sum_{i=1}^{\ell-1} n_i^\ell w_i w_i]^{-1}.
\]

Also,

\[
\frac{dq_{\ell}^{*o}}{dn_{\ell}} = \frac{\mathbf{w}_{\ell}^o \mathbf{F}_{\ell-1}^o \mathbf{w}_{\ell}}{(1 + n_{\ell}^o \mathbf{w}_{\ell}^o \mathbf{F}_{\ell-1}^o \mathbf{w}_{\ell})^2},
\]

(5.5.1.3)

and

\[
Q_{\ell}^{*o} = \sum_{i=1}^{\ell} q_i^{*o}.
\]

(5.5.1.4)
Define

\[ Q_x^* = \left( \frac{\langle dq_x^O/dn_x^O \rangle}{\langle dq_x^O/dn_x \rangle} - 1 \right)^2. \]  \hspace{1cm} (5.5.1.5)

When the problem has similar characteristics to an orthogonal situation, \( Q_x^* \) will be zero. As the orthogonal situation is the most stable situation by definition, the optimum \( n \)-weighting is that which minimizes \( Q_x^* \) of equation (5.5.1.5).

Substitution of equations (5.5.1.1) and (5.5.1.3) into equation (5.5.1.5), and minimization of \( Q_x^* \) yields

\[ n_x^2 = \frac{1}{\langle \psi_x^O F_{x-1}^O \psi_x^O \rangle \langle \psi_x^F F_{x-1} \psi_x^F \rangle}. \]  \hspace{1cm} (5.5.1.6)

Equation (5.5.1.6) implies either a positive or negative value of \( n_x \) is possible, depending on the precise structure of \( F_{x-1}^O \) and \( F_{x-1} \). Secondary criterion such as variance or bias-squared considerations could be brought into play to determine the appropriate \( n \)-weighting.

5.5.2. The \( Q_x^{**} \) standard. The \( Q_x^{**} \) standard is an extension of the \( Q_x^* \) standard in that the theoretical, orthogonal situation is derived from the actual situation being investigated. The advantage of using a specific orthogonal matrix lies in the analyst being able to visualize his experiment as being a "fractional replicate" of a much larger orthogonal experiment. This orthogonal matrix uses the original design matrix \( S \), and the O.L.S. estimates. The "new" design matrix \( X_a \) is such that
\[ X_a^T X_a = cI, \]

where \( c \) is a constant, and \( I \) is the identity matrix.

Define

\[
X_a = \begin{bmatrix}
X \\
\vdots \\
X_{a1} \\
\vdots \\
X_{a(2^p-1)}
\end{bmatrix}
\text{ of size } (2^p n_0 \times p), \quad (5.5.2.1)
\]

where, for example,

\[
X_{aj} = \begin{bmatrix}
w_{11} & w_{12} & \cdots & w_{1j} & \cdots & w_{1p} \\
w_{21} & w_{22} & \cdots & w_{2j} & \cdots & w_{2p} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
w_{n_01} & w_{n_02} & \cdots & w_{n_0j} & \cdots & w_{n_0p}
\end{bmatrix}
\]

With the "new" rows of \( X_a \) denoted by \( \hat{w}_{a1}, \hat{w}_{a2}, \ldots, \hat{w}_{a(2^p-1)} \), the "new" \( Y \)-values at these points are then,

\[
Y_a^* = [Y, \hat{Y}_{a1}, \hat{Y}_{a2}, \ldots, \hat{Y}_{a(2^p-1)}].
\quad (5.5.2.2)
\]

This concept of "creating" an orthogonal design may be seen clearly in Figure 5.1.

If ordinary least-squares technique is used with \( X_a \) and \( Y_a \),

\[
\hat{\beta}_{\text{augmented}} = (X_a^T X_a)^{-1} X_a^T Y_a, \quad (5.5.2.3)
\]

where
* denotes original points
  denotes artificial (augmented) points

Figure 5.1. An augmented design for $p = 2$, $n_o = 6$. 
\[ X^\dagger_a X_a = \sum_{i=1}^{n_0} w_i w_i^\dagger + \sum_{i=1}^{n_0 2^p-1} w_{a_i} w_{a_i}^\dagger, \]

\[ = 2^p I, \]

by definition of orthogonality. Also,

\[ X^\dagger_a w_a = X^\dagger Y + \sum_{i=1}^{n_0 2^p-1} w_{a_i} w_{a_i}^\dagger \hat{\beta}, \]

\[ = X^\dagger Y + 2^p \hat{\beta} - X^\dagger \hat{\beta}, \]

\[ = 2^p \hat{\beta}. \]

If the ridge technique is used with this augmented data set,

\[ \hat{\beta}_{\text{augmented}}^* = (X^\dagger_a X_a + kI)^{-1} X^\dagger_a Y_a, \]

\[ = [1 + k(X^\dagger_a X_a)^{-1}]^{-1} \hat{\beta}_{\text{augmented}}, \]

\[ = [1 + \frac{k}{2^p}] \hat{\beta}. \tag{5.5.2.4} \]

The following quantities are established in the same way as their analogs in Section 5.5.1,

\[ \hat{\beta}_{(\ell)\text{augmented}}^* = [(2^p+k)I + \sum_{i=1}^{l} n_i w_i w_i^\dagger \gamma^{-1}(2^p \hat{\beta} + \sum_{i=1}^{l} n_i w_i y_i), \tag{5.5.2.5} \]

\[ \frac{d\hat{\beta}_{a}^*}{dn} = \frac{\mathbf{F}^a_{\ell-1} w_{\ell} (y_{\ell} - y_{\ell-1}^\star)}{(1 + n_k w_{\ell} F_{\ell-1} w_{\ell})^2}, \tag{5.5.2.6} \]
and
\[ q^a_k = \frac{n_k w_k^a F_k^a w\hat{1}_k}{1 + n_k w_k^a F_{k-1} w\hat{1}_k} \]  
(5.5.2.7)

where
\[ 2 F_{2-1} = \left[ (2^P + k)I + \sum_{i=1}^{k-1} n_i w_i w\hat{1}_i \right]^{-1} \]

Define
\[ Q^*_{\hat{k}} = \left\{ \frac{\left( \frac{dq_k}{dn_k} \right)^2}{\left( \frac{dd_k}{dn_k} \right)} - 1 \right\} . \]  
(5.5.2.8)

Substitution of equations (5.5.1.1) and (5.5.2.6) into equation (5.5.2.8), and minimization of \( Q^*_{\hat{k}} \) with respect to \( n_k \) yields
\[ n_k^2 = \frac{1}{(w_k F_{k-1} w\hat{1}_k)(w_k F_{k-1} w\hat{1}_k)} . \]  
(5.5.2.9)

It is possible the choice of a negative value for \( n_k \) from equation (5.5.2.9) results in a non-viable n-weighting \( (n_k < -1) \). Such values should be automatically excluded and the positive value of \( n \) used in the next stage of the sequential estimation process.

5.5.3. The \( S^* \) standard. Vinod (1976) proposes a measure of stability of estimates that is not unduly affected by magnitudes of k-values or estimated parameters. Using a new variable \( (m) \), defined in terms of the eigenvalues of \( X^t X \) and k-value chosen, a standard \( (S) \) is proposed which compares the stability of situations being investigated.
with an orthogonal situation. The k-value is then altered until the
S standard is a minimum; this value of k is then known as \( k_{\text{stability}} \).

The \( S_k^* \) standard is an adaptation of the S standard except
that the focus is on \( n \)-weighting for stability, rather than \( k \)-value
for stability. By definition of the weighted ridge regression estimator,

\[
\tilde{\mathbf{S}}^*(\xi) = (\tilde{\mathbf{X}}_k \tilde{\mathbf{X}} + k I)^{-1} \tilde{\mathbf{X}}_k \tilde{\mathbf{Y}},
\]

\[
= [I + k (\tilde{\mathbf{X}}_k \tilde{\mathbf{X}})^{-1}]^{-1} \tilde{\mathbf{S}}(\xi).
\]

The equation (5.5.3.1) may be written in canonical form as

\[
\tilde{\mathbf{S}}^*(\xi) = \tilde{\mathbf{P}}_k \tilde{\mathbf{D}}_k \tilde{\mathbf{P}}_k \tilde{\mathbf{S}}(\xi),
\]

where \( \tilde{\mathbf{D}}_k \) is diagonal with elements,

\[
\begin{pmatrix}
\frac{\tilde{\lambda}_{kj}}{\tilde{\lambda}_{kj} + k} \\
\frac{\tilde{\lambda}_{kj}}{\tilde{\lambda}_{kj} + k}
\end{pmatrix}, \quad j = 1, \ldots, p.
\]

Following Vinod, define

\[
\tilde{m}_k = p - \sum_{j=1}^{p} \frac{\tilde{\lambda}_{kj}}{\tilde{\lambda}_{kj} + k}, \quad 0 < \tilde{m}_k < p.
\]

The orthogonal counterpart of equation (5.5.3.2) may be written,

\[
\tilde{m}_k^o = p - \sum_{j=1}^{p} \left( \frac{1}{1 + k} \right),
\]

\[
= \frac{pk}{1 + k}.
\]

The orthogonal weighted ridge estimate is related to the orthogonal
weighted estimator by
\[
\hat{\beta}^o(t) = \left( \frac{1}{1 + k} \right) \hat{\beta}^o(t).
\]

Following the development of equation (5.5.4) then,

\[
\frac{d\hat{\beta}^o(\ell)}{d\bar{m}_k} = \begin{bmatrix} \frac{d\hat{\beta}^o(\ell)}{d\bar{n}_k} \\ \frac{d\bar{n}_k}{d\bar{m}_k} \end{bmatrix}.
\]

(5.5.3.4)

but the quantity \( \frac{d\bar{n}_k}{d\bar{m}_k} \) of equation (5.5.3.4) cannot be easily obtained from equation (5.5.3.2) due to the extreme difficulty of expressing \( \tilde{\lambda}_{k,j} \) in terms of \( \lambda_j \) and \( n_k \). If each \( \tilde{\lambda}_{k,j} \) is bounded by the maximum each root could attain, a solution to equation (5.5.3.4) may be obtained.

From equation (4.3.9) it may be deduced that \( \tilde{\lambda}_{k,j} \leq n_k \lambda_j \) and therefore the equation (5.5.3.2) may be written as

\[
\bar{m}_k = p \left[ \frac{p \frac{(n_k+1)\lambda_j}{j \ (n_k+1)\lambda_j+1}}{j \ (n_k+1)\lambda_j+1} \right],
\]

(5.5.3.5)
and the corresponding sequential weighted estimator as

$$\tilde{B}^*_{(\xi)} = \tilde{P}^\top \tilde{D}^{-1}_\lambda \tilde{P} \tilde{B}_{(\xi)},$$  \hspace{1cm} (5.5.3.6)

where $D^{-1}_\lambda$ is a diagonal matrix, with $j$th diagonal element

$$\left( \frac{(n_{(l+1)}\lambda_j^j)}{(n_{(l+1)}\lambda_j^j+k)} \right).$$

If $\tilde{m}_{(\xi)}$ of equation (5.5.3.5) is differentiated with respect to $n_{(l)}$, then

$$\frac{d\tilde{m}_{(l)}}{dn_{(l)}} = -k \sum_{j=1}^{p} \lambda_j \frac{j}{n_{(l+1)}\lambda_j^j+k}.$$

If $\tilde{B}^*_{(\xi)}$ of equation (5.5.3.6) is minimized with respect to $n_{(l)}$,

$$\frac{d\tilde{B}^*_{(\xi)}}{dn_{(l)}} = \frac{n_{(l+1)}\lambda_j^j}{1+n_{(l)}w_{(l)}\tilde{P}^\top \tilde{D}^{-1}_\lambda \tilde{P} \tilde{B}_{(\xi)}}.$$  \hspace{1cm} (5.5.3.8)

When the orthogonal forms of equations (5.5.3.7) and (5.5.3.8) are substituted into equation (5.5.3.4),

$$\frac{d\tilde{B}^*_{(\xi)}}{dn_{(l)}} \text{ orthogonal} = -\frac{1}{pk} \tilde{w}_{(l)}(\tilde{v}_{(l)} - \tilde{v}_{(l-1)})$$  \hspace{1cm} (5.5.3.9)

and the right-hand side of equation (5.5.3.9) is independent of $n_{(l)}$.

Define the $S^*_{(l)}$ measure of stability as

$$S^*_{(l)} = \sum_{j=1}^{p} \frac{kp\lambda_j}{n_{(l+1)}\lambda_j^j+k} \left[ \frac{1}{S_{(l)}[(n_{(l+1)}\lambda_j^j+k)^2] - 1} \right]^2$$  \hspace{1cm} (5.5.3.10)
where

\[
S^*_\ell = \begin{vmatrix} \frac{d m_\ell}{d n_\ell} \end{vmatrix},
\]

(5.5.3.11)

\[
= k \sum_{j=1}^{p} \frac{\lambda_j}{[(n_\ell - 1)\lambda_j + k]^2}.
\]

S^*_\ell will be zero for a (replaced) orthogonal design matrix X and a reasonable criterion would be to choose an n_\ell that minimized S^*_\ell of equation (5.5.3.10). This would have to be done by numerical approximation as a closed-form solution to equation (5.5.3.10) could not be found for p > 2.

5.5.4. The \(S^{**}_{\ell}\) standard. The \(S^{**}_{\ell}\) standard is similar to the \(S^*_\ell\) standard but differs in the concept of obtaining the maximum increase in any given root. From equation (4.3.9), using trace \([\tilde{X}^T \tilde{X}]\) considerations, it may be deduced that \(\tilde{\lambda}_j = \lambda_j + n_\ell \tilde{\nu}_j\).

Define

\[
m^*_\ell = p - \sum_{j=1}^{p} \frac{\lambda_j + n_\ell \tilde{\nu}_j \tilde{\nu}_\ell}{\lambda_j + k + n_\ell \tilde{\nu}_\ell \tilde{\nu}_\ell},
\]

(5.5.4.1)

and

\[
\tilde{\beta}^{**}_{\ell}(\ell) = \tilde{P}_\ell D^w_{\ell} \tilde{P}_\ell \tilde{\beta}^{**}_{\ell}(\ell),
\]

(5.5.4.2)

where \(D^w_{\ell}\) is a diagonal matrix with jth element

\[
\begin{bmatrix}
\frac{\tilde{\lambda}_j + n_\ell \tilde{\nu}_j \tilde{\nu}_\ell}{\tilde{\lambda}_j + k + n_\ell \tilde{\nu}_\ell \tilde{\nu}_\ell}
\end{bmatrix}.
\]
If $m_w^\ell$ of equation (5.5.4.1) is differentiated with respect to $n_\ell^W$,

$$\frac{dm_w^\ell}{dn_\ell^W} = -\sum_{i=1}^{p} \frac{k_w^\ell \dot{w}_i^\ell}{(\lambda_j^W + k + n_\ell^W \dot{w}_i^\ell)^2}. \quad (5.5.4.3)$$

If $\tilde{\beta}^{**}(\ell)$ of equation (5.5.4.2) is differentiated with respect to $n_\ell^W$,

then

$$\frac{d\tilde{\beta}_w^{**}(\ell)}{dn_\ell^W} = \frac{\tilde{p}_\ell^W \dot{D}_w^W \tilde{p}_\ell^W \dot{D}_w^W (y_\ell^W - y_{\ell-1}^{**})}{[1 + n_\ell^W \dot{w}_i^\ell \tilde{p}_\ell^W \dot{D}_w^W \tilde{p}_\ell^W \dot{w}_i^\ell]^2}. \quad (5.5.4.4)$$

If the orthogonal forms of equations (5.5.4.3) and (5.5.4.4) are substituted into equation (5.5.3.4), then

$$\left. \frac{d\tilde{\beta}_w^{**}(\ell)}{dm_w^\ell} \right|_{\text{orthogonality}} = \begin{pmatrix} -1 \\ \frac{pk_w^\ell \dot{w}_i^\ell}{S_\ell^{**}[\lambda_j^W + k + n_\ell^W \dot{w}_i^\ell]} \end{pmatrix} \left(\frac{w_i^\ell \dot{w}_i^\ell}{y_\ell^W - y_{\ell-1}^{**}}\right), \quad (5.5.4.5)$$

and the right hand side of equation (5.5.4.5) is independent of $n_\ell^W$.

Define the $S^{**}$ measure of stability as

$$S^{**}_\ell = \left[ \frac{pk_w^\ell \dot{w}_i^\ell}{S_\ell^{**}[\lambda_j^W + k + n_\ell^W \dot{w}_i^\ell]} - 1 \right]^2 \quad (5.5.4.6)$$

where

$$S_\ell^{**} = \left| \frac{dm_w^\ell}{dn_\ell^W} \right|, \quad (5.5.4.7)$$

$$= \frac{pk_w^\ell \dot{w}_i^\ell}{\sum_{j=1}^{p} \frac{k_w^\ell \dot{w}_i^\ell}{(\lambda_j^W + k + n_\ell^W \dot{w}_i^\ell)^2}}.$$
Although $S^*_L$ and $S^{**}_L$ are appealing in their approach to stability of estimates, the computational aspects of repeated iterations will tend to discourage an analyst from using them.

5.6 Sign Change of an Estimated Parameter.

There are instances where the ridge regression solution to a problem contains an estimate with an incorrect sign. A reversal in the sign of a particular estimate is possible by judicious weighting of an observation. A single point ($w_i$) may be weighted such that an estimate known to carry an incorrect sign will have an acceptable sign after the weighting process. Let $\beta^*_j$ be the coefficient with the incorrect sign, and consider weighting the $i$th observation. By an argument that parallels the derivation of equation (3.6.9), it may be shown that the required weighting is any value of $n$ such that

\[
 n < \frac{\beta^2_j}{\beta^*_j w_i^T F w_i + \beta^*_j q_j^T F w_i (y_i - y_0^*)}.
\]  

(5.6.1)

5.7 Iterative Sequential n-weighting.

The iterative sequential technique starts by first determining an initial set of n-weightings obtained by reference to some criterion (for example, D-optimality of the variance-covariance matrix). The points are then reweighted in a sequential fashion, and the process repeated until no further improvement on the solution can be made. Consider the $l$th stage in the $m$th iteration and let the estimator, variance-covariance matrix, and bias-matrix be expressed in terms of the $\Omega$-matrices defined in Chapter 3.
Define

\[ U_m \] to be the \( \Omega^{-1} \) weighting for the mth iteration,

\[ X_m \] to be the X-matrix for the mth iteration,

\[ X_m = U_m \frac{1}{2} U_{m-1} \frac{1}{2} \cdots U_2 \frac{1}{2} U_1 \frac{1}{2} X \, , \]

\[ \hat{\beta}^* \{ m \} \] to be the weighted ridge regression estimate for the mth iteration,

For the first iteration define

\[ X_1 = \Omega^{-\frac{1}{2}} X, \]

\[ \hat{\beta}^* \{ 1 \} = (X_1' X_1 + kI)^{-1} X_1' Y_1, \]

\[ = (X'u_1' X + kI)^{-1} X'u_1' Y. \] \hfill (5.7.1)

It may be shown that

\[ \text{Var}[\hat{\beta}^* \{ 1 \}] = (X'u_1' X + kI)^{-1} X'u_1'^2 X(X'u_1' X + kI)^{-1} \sigma^2 \]

\hfill (5.7.2)

and reduces to equation (4.1.7) as

\[ U_1 = \Omega^{-1}. \]

Similarly, the bias-matrix

\[ B[\hat{\beta}^* \{ 1 \}] = k^2 (X'u_1' X + kI)^{-1} \hat{\beta}^* (X'u_1' X + kI)^{-1}. \]

\hfill (5.7.3)

reduces to equation (4.1.10).
For the $m$th iteration define,

$$
\tilde{\beta}^*[m] = (X^*U_{m}X + kI)^{-1}X^*U_{m}Y, \quad (5.7.4)
$$

$$
V[\tilde{\beta}^*[m]] = (X^*U_{m}X + kI)^{-1}U_{m}^2(X^*U_{m}X + kI)^{-1}G, \quad (5.7.5)
$$

$$
B[\tilde{\beta}^*[m]] = k^2(X^*U_{m}X + kI)^{-1}(X^*U_{m}X + kI)^{-1}, \quad (5.7.6)
$$

where

$$
U_{m} = \prod_{i=1}^{m}U_{i}.
$$

Now consider the $l$th sequential point in the $(m+1)$th iteration.

Define

$$
X_m = U_{m}^l X,
$$

and

$$
X_m^* = U_{m} X.
$$

In terms of the actual points, it should be noted that

$$
X_m^* = [w_{1,m}, \ldots, w_{n,m}],
$$

and

$$
X_m^* = [w_{1,m}^*, \ldots, w_{n,m}^*].
$$

Let

$$
F_\lambda^1(m+1) = \left[ (X_m^*X_m + kI) + \sum_{i=1}^{m} n_i w_{i,m-1,m}^* \right]^{-1}
$$

$$
= \left[ F_{\lambda-1}^1(m+1) + \sum_{i=m-k}^{m} n_i w_{i,m-k,m}^* \right]^{-1}, \quad (5.7.7)
$$
and

\[ G_{\ell}(m+1) = X_{m}^{*}X_{m}^{*} \]

\[ = G_{\ell-1}(m+1) + n_{\ell}(n_{\ell}+2)\mathcal{W}_{\ell,m}^{*} \mathcal{W}_{\ell,m}^{*}. \]  

(5.7.11)

The iterative analogs of Sections 5.1 and 5.3 are then easily established. For A-optimality of the variance-covariance matrix, the \( n \)-weighting for the \( \ell \)th point, \((m+1)\)th iteration, is given by

\[ n_{\ell} = \frac{d}{Ze - df}, \]

(5.7.12)

where

\[ d = \left[ c_{\ell}^{2} \mathcal{W}_{\ell,m}^{2} F_{\ell-1}(m) \mathcal{W}_{\ell,m} \right] - \left[ \mathcal{W}_{\ell,m}^{2} F_{\ell-1}(m) G_{\ell-1}(m) F_{\ell-1}(m) \mathcal{W}_{\ell,m} \right] \]

\[ - \left[ \mathcal{W}_{\ell,m}^{2} F_{\ell-1}(m) G_{\ell-1}(m) F_{\ell-1}(m) \mathcal{W}_{\ell,m} \right], \]

\[ e = \left[ c_{\ell}^{2} \mathcal{W}_{\ell,m}^{2} F_{\ell-1}(m) \mathcal{W}_{\ell,m} \right] + \left[ \mathcal{W}_{\ell,m}^{2} F_{\ell-1}(m) G_{\ell-1}(m) F_{\ell-1}(m) \mathcal{W}_{\ell,m} \right] \]

\[ - \left[ \mathcal{W}_{\ell,m}^{2} F_{\ell-1}(m) G_{\ell-1}(m) F_{\ell-1}(m) \mathcal{W}_{\ell,m} \right] \]

\[ f = \mathcal{W}_{\ell,m}^{2} F_{\ell-1}(m) \mathcal{W}_{\ell,m} \],

and where \( c_{\ell} \) is the \( \ell \)th element of \( U_{m} \).
For the criterion of D-optimality, the weighting for the \( l \)th point on the \((m+1)\)th iteration is given by

\[
\eta^*_l = \frac{w_{\xi,m}^* P^*_{\xi-1} \{m\} w_{\xi,m}^* - w_{\xi,m}^* C^*_{\xi-1} \{m\} w_{\xi,m}^*}{[1 - w_{\xi,m}^* P^*_{\xi-1} \{m\} w_{\xi,m}^*][w_{\xi,m}^* C^*_{\xi-1} \{m\} w_{\xi,m}^*]}.
\] (5.7.13)

The advantage in using an iterative scheme is that it may be used in conjunction with a secondary criterion to produce a given optimal stopping-rule of iterative sequences.

In summary, five different criteria for the selection of \( n \)-weightings have been discussed. The complexity in solution for optimal \( n \)-weighting with the stability criterion will cause the analyst to consider A-optimality and D-optimality in a more favorable light. If the final criterion of iterative sequential \( n \)-weighting is used, the technique often finds \( n \)-weightings with favorable secondary features.
6. SPECIFIC WEIGHTED RIDGE REGRESSIONS

The concept of weighting the observations will be extended to three specific situations. For each situation comparisons with the ridge estimator will be made. The first situation concerns the estimation of a function of the unknown parameters; the remaining situations concern differing variance structures of a regression model.

6.1 Functional Weighted Ridge Regression

The estimation of $\beta$ is sometimes not of primary importance, the principal interest being the estimation of a linear function of parameters which may be expressed as

$$ T = D' \hat{\beta}. $$

Suppose $t$ is the minimum mean square error estimator of $T$ with the understanding that the M.S.E.($t$) is equivalent to the expected Euclidean distance squared from $t$ to $T$. The eigenvalues $X'X$ are $\lambda_1, \ldots, \lambda_p$ and have the associated eigenvectors $\xi_1, \xi_2, \ldots, \xi_p$. The eigenvalues of $\tilde{X}'\tilde{X}$ are those of $X'X$ augmented by $(n+n_0 - p)$ zero eigenvalues with the eigenvectors corresponding to the "p" real roots being $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_p$, and the arbitrary eigenvectors corresponding to the $(n+n_0 - p)$ zero roots, $-p+1, -p+2, \ldots, -n+n_0-p$.

To establish the form of the estimator, first consider the single function case, with the estimator denoted by $\hat{\beta}^*_{-F(1)}$.

6.1.1. The single eigenvector case. Suppose $T$ is simply a function of the $i$th eigenvector and the unknown vector of coefficients;
\[ T = \tilde{\xi}_1^\prime \tilde{\beta}, \]

The estimator \( t \) may be expressed as

\[
t = \sum_{i=1}^{p} c_{i} \tilde{\xi}_i^\prime \tilde{X}^\prime \tilde{Y} + \sum_{v=1}^{n^*-p} d_{j} \tilde{\xi}_j^\prime \tilde{Y} \tag{6.1.1.1} \]

where \( c_i \) and \( d_j \) are constants, and \( n^* = n + n_o \). Then

\[
\text{M.S.E.}(t) = E(t - T)^2, \tag{6.1.1.2}
\]

where

\[
\sum_{i=1}^{p} c_{i} \tilde{\xi}_i^\prime \tilde{X}^\prime \tilde{Y} = \left[ c_{1} \tilde{\xi}_{1}^\prime \tilde{X}^\prime \tilde{Y} + \ldots + c_{p} \tilde{\xi}_{p}^\prime \tilde{X}^\prime \tilde{Y} \right]
\]

\[
+ c_{p+1} \tilde{\xi}_{p+1}^\prime \tilde{X}^\prime \tilde{Y} + \ldots + c_{p} \tilde{\xi}_{p}^\prime \tilde{X}^\prime \tilde{Y}.
\]

If equation (6.1.1.2) is expanded and all terms containing \( \tilde{\xi}_i \tilde{\xi}_j \) (where \( i \neq j \)) eliminated due to orthogonality of \( \tilde{\xi}_i \) with \( \tilde{\xi}_j \), then

\[
E(t - T)^2 = E\left( \sum_{i=1}^{p^*} c_{i} \tilde{\xi}_i^\prime \tilde{X}^\prime \tilde{Y} \right)^2 + E\left( \sum_{j=1}^{n^*-p} d_{j} \tilde{\xi}_j^\prime \tilde{Y} \right)^2 \tag{6.1.1.3}
\]

\[
E(c_{l} \tilde{\xi}_l \tilde{X}^\prime \tilde{Y} - \tilde{\xi}_l \tilde{\beta})^2.
\]
Consider each term of equation (6.1.1.3) separately. The term

\[ E \left\{ \sum_{i=1}^{p^*} c_i \tilde{\xi}_i X_i \tilde{Y} \right\}^2 \]

may be simplified by using the definition of the estimator \( \tilde{\beta} \),

\[ \tilde{X} \tilde{Y} = \tilde{X} \tilde{X} \tilde{\beta} \]

and so it follows that

\[ E \left\{ \sum_{i=1}^{p^*} c_i \tilde{\xi}_i X_i \tilde{X} \tilde{Y} \right\}^2 = E \left[ \sum_{i=1}^{p^*} c_i \tilde{\xi}_i X_i \tilde{X} \tilde{X} \tilde{\beta} \right] \]

(6.1.1.4)

Now, using the fact that, in general,

\[ E(z^2) = V(z) + [E(z)]^2 \]

it follows that

\[ E \left\{ \sum_{i=1}^{p^*} c_i \tilde{\xi}_i X_i \tilde{X} \tilde{X} \tilde{\beta} \right\}^2 = V \left[ \sum_{i=1}^{p^*} c_i \tilde{\xi}_i X_i \tilde{X} \tilde{\beta} \right] + E \left[ \sum_{i=1}^{p^*} c_i \tilde{\xi}_i X_i \tilde{X} \tilde{X} \tilde{\beta} \right]^2 \]

(6.1.1.5)

\[ = \sum_{i=1}^{p^*} c_i^2 \tilde{\xi}_i \tilde{X} \tilde{X}(\tilde{\beta}) \tilde{X} \tilde{X} \tilde{\beta} + \sum_{i=1}^{p^*} c_i^2 (\tilde{\xi}_i \tilde{X} \tilde{X} \tilde{\beta})^2 \]

which may be rewritten as

\[ E \left\{ \sum_{i=1}^{p^*} c_i \tilde{\xi}_i X_i \tilde{X} \tilde{X} \tilde{\beta} \right\}^2 = \sum_{i=1}^{p^*} c_i^2 \tilde{\xi}_i^2 \tilde{X} \tilde{X} \tilde{\beta} + \sum_{i=1}^{p^*} c_i^2 (\tilde{\xi}_i \tilde{X} \tilde{X} \tilde{\beta})^2 \]

(6.1.1.6)

using equation (3.4.8) and the fact that
\[ \bar{\xi}^* \bar{X}^* \bar{X}_\beta = \bar{\xi}^* \bar{\lambda} \bar{\lambda}_\beta \]

\[ = \lambda^* \alpha_1. \]

The term

\[ E \left[ \sum_{j=1}^{n^*} d_{j-j} \bar{Y} \right]^2 \]

may be expanded directly to obtain

\[ E \sum_{j=1}^{n^*} d_{j-j} \bar{Y}^2 = E \left[ \sum_{j=1}^{n^*} d_{j-j} \bar{X}_\beta \right]^2 + E \left[ \sum_{j=1}^{n^*} d_{j-j} \bar{\xi} \right]^2 \]

\[ = \sum_{j=1}^{n^*} d_{j-j} \bar{Y}^2 \]

(6.1.1.7)

The term \( E(c_{\lambda^*} \bar{X}^{\gamma*} \bar{Y} - \bar{\xi}^{\beta*})^2 \) of equation (6.1.1.3) is simplified if \( \bar{X}^{\gamma*} \bar{Y} \) is replaced by \( \bar{X}^* \bar{X}_\beta \), i.e.,

\[ E(c_{\lambda^*} \bar{X}^{\gamma*} \bar{Y} - \bar{\xi}^{\beta*})^2 = E\left( c_{\lambda^*} \bar{X}^{\gamma*} \bar{X}_\beta \right)^2 - 2(c_{\lambda^*} \bar{X}^{\gamma*} \bar{X}_\beta)(\bar{\xi}^{\beta*}) + (\bar{\xi}^{\beta*})^2 \]

\[ = c_{\lambda^*} \bar{X}^{\gamma*} \bar{X}_\beta \bar{X}_\beta \bar{\xi}^2 + (c_{\lambda^*} \bar{X}^{\gamma*} \bar{X}_\beta - \bar{\xi}^{\beta*})^2 \]

(6.1.1.8)

Now,

\[ \bar{\xi}^* \bar{X}^* \bar{X}_\beta = \lambda_1 \tilde{\alpha}_1, \]

and

\[ \bar{\xi}^{\beta*} = \tilde{\alpha}_1. \]
Equation (6.1.1.8) may therefore be written as

\[ E(c\tilde{\xi}_k\tilde{\xi}_k \tilde{X} \tilde{Y} - \tilde{\xi}_k \tilde{\xi}_k)^2 = (c\tilde{\lambda}_k - 1)^2 + c\tilde{\xi}_k \tilde{X} \tilde{V} \tilde{X} \tilde{\xi}_k \sigma^2. \]  

(6.1.1.9)

Equation (6.1.1.3) may be expressed as the sum of equations (6.1.1.6), (6.1.1.6), (6.1.1.7) and (6.1.1.9)

\[ E(t-T)^2 = \sum_{i=1}^{p^*} c_i^2 \left[ \tilde{\lambda}_1^2 a_1^2 + \tilde{\xi}_1^2 \tilde{X} \tilde{V} \tilde{X} \sigma^2 \right] + \sum_{j=1}^{n^*-p} C_d \sum_{j=*-1}^{n^*} \tilde{v}^2_{j-j} \tilde{v}_{j-j}^2 \]

\[ + (c\tilde{\lambda}_k - 1)^2 + c\tilde{\xi}_k \tilde{X} \tilde{V} \tilde{X} \tilde{\xi}_k \sigma^2. \]

(6.1.1.10)

Now, \( E(t-T)^2 \) of equation (6.1.1.10) should be minimized with respect to \( c_k \). As \( \tilde{V} \) and \( \tilde{X} \tilde{V} \tilde{X} \) are positive definite, it follows that \( c_i \) (for \( i = 1, \ldots, p \) except \( k \)), and \( d_j \) (for \( j = 1, \ldots, n^*-p \)), should be set equal to zero, and the \( c_k \) term minimized directly. Differentiating equation (6.1.1.10) with respect to \( n_k \), equating to zero and solving for \( c_k \) yields

\[ c_k = \frac{\tilde{\xi}_k \tilde{X} \tilde{V} \tilde{X} \tilde{\xi}_k \sigma^2}{\tilde{\lambda}_k a_k^2 + \tilde{\xi}_k \tilde{X} \tilde{V} \tilde{X} \tilde{\xi}_k \sigma^2} \]

\[ = \left\{ \frac{\tilde{\xi}_k \tilde{X} \tilde{V} \tilde{X} \tilde{\xi}_k \sigma^2}{\tilde{\lambda}_k a_k} \right\}^{-1}. \]

(6.1.1.11)

with all other coefficients zero. The minimum M.S.E. estimator of

\[ T = \tilde{\xi}_k \tilde{\lambda}_k \] is therefore
\[ t = c \tilde{X}^\top \tilde{Y} \]
\[ = \left[ \lambda_\tilde{\xi} + \frac{\tilde{X}^\top \tilde{V} \tilde{X} \lambda_\tilde{\xi} \sigma^2}{\tilde{\lambda}_\xi \tilde{\alpha}_\xi^2} \right]^{-1} \tilde{X}^\top \tilde{Y} \]
\[ = \left( \frac{1}{\tilde{\lambda}_\xi + \tilde{k}_\xi} \right) \tilde{\xi} \tilde{X}^\top \tilde{Y}. \]

For any \( \tilde{\xi}_\tilde{\beta} \),

\[ \tilde{\xi}_\tilde{\beta} \tilde{\beta} = \frac{1}{\tilde{\lambda}_1} \tilde{\xi}_\tilde{\beta} \tilde{X}^\top \tilde{Y}, \]

and is unbiased for \( \tilde{\xi}_\tilde{\beta} \).

Equations (6.1.1.12) and (6.1.1.13) may be combined to form a single estimator by noting that

\[ \tilde{P} \tilde{\beta} \] is estimated by \([\Lambda + K]^{-1} \tilde{P} \tilde{X} \tilde{Y}, \]

where \( K \) is a diagonal matrix with diagonal elements,

\[ \tilde{\lambda}_1^{-1}, \tilde{\lambda}_2^{-1}, \ldots, \tilde{\lambda}_{p-1}^{-1}, \tilde{\lambda}_p^{-1}, (\tilde{\lambda}_\xi + \tilde{k}_\xi)^{-1}, \tilde{\lambda}_{p+1}^{-1}, \ldots, \tilde{\lambda}_p^{-1} \]

and the \( \tilde{k}_\xi \) defined as

\[ \tilde{k}_\xi = \frac{\sigma^2}{\tilde{\alpha}_\xi^2}. \]
It follows that the minimum M.S.E. estimator of \( T = \hat{\beta} \) is given by
\[
t = \hat{\hat{\beta}}^*_{LF} ,
\]
where
\[
\hat{\hat{\beta}}_{LF} = (\hat{X}'\hat{X} + k \hat{\hat{\beta}}_{LF}' \hat{\hat{\beta}}_{LF})^{-1} \hat{X}'\hat{Y} .
\]
(6.1.1.16)

Notice how equation (6.1.1.16) is similar to the ordinary weighted ridge estimator except that a specific value for \( k \) is demanded and that the identity matrix used in \( \hat{\hat{\beta}}^* \) is replaced by \( \hat{\hat{\beta}}_{LF}' \hat{\hat{\beta}}_{LF} \).

6.1.2. The \( p \)-eigenvectors case. The extension to the more general case where \( T \) cannot be expressed as a single function of one of the eigenvectors will now be investigated. Again, the minimum M.S.E. estimator, \( t = \hat{d}'\hat{\hat{\beta}}^* \), of \( T = \hat{d}'\hat{\beta} \) is required.

Notice that \( \hat{d} \) may be written as a linear function of the eigenvectors of \( \hat{X}'\hat{X} \), i.e.,
\[
d = \sum_{i=1}^{p} \hat{e}_i \hat{x}_i ,
\]
(6.1.2.1)
\[
= \hat{P}_{\hat{\hat{\beta}}} ,
\]
where \( \hat{e}_i \) is a vector of constants.

The estimator \( t \) may therefore be written as
\[
t = \sum_{i=1}^{p} \hat{e}_i \hat{c}_i \hat{x}_i \hat{Y} + \sum_{j=1}^{n-p} \hat{d}_j \hat{\hat{\beta}}^*_{LF} ,
\]
(6.1.2.2)

and \( E(t-T)^2 \) may be expressed as
\[ E(t-T)^2 = E[\sum_{i=1}^{p} f_i \xi_i f_i \Xi_i \Upsilon - f_i \xi_i \beta] + \sum_{j=1}^{n^* - p} d_j \xi_j \gamma_j^2. \quad (6.1.2.3) \]

Using arguments similar to those used to derive equations (6.1.1.6), (6.1.1.7) and (6.1.1.9),

\[ E(t-T)^2 = \sum_{i=1}^{p} f_i^2 (c_i \lambda_i - 1) \alpha_i ^2 + \sum_{i=1}^{p} c_i f_i \lambda_i \Xi_i \Upsilon_i \xi_i \alpha_i ^2 + \sum_{j=1}^{n^* - p} d_j \gamma_j \beta_j^2. \quad (6.1.2.4) \]

By an argument paralleling that used to derive equation (6.1.1.11), it may be shown that

\[ c_i = \left[ \lambda_i - \frac{\xi_i \lambda_i \Xi_i \Upsilon_i}{\lambda_i \kappa_i} \right]^{-1} \quad i=1, \ldots, p \quad (6.1.2.5) \]

It follows that

\[ t = \frac{1}{\lambda_i \kappa_i} \xi_i \Upsilon_i \]

\[ = \left[ \frac{1}{\lambda_i + \kappa_i} \right] \xi_i \Upsilon_i, \]

and so, \( \xi_i \) is estimated by

\[ \xi_i = \left[ \frac{1}{\lambda_i + \kappa_i} \right] \Upsilon_i. \]

It follows the complete form of the estimator may be written as

\[ \hat{\beta}_F = \left( \hat{\Xi}_i + \sum_{i=1}^{p} \xi_i \hat{\xi}_i \right) \left( \frac{1}{\lambda_i + \kappa_i} \right) \hat{\Upsilon}_i. \quad (6.1.2.6) \]
The estimator $\hat{\beta}_F^*$ may be written in canonical form as

$$\hat{\beta}_F^* = \tilde{P}(\tilde{\Lambda} + K)^{-1}\tilde{P}^\tau \tilde{X}^\tau \tilde{Y},$$

and if

$$\tilde{a}_F^* = \tilde{P}^\tau \tilde{a}_F^*, $$

then

$$\tilde{a}_F^* = (\tilde{\Lambda} + K)^{-1}\tilde{P}^\tau \tilde{X}^\tau \tilde{Y}. \quad (6.1.2.7)$$

If $\tilde{Z} = \tilde{X}\tilde{P}$ such that $\tilde{Z}^\tau \tilde{Z} = \Lambda$, equation (6.1.2.6) may be written as

$$\tilde{a}_F^* = (\tilde{Z}^\tau \tilde{Z} + K)^{-1}\tilde{Z}^\tau \tilde{Y},$$

and hence $\tilde{a}_F^*$ may be regarded as being ridge regression applied to weighted canonical variables.

Further, notice that $\tilde{X}_{1i}^\tau \tilde{X}_{1i}$ is only another form of $\tilde{P}^\tau \tilde{P}$, and so $\tilde{\beta}_F^* = (\tilde{X}^\tau \tilde{X} + \tilde{P}^\tau \tilde{P})^{-1}\tilde{X}^\tau \tilde{Y}$ is then generalized ridge regression with specific values for $k_1$. From consideration of equation (6.1.2.5),

$$k_1 = \frac{\tilde{X}_{1i}^\tau \tilde{X}_{1i} \tilde{\lambda}_1^2}{\tilde{\lambda}_1^2}, \quad (6.1.2.8)$$

For ridge regression, Hoerl and Kennard (1976) argued for the use of a single $k$-value of

$$k = \frac{p\sigma^2}{\tilde{\beta}^2} . \quad (6.1.2.9)$$
and if equation (6.1.2.8) is considered for the case of standard ridge regression, then

$$k_i = \frac{\xi_i^2 \lambda_i \sigma^2}{\lambda_i \alpha_i^2}, \quad (6.1.2.10)$$

and equation (6.1.2.9) may only be attained if all $\alpha_i^2$ are assumed equal, which is rarely to be found in practice. The disadvantage functional ridge regression is that parameter estimates are demanded for determination of $k$-value.

6.2 Prior Ridge Regression.

Consider the situation where it is known that some observations are more precise (in a variance sense) than others. Assume that $E(\xi) = 0$ and $V(\xi) = \sigma^2 M^{-2}$ where $M$ is a known non-singular matrix of correlation coefficients.

The Aitken least squares unbiased estimate of $\beta$ is

$$\hat{\beta}_A = (X'M^{-1}X)^{-1}X'M^{-1}Y,$$

with variance-covariance $(X'M^{-1}X)^{-1} \sigma^2$.

Use of the generalized ridge regression technique on the Aitken least squares structure yields

$$\hat{\beta}^*_A = (X'M^{-1}X + K)^{-1}X'M^{-1}Y, \quad (6.2.1)$$

and the variance-covariance matrix, and bias-matrix are given by

$$V(\hat{\beta}^*_A) = (X'M^{-1}X + K)^{-1}X'M^{-1}X(X'M^{-1}X + K)^{-1} \sigma^2 \quad (6.2.2)$$
and
\[ B(\beta_{AK}^*) = (X'M^{-1}X + K)^{-1}K\beta\gamma K(XM^{-1}X + K)^{-1} \quad (6.2.3) \]

The disturbing feature lies in \(M^{-1}\). If the initial set of observations are such that there exists high correlations between responses, then this \(M\) will be nearly singular and hence the computed \(M^{-1}\) unreliable. A possible solution would be to investigate the use of ridge-technique on \(M\) and then continue with standard ridge regression on this result.

Let \(K_A^*\) denote the \(n_0 \times n_0\) matrix to be added to \(M\) prior to its inversion in calculating equation (6.2.1). Then
\[ \beta_{AK}^* = [X'(M + K_A^*)^{-1}X + K^{-1}X'(M + K_A^*)^{-1}Y] \quad (6.2.4) \]

with
\[ V(\beta_{AK}^*) = M^*[X'(M + K_A^*)^{-1}M(M + K_A^*)^{-1}X]M^* \sigma^2 \quad (6.2.5) \]

where
\[ M^* = [X'(M + K_A^*)^{-1}X + K]^{-1} \]

The bias-matrix is given by
\[ B(\beta_{AK}^*) = M^*K\beta\gamma K^*M^* \quad (6.2.6) \]

Suppose \(K_A^*\) is a diagonal matrix, all elements equal to \(k_\gamma^*\) and \(K\) a diagonal matrix, all elements equal to the positive value \(k\).
Equation (6.2.4) may then be written

$$\beta^*_A = M_A [X' (M + k_A I)^{-1} X]^{-1} Y,$$  \hspace{1cm} (6.2.7)

where

$$M_A = [X' (M + k_A I)^{-1} X + kI]^{-1}.$$  

The variance-covariance matrix and bias-matrix are

$$V(\beta^*_A) = M_A [X' (M + k_A I)^{-1} M (M + k_A I)^{-1} X] M_A^{-1},$$ \hspace{1cm} (6.2.8)

and

$$B(\beta^*_A) = k^2 M_A \beta^*_A M_A^{-1},$$ \hspace{1cm} (6.2.9)

respectively.

Define the estimator resulting from using ridge regression on the Aitken least squares structure as

$$\beta^*_A = (X' M^{-1} X + kI)^{-1} X' M^{-1} Y,$$ \hspace{1cm} (6.2.10)

with bias-matrix

$$B(\beta^*_A) = k^2 (X' M^{-1} X + kI)^{-1} \beta^*_A (X' M^{-1} X + kI)^{-1}.$$ \hspace{1cm} (6.2.11)

Consider the difference \(D_B\) between traces of the bias-matrices of the estimators. Equation (6.2.11) minus equation (6.2.8) is then

$$D_B = k^2 \beta^*_A (X' M^{-1} X + kI)^{-2} - (X' (M + k_A I)^{-1} X + kI)^{-2} \beta^*_A.$$ \hspace{1cm} (6.2.12)
Equation (6.2.12) will be positive definite if the kernel

\[ [X^{\prime}X^{-1}X + kI]^{-1} - [X^{\prime}(M + k_* I)^{-1} X + kI]^{-1} \] (6.2.13)

is positive definite. From Graybill (1969, p. 330), equation (6.2.13) will be positive definite if

\[ Q = [X^{\prime}(M + k_* I)^{-1} X + kI] - [X^{\prime}X^{-1}X + kI] \] , (6.2.14)

is positive definite.

Now,

\[ (M + k_* I)^{-1} = M^{-1} - k_* M^{-1} [I + k_* M^{-1}]^{-1} M^{-1} \] , (6.2.15)

and so equation (6.2.14) may be expressed as

\[ Q = -k_* X^{\prime}X^{-1}P_m [I + k_* A_m^{-1}]^{-1} P_m^{\prime}X^{-1} \] , (6.2.16)

where \( M = P_m A_m P_m^{\prime} \). Notice that \( Q \) of equation (6.6.16) can be positive definite only if \( k_* \) is negative and \( [I + k_* A_m^{-1}]^{-1} \) is positive definite and is possible only if

\[ -\lambda_{pm} < k_* < 0 \] , (6.2.17)

where \( \lambda_{pm} \) is the smallest root of \( M \). The implication of equation (6.2.17) is clear. As \( M \) becomes more and more ill-conditioned \( (\lambda_{pm} \to 0) \), the existence of a \( k_* \) to achieve a decrease in trace bias-

matrix becomes confined to a very small negative range.

The variance-covariance matrix for the estimator defined by equation (6.2.10) is

\[ V(\beta_A) = [X^{\prime}X^{-1}X + kI]^{-1}X^{\prime}X^{-1}\Sigma^{-1}X[X^{\prime}X^{-1}X + kI]^{-1} \] , (6.2.18)
Consider the difference \((D_v)\) between the traces of the variance-covariance matrices. Equation (6.2.8) minus equation (6.2.18) is then

\[
D_v = \text{tr}[M_A (X'M + k_* I)^{-1} M (M + k_* I)^{-1} X'M_A] \sigma^2
\]

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

(6.2.19)

As \(M_A = (X'M^{-1}X + kI)^{-1}\) is positive definite

\[
D_v > \text{tr}[(X'M^{-1}X + kI)^{-1} X(M + k_* I)^{-1} X'(M + k_* I)^{-1} X^{-1} X'(X'M^{-1}X + kI)^{-1} X^{-1} X] \sigma^2.
\]

(6.2.20)

\(D_v\) of equation (6.2.20) may now be made positive by making the kernel of equation (6.2.20) positive definite and examination of this kernel reveals the condition

\[k_* < -\frac{1}{2} \lambda_{pm}.
\]

(6.2.21)

From equations (6.2.17) and (6.2.21) the conclusion to be drawn is that prior ridge regression will be a trace mean square error improvement over ridge regression for any choice of \(k_*\) such that

\[-\lambda_{pm} < k_* < -\frac{1}{2} \lambda_{pm}.
\]

(6.2.22)

6.3 Variance Weighted Ridge Regression.

Suppose the data points are uncorrelated but the variance (i.e., in a sense, precision) for point \(i\) equals \(\sigma_i^2\). Define

\[
\beta_C^* = (X'C^{-1}X + kI)^{-1} X'C^{-1}Y
\]

(6.3.1)
where $C$ is a diagonal matrix with $i$th diagonal element $c_i^2$. Define the variance weighted ridge regression as

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

$$
= (X'C^{-1}X + \sum_{i=1}^{n o} \frac{n_i}{c_i^2} w_i w_i') + kI)^{-1}(X'C^{-1}Y + \sum_{i=1}^{n o} \frac{n_i}{c_i^2} w_i w_i')
$$

The sequential variance weighted ridge estimator is defined as

$$
\hat{\beta}_c(t) = F_{c t}(X'C^{-1}Y + \sum_{i=1}^{t} n_i w_i w_i') ,
$$

(6.3.3)

where

$$
F_{ct} = (X'C^{-1}X + kI + \sum_{i=1}^{t} n_i w_i w_i')^{-1}
$$

and

$$
n_* = n c^{-2}
$$

This estimator has a variance-covariance matrix

$$
V(\hat{\beta}_c(t)) = F_{ct}(X'C^{-1}X + \sum_{i=1}^{t} n_i^* (n_i^* + 2) y_i w_i) F_{ct}^T
$$

(6.3.5)

and a bias-matrix

$$
B(\hat{\beta}_c(t)) = k^2 F_{ct} 3 \beta F_{ct}
$$

(6.3.6)
The specific \( n \)-weightings may be determined in a manner analogous to those found in Sections 5.1 to 5.3. As an example, for the criterion of minimal trace variance-covariance matrix, the \( n_{\ell} \)-weighting is

\[
 n_{\ell} = \frac{-k \bar{w}^{3}_{\ell} F_{c(\ell-1)} \bar{w}_{\ell}}{k \{ (\bar{w}^{3}_{\ell} F_{c(\ell-1)} \bar{w}_{\ell}) (\bar{w}^{3}_{\ell} F_{c(\ell-1)} \bar{w}_{\ell}) - (\bar{w}^{2}_{\ell} F_{c(\ell-1)} \bar{w}_{\ell})^2 \} + (c_{\ell}^2 - \bar{w}^{2}_{\ell} F_{c(\ell-1)} \bar{w}_{\ell}) (\bar{w}^{2}_{\ell} F_{c(\ell-1)} \bar{w}_{\ell})}
\]

(6.3.6)

The trace of equation (6.3.4), with the \( n_{\ell} \) of equation (6.3.6) is given by

\[
 \text{tr} \, V[\tilde{\rho}_{c(\ell)}] = \text{tr}[V[\tilde{\rho}_{c(\ell-1)}]] - \frac{k^2 (\bar{w}^{3}_{\ell} F_{c(\ell-1)} \bar{w}_{\ell})^2}{(\bar{w}^{2}_{\ell} F_{c(\ell-1)} \bar{w}_{\ell}) [c_{\ell}^2 - \bar{w}^{2}_{\ell} (I+k F_{c(\ell-1)}) F_{c(\ell-1)} \bar{w}_{\ell}]}
\]

(6.3.7)

Equation (6.3.7) may be shown to be greater than or equal to zero, thus indicating negative \( n_{\ell} \)-weightings achieve a diminution in sequential trace variance-covariance. Similar expressions may be found for the criteria of D-optimality of variance and A-optimality of bias-matrix.
7. AN EXAMPLE

Some of the techniques developed in earlier chapters will be demonstrated using an example. The parameter values are assumed known in order for a comparison of bias to be made. These parameters chosen came from repeated clinical experiments involving mice. There were nine observations and five parameters to be estimated. The data had been standardized so as to produce $X'X$ in correlation form and the error terms were drawn from a $N(0,1)$ population. Table 7.1 shows the initial values used and Table 7.2 gives the transformed values.

Applying O.L.S. to the transformed data yields

$$
\hat{\beta} = \begin{bmatrix}
-3.30352 \\
5.83488 \\
4.18419 \\
0.58846 \\
3.42357
\end{bmatrix},
$$

The residual sum of squares is 2.7061 and the estimated variance, 0.67653. The Euclidean distance from estimate to parameter is 12.2374. The F-test for regression is significant for $\alpha = 0.01$. The variance-covariance matrix of the estimator is

$$
V(\hat{\beta}) = \begin{bmatrix}
12.27670 & 0.72300 & -11.86151 & 3.07188 & -0.59535 \\
5.08051 & -5.13252 & 1.79507 & 0.29601 & \\
16.58890 & -4.54279 & -0.24945 & & \\
\text{symmetric} & 2.34224 & 0.22013 & & \\
& & & 1.31964 &
\end{bmatrix}.
$$

The trace of the variance-covariance matrix is 37.6079, and the determinant, 78.6356. For an orthogonal situation, the trace of the variance-covariance matrix should equal 5, and the determinant equal 1. As the O.L.S. values are much larger than their orthogonal equivalents,
Table 7.1. Initial values for the example \((p = 5, n_o = 9)\).

\[
\begin{bmatrix}
1 & 1 & 1 & -1 & 1 \\
1 & 2 & 1 & 0 & 0 \\
1 & 3 & 1 & -2 & -2 \\
2 & 2 & 2 & -1 & 3 \\
2 & 3 & 3 & 0 & 1 \\
2 & 4 & 2 & -2 & 4 \\
4 & 4 & 3 & -2 & 3 \\
4 & 5 & 4 & -1 & 0 \\
4 & 6 & 4 & -1 & 4
\end{bmatrix}, \quad
\begin{bmatrix}
\beta_{\text{initial}} \\
[-0.2] \\
0.8 \\
1.0 \\
0.4 \\
0.6 \end{bmatrix}, \quad
\begin{bmatrix}
Y(\text{initial}) \\
0.21549 \\
3.36302 \\
2.73027 \\
4.52873 \\
6.15905 \\
6.32341 \\
5.27558 \\
7.67087 \\
11.21620
\end{bmatrix}
\]

\[
\text{Variance}(\sigma^2) = 1.0
\]
Table 7.2. Transformed values for the example \((p = 5, \, n_o = 9)\).

\[
X = \begin{bmatrix}
-0.356348 & -0.521749 & -0.384900 & 0.050252 & -0.094967 \\
-0.356348 & -0.298142 & -0.384900 & 0.502520 & -0.265908 \\
-0.356348 & -0.074536 & -0.384900 & -0.402015 & -0.607790 \\
-0.089087 & -0.298142 & -0.096225 & 0.050252 & 0.246915 \\
-0.089087 & -0.074536 & 0.192450 & 0.502520 & -0.094967 \\
-0.089087 & 0.149071 & -0.096225 & -0.402015 & 0.417855 \\
0.445435 & 0.149071 & 0.192450 & -0.402015 & 0.246915 \\
0.445435 & 0.372678 & 0.481125 & 0.050252 & -0.265908 \\
0.445435 & 0.596285 & 0.481125 & 0.050252 & 0.417855
\end{bmatrix}
\]

\[
Y = \begin{bmatrix}
-5.06036 \\
-1.91282 \\
-2.54558 \\
-0.74711 \\
0.88320 \\
1.04756 \\
0.00027 \\
2.39503 \\
5.94034
\end{bmatrix}, \quad \text{Beta} = \begin{bmatrix}
-0.74833 \\
3.57771 \\
3.46410 \\
0.88443 \\
3.50999
\end{bmatrix}
\]

\[
X^\prime X = \begin{bmatrix}
1.00000 & 0.836660 & 0.925820 & -0.201456 & 0.472087 \\
0.836660 & 1.00000 & 0.839146 & -0.269680 & 0.356753 \\
0.925820 & 0.839146 & 1.00000 & 0.043519 & 0.411220 \\
-0.201456 & -0.269680 & 0.043519 & 1.00000 & -0.188982 \\
0.472087 & 0.356753 & 0.411220 & -0.188982 & 1.00000
\end{bmatrix}
\text{ (symmetric)}
\]

The eigen values of \(X^\prime X\) are 3.036040, 1.045120, 0.725906, 0.157982, 0.034948.
ridge regression is used. Table 7.3 illustrates the major properties of ridge estimation for various k-values. The behavior of the individual regression coefficients are displayed as a ridge-plot in Figure 7.1. As the k-value increases in magnitude, the coefficients tend to zero non-monotonically.

Weighted ridge regression can be used to produce an estimator that is superior (with a given criterion) to that obtained by ridge regression. With a non-stochastic k-value, the n-weightings are selected sequentially according to the methods of Chapter 5. The specific order of points weighted is by that point that yields maximum improvement for a selected criterion. After all points have been weighted, the procedure is repeated until further improvement cannot be achieved.

For the criterion of A-optimality of the variance-covariance matrix, negative n-weightings decline from zero to -1 on each successive iteration (Table 7.4). The physical interpretation of this decline is that the data points are being "discounted" in importance on each successive iteration until \( \hat{\beta} \) is estimated by the zero vector owing to the complete absence of data. The decreasing trace of the variance matrix is offset to some degree by the increasing of bias, and is illustrated in Figure 7.2. However, caution must be used in its interpretation. It is tempting to conclude that monotonicity holds for both variance and bias-squared as the points are sequentially weighted. This is not the case as can be seen in Figure 7.3. The behavior of individual estimated coefficients is illustrated in Figure 7.4. In each instance, the final beta-value is closer to zero than its initial value, but the path taken by intermediate steps is not monotone. If Figure 7.2 is examined, the maximum value of \( \text{tr}(\mathcal{R}) \) is \( \hat{\beta}^T \hat{\beta} \), equal to 38.4622 in
<table>
<thead>
<tr>
<th>k</th>
<th>0.0005</th>
<th>0.010</th>
<th>0.015</th>
<th>0.020</th>
<th>0.025</th>
<th>0.030</th>
<th>0.035</th>
<th>0.040</th>
<th>0.045</th>
<th>0.050</th>
</tr>
</thead>
<tbody>
<tr>
<td>tr(V)</td>
<td>37.6</td>
<td>30.4</td>
<td>25.5</td>
<td>21.9</td>
<td>19.1</td>
<td>16.3</td>
<td>14.6</td>
<td>13.5</td>
<td>12.7</td>
<td>12.0</td>
</tr>
<tr>
<td>det(V)</td>
<td>78.6</td>
<td>55.0</td>
<td>39.9</td>
<td>29.9</td>
<td>22.6</td>
<td>17.0</td>
<td>13.4</td>
<td>10.9</td>
<td>8.8</td>
<td>7.3</td>
</tr>
<tr>
<td>tr(B)</td>
<td>0.081</td>
<td>0.263</td>
<td>0.490</td>
<td>0.736</td>
<td>0.987</td>
<td>1.236</td>
<td>1.480</td>
<td>1.714</td>
<td>1.943</td>
<td>2.169</td>
</tr>
<tr>
<td>det(M.S.E.)</td>
<td>78.6</td>
<td>55.0</td>
<td>39.9</td>
<td>29.9</td>
<td>22.6</td>
<td>17.0</td>
<td>13.4</td>
<td>10.9</td>
<td>8.8</td>
<td>7.3</td>
</tr>
<tr>
<td>E.dist.sq.</td>
<td>12.2</td>
<td>9.8</td>
<td>8.1</td>
<td>6.9</td>
<td>5.9</td>
<td>5.2</td>
<td>4.6</td>
<td>4.2</td>
<td>3.8</td>
<td>3.6</td>
</tr>
<tr>
<td>Res.Sq.</td>
<td>2.706</td>
<td>2.719</td>
<td>2.752</td>
<td>2.850</td>
<td>2.908</td>
<td>2.978</td>
<td>3.036</td>
<td>3.103</td>
<td>3.177</td>
<td>3.241</td>
</tr>
</tbody>
</table>

Note: k = 0 is the OLS solution.  
tr(V) = trace(variance-covariance matrix).  
det(V) = determinant(variance-covariance matrix).  
tr(B) = trace(bias matrix).  
det(M.S.E.) = |det(V)| + tr(B).  
E.dist.sq. = Euclidean distance-squared, estimate to parameter.  
Res.Sq. = Residual sum of squares.
Table 7.4. Iterative n-weightings for A-optimality of the variance-covariance matrix.

<table>
<thead>
<tr>
<th>Point Number</th>
<th>0th iteration</th>
<th>1st iteration</th>
<th>2nd iteration</th>
<th>3rd iteration</th>
<th>4th iteration</th>
<th>5th iteration</th>
<th>6th iteration</th>
<th>7th iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>-0.554519</td>
<td>-0.669375</td>
<td>-0.786939</td>
<td>-0.771083</td>
<td>-0.901830</td>
<td>-0.986857</td>
<td>-0.999985</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>-0.595583</td>
<td>-0.808176</td>
<td>-0.818234</td>
<td>-0.841049</td>
<td>-0.910934</td>
<td>-0.992629</td>
<td>-0.999970</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>-0.596144</td>
<td>-0.749709</td>
<td>-0.728652</td>
<td>-0.876286</td>
<td>-0.941275</td>
<td>-0.996484</td>
<td>-0.999999</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>-0.523406</td>
<td>-0.617180</td>
<td>-0.687183</td>
<td>-0.792396</td>
<td>-0.932013</td>
<td>-0.997265</td>
<td>-0.999998</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>-0.587026</td>
<td>-0.828034</td>
<td>-0.729486</td>
<td>-0.853187</td>
<td>-0.966827</td>
<td>-0.999999</td>
<td>-1.000000</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>-0.661790</td>
<td>-0.595277</td>
<td>-0.701792</td>
<td>-0.793460</td>
<td>-0.951005</td>
<td>-0.996362</td>
<td>-0.999997</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>-0.736564</td>
<td>-0.831780</td>
<td>-0.863564</td>
<td>-0.829468</td>
<td>-0.884204</td>
<td>-0.998868</td>
<td>-1.000000</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>-0.665498</td>
<td>-0.775772</td>
<td>-0.790434</td>
<td>-0.817356</td>
<td>-0.934152</td>
<td>-0.998843</td>
<td>-0.999999</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>-0.466295</td>
<td>-0.663489</td>
<td>-0.680536</td>
<td>-0.779326</td>
<td>-0.837266</td>
<td>-0.985826</td>
<td>-0.999849</td>
</tr>
<tr>
<td>tr(V)</td>
<td>25.5034</td>
<td>16.8648</td>
<td>6.20498</td>
<td>2.33323</td>
<td>0.519097</td>
<td>0.029791</td>
<td>0.000090</td>
<td>0.000000</td>
</tr>
<tr>
<td>tr(B)</td>
<td>0.263772</td>
<td>1.18031</td>
<td>4.7156</td>
<td>9.24916</td>
<td>15.393</td>
<td>29.2681</td>
<td>38.2942</td>
<td>38.4622</td>
</tr>
<tr>
<td>tr(M.S.E.)</td>
<td>25.7611</td>
<td>18.0451</td>
<td>10.9206</td>
<td>11.5824</td>
<td>15.9121</td>
<td>29.2979</td>
<td>38.2942</td>
<td>38.4622</td>
</tr>
</tbody>
</table>

Note: The ith n-weight refers to the ith row of X, 

\[ k = 0.01 \].
Figure 7.1. The ridge-plot for estimates in standard ridge regression.
Figure 7.2. Iterative n-weightings for A-optimality of the variance-covariance matrix.
Figure 7.3. Individual weighting of points for A-optimality of the variance-covariance matrix.
Figure 7.4. Beta-coefficients as a function of n-weights chosen. $k = 0.01$. 
in this instance, the minimum value of \( tr(V) \) being zero and occurring together with the maximum value of \( tr(B) \). Empirically, it should be noted that the rate of decrease in trace of the variance matrix is large for the first two or three iterations, while the rate of increase for trace of the bias matrix is relatively small. This suggests the use of only two iterations to reduce variance when actual values for bias-squared are not known.

For the criterion of D-optimality of the variance-covariance matrix, negative n-weightings are determined. On the sequential recalculation of n-weightings, it is found that weighting further than the second iteration fails to decrease the determinant of the variance-covariance matrix (Table 7.5).

The specific n-weightings are a function of the k-value initially chosen and Table 7.6 illustrates the steady increase of negative initial n-weightings as the k-value increases. Somewhat similar characteristics may be observed for the D-optimality of variance (Table 7.7). For this particular example, the Euclidean distance-squared from estimate to parameter decreases as a function of the number of points included in the iteration step. Investigation of other examples showed this not to be a consistent phenomenon however.

Reference to Figure 7.2 indicates the largest reduction in the trace of the variance-covariance matrix occurs when proceeding from iteration 1 to iteration 2 (a 42.88\% reduction). If the problem is recalculated as a standard ridge regression problem with a k-value chosen such that the new trace of the variance-covariance matrix equals the trace of the variance-covariance matrix of the weighted ridge estimator (2nd iteration), bias comparisons may be made. Direct comparison of the trace of the bias matrix by inspection of the difference of two quadratic forms is
Table 7.5. Iterative n-weightings for D-optimality of the variance-covariance matrix.

<table>
<thead>
<tr>
<th>Point Number</th>
<th>n-weightings</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0&lt;sup&gt;th&lt;/sup&gt; iteration</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; iteration</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; iteration</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>-0.263390</td>
<td>-0.375763</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>-0.449882</td>
<td>-0.467134</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>-0.271684</td>
<td>-0.368888</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>-0.208880</td>
<td>-0.366138</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>-0.355609</td>
<td>-0.437679</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>-0.288008</td>
<td>-0.407156</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>-0.452579</td>
<td>-0.504362</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>-0.333887</td>
<td>-0.445438</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>-0.204051</td>
<td>-0.338000</td>
</tr>
<tr>
<td>det(V)</td>
<td>39.8767</td>
<td>28.5215</td>
<td>15.4789</td>
</tr>
<tr>
<td>tr(B)</td>
<td>0.263772</td>
<td>0.571228</td>
<td>1.25784</td>
</tr>
<tr>
<td>det(M.S.E.)</td>
<td>40.1404</td>
<td>29.0927</td>
<td>16.7367</td>
</tr>
</tbody>
</table>

Note: The i<sup>th</sup> n-weight refers to the i<sup>th</sup> row of X.

k = 0.01.
<table>
<thead>
<tr>
<th>Point Number</th>
<th>k = 0.005</th>
<th>k = 0.010</th>
<th>k = 0.015</th>
<th>k = 0.020</th>
<th>k = 0.025</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-0.360429</td>
<td>-0.554319</td>
<td>-0.654249</td>
<td>-0.692387</td>
<td>-0.712682</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.428312</td>
<td>-0.595583</td>
<td>-0.683996</td>
<td>-0.733219</td>
<td>-0.765391</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-0.440258</td>
<td>-0.595414</td>
<td>-0.686941</td>
<td>-0.736271</td>
<td>-0.769325</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.445491</td>
<td>-0.595674</td>
<td>-0.687494</td>
<td>-0.736704</td>
<td>-0.769325</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.453066</td>
<td>-0.596069</td>
<td>-0.689794</td>
<td>-0.738992</td>
<td>-0.770237</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-0.463727</td>
<td>-0.596934</td>
<td>-0.692656</td>
<td>-0.742866</td>
<td>-0.773655</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-0.469271</td>
<td>-0.597537</td>
<td>-0.694253</td>
<td>-0.744666</td>
<td>-0.774953</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-0.469285</td>
<td>-0.597537</td>
<td>-0.694253</td>
<td>-0.744666</td>
<td>-0.774953</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-0.469285</td>
<td>-0.597537</td>
<td>-0.694253</td>
<td>-0.744666</td>
<td>-0.774953</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Point Number</th>
<th>k = 0.035</th>
<th>k = 0.040</th>
<th>k = 0.045</th>
<th>k = 0.050</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-0.702640</td>
<td>-0.737255</td>
<td>-0.758390</td>
<td>-0.763277</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.747428</td>
<td>-0.785147</td>
<td>-0.828126</td>
<td>-0.846814</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-0.750526</td>
<td>-0.785147</td>
<td>-0.828126</td>
<td>-0.846814</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.750526</td>
<td>-0.785147</td>
<td>-0.828126</td>
<td>-0.846814</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.759999</td>
<td>-0.791881</td>
<td>-0.837211</td>
<td>-0.857891</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-0.760982</td>
<td>-0.791881</td>
<td>-0.837211</td>
<td>-0.857891</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-0.760982</td>
<td>-0.791881</td>
<td>-0.837211</td>
<td>-0.857891</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-0.759999</td>
<td>-0.791881</td>
<td>-0.837211</td>
<td>-0.857891</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-0.759999</td>
<td>-0.791881</td>
<td>-0.837211</td>
<td>-0.857891</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) = Standard ridge regression.

(b) = Weighted ridge regression.
Table 7.7. n-weightings for 1st iteration. D-optimality of the variance-covariance matrix.

<table>
<thead>
<tr>
<th>Point Number</th>
<th>k = 0.005</th>
<th>k = 0.010</th>
<th>k = 0.015</th>
<th>k = 0.020</th>
<th>k = 0.025</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.145790</td>
<td>-0.263390</td>
<td>-0.357882</td>
<td>-0.454552</td>
<td>-0.500406</td>
</tr>
<tr>
<td>2</td>
<td>-0.306620</td>
<td>-0.459882</td>
<td>-0.539324</td>
<td>-0.699043</td>
<td>-0.641837</td>
</tr>
<tr>
<td>3</td>
<td>-0.150947</td>
<td>-0.271684</td>
<td>-0.367696</td>
<td>-0.445429</td>
<td>-0.590424</td>
</tr>
<tr>
<td>4</td>
<td>-0.113432</td>
<td>-0.208880</td>
<td>-0.287447</td>
<td>-0.352964</td>
<td>-0.408328</td>
</tr>
<tr>
<td>5</td>
<td>-0.208422</td>
<td>-0.355609</td>
<td>-0.455618</td>
<td>-0.524827</td>
<td>-0.574357</td>
</tr>
<tr>
<td>6</td>
<td>-0.163237</td>
<td>-0.288008</td>
<td>-0.381809</td>
<td>-0.452575</td>
<td>-0.506659</td>
</tr>
<tr>
<td>7</td>
<td>-0.266037</td>
<td>-0.452579</td>
<td>-0.578980</td>
<td>-0.665335</td>
<td>-0.725908</td>
</tr>
<tr>
<td>8</td>
<td>-0.183511</td>
<td>-0.333887</td>
<td>-0.468740</td>
<td>-0.535257</td>
<td>-0.601079</td>
</tr>
<tr>
<td>9</td>
<td>-0.113636</td>
<td>-0.204051</td>
<td>-0.277891</td>
<td>-0.340881</td>
<td>-0.396287</td>
</tr>
</tbody>
</table>

| det(V)       | 55.0644   | 50.118   | 39.8767   | 28.5215   | 29.6629   | 15.3071   | 22.5551   | 8.0493    | 17.4680   | 4.2360    |
| tr(B)        | 0.0818    | 0.1339   | 0.2637    | 0.5712    | 0.4904    | 1.2366    | 0.7364    | 2.0098    | 0.9872    | 2.8082    |
| det(H.S.E.)  | 55.1462   | 50.2519  | 29.0927   | 16.5437   | 30.1538   | 23.2915   | 10.0592   | 18.4552   | 7.0442    |           |

<table>
<thead>
<tr>
<th>Point Number</th>
<th>k = 0.030</th>
<th>k = 0.035</th>
<th>k = 0.040</th>
<th>k = 0.045</th>
<th>k = 0.050</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.570468</td>
<td>-0.621630</td>
<td>-0.660742</td>
<td>-0.698786</td>
<td>-0.732201</td>
</tr>
<tr>
<td>2</td>
<td>-0.647074</td>
<td>-0.699282</td>
<td>-0.719569</td>
<td>-0.736278</td>
<td>-0.750300</td>
</tr>
<tr>
<td>3</td>
<td>-0.604938</td>
<td>-0.615241</td>
<td>-0.661047</td>
<td>-0.695295</td>
<td>-0.724753</td>
</tr>
<tr>
<td>4</td>
<td>-0.493192</td>
<td>-0.53977</td>
<td>-0.553022</td>
<td>-0.586918</td>
<td>-0.616679</td>
</tr>
<tr>
<td>5</td>
<td>-0.61075</td>
<td>-0.639191</td>
<td>-0.616683</td>
<td>-0.632880</td>
<td>-0.646579</td>
</tr>
<tr>
<td>6</td>
<td>-0.652645</td>
<td>-0.686848</td>
<td>-0.660218</td>
<td>-0.684941</td>
<td>-0.705541</td>
</tr>
<tr>
<td>7</td>
<td>-0.769779</td>
<td>-0.802577</td>
<td>-0.790074</td>
<td>-0.806415</td>
<td>-0.825506</td>
</tr>
<tr>
<td>8</td>
<td>-0.557124</td>
<td>-0.602517</td>
<td>-0.733898</td>
<td>-0.760343</td>
<td>-0.782205</td>
</tr>
<tr>
<td>9</td>
<td>-0.469110</td>
<td>-0.515938</td>
<td>-0.549420</td>
<td>-0.589257</td>
<td>-0.625609</td>
</tr>
</tbody>
</table>

| det(V)       | 13.7408   | 2.0936   | 10.9552   | 1.0964    | 8.8373    | 0.6899    | 7.2028    | 0.3823    | 5.9249    | 0.2142    |
| tr(B)        | 1.2326    | 2.7556   | 1.4802    | 4.2543    | 1.7174    | 4.9016    | 1.9472    | 5.5693    | 2.1695    | 6.2088    |

Note: (a) = Ridge Regression quantities; (b) = Weighted Ridge Regression quantities.
inconclusive as mixed eigenvalues of the resulting difference matrix makes interpretation of bias-squared difficult. Transformation to the equivalent bias-squared of Sections 4.6 and 4.7 shows weighted ridge regression to be clearly superior to standard ridge regression; this result is holding true for both variance criteria.

If the A-optimality of the bias-matrix is investigated, the question arises as to which five points can carry infinite weighting. Selection of that point which gives the maximum reduction in bias-squared leaves the problem of the n-weighting for the remaining points unanswered. At each stage of the iterative scheme, the remaining points are weighted according to the A-optimality of variance criterion. In this way the increase in variance due to infinite n-weighting may be partially compensated by the decrease in variance due to the negative n-weighting. When the maximum of five points are given infinite n-weighting, the remaining points cannot be weighted as the fitted hyperplane is now fixed. Tables 7.8 to 7.11 illustrate these concepts. In each of the A-optimality of bias schemes, the iteration was terminated after the third cycle as being unable to improve the trace of the variance.
Table 7.8. One point infinitely weighted. A-optimality of the bias matrix, then A-Optimality of the variance-covariance matrix.

<table>
<thead>
<tr>
<th>Point Number</th>
<th>Ridge Regression</th>
<th>Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>-0.402075</td>
<td>-0.543640</td>
</tr>
<tr>
<td>2</td>
<td>-0.387483</td>
<td>-0.506198</td>
</tr>
<tr>
<td>3</td>
<td>-0.444940</td>
<td>-0.491564</td>
</tr>
<tr>
<td>4</td>
<td>-0.325724</td>
<td>-0.491166</td>
</tr>
<tr>
<td>5</td>
<td>(\infty)</td>
<td>(\infty)</td>
</tr>
<tr>
<td>6</td>
<td>-0.411964</td>
<td>-0.549213</td>
</tr>
<tr>
<td>7</td>
<td>-0.319462</td>
<td>-0.492398</td>
</tr>
<tr>
<td>8</td>
<td>-0.481980</td>
<td>-0.568463</td>
</tr>
<tr>
<td>9</td>
<td>-0.337697</td>
<td>-0.503713</td>
</tr>
<tr>
<td>tr(V)</td>
<td>25.5034</td>
<td>37.758</td>
</tr>
<tr>
<td>tr(B)</td>
<td>0.2638</td>
<td>0.09421</td>
</tr>
<tr>
<td>tr(M.S.E.)</td>
<td>25.7671</td>
<td>37.8522</td>
</tr>
<tr>
<td>E.dist.sq.</td>
<td>8.1165</td>
<td>18.6119</td>
</tr>
<tr>
<td>Res.Sq.</td>
<td>2.7528</td>
<td>3.10409</td>
</tr>
<tr>
<td>(\hat{\beta}_1)</td>
<td>-2.59918</td>
<td>-4.09789</td>
</tr>
<tr>
<td>(\hat{\beta}_2)</td>
<td>5.72552</td>
<td>5.2761</td>
</tr>
<tr>
<td>(\hat{\beta}_3)</td>
<td>3.61112</td>
<td>5.57199</td>
</tr>
<tr>
<td>(\hat{\beta}_4)</td>
<td>0.70566</td>
<td>0.71702</td>
</tr>
<tr>
<td>(\hat{\beta}_5)</td>
<td>3.35435</td>
<td>3.31833</td>
</tr>
</tbody>
</table>
Table 7.9. Two points infinitely weighted. A-optimality of the bias matrix, then A-optimality of the variance-covariance matrix.

<table>
<thead>
<tr>
<th>Point Number</th>
<th>Ridge Regression</th>
<th>Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tr(V)</td>
<td>25.5034</td>
<td>50.4521</td>
</tr>
<tr>
<td>tr(B)</td>
<td>0.2638</td>
<td>0.0124</td>
</tr>
<tr>
<td>tr(H.S.E.)</td>
<td>25.7671</td>
<td>50.4645</td>
</tr>
<tr>
<td>E.dist.sq.</td>
<td>8.1165</td>
<td>18.0960</td>
</tr>
<tr>
<td>Res.Sq.</td>
<td>2.7528</td>
<td>3.4851</td>
</tr>
<tr>
<td>Estimated β₁</td>
<td>-2.59918</td>
<td>-3.97284</td>
</tr>
<tr>
<td>β₂</td>
<td>5.72552</td>
<td>5.01805</td>
</tr>
<tr>
<td>β₃</td>
<td>3.61112</td>
<td>5.80482</td>
</tr>
<tr>
<td>β₄</td>
<td>0.70566</td>
<td>0.83804</td>
</tr>
<tr>
<td>β₅</td>
<td>3.35430</td>
<td>3.13205</td>
</tr>
</tbody>
</table>
Table 7.10. Three points infinitely weighted. A-optimality of the bias matrix, then A-optimality of the variance-covariance matrix.

<table>
<thead>
<tr>
<th>Point Number</th>
<th>Ridge Regression</th>
<th>Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>∞</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tr(V)</td>
<td>25.5034</td>
<td>62.038</td>
</tr>
<tr>
<td>tr(B)</td>
<td>0.2638</td>
<td>0.0075</td>
</tr>
<tr>
<td>tr(M:S.E.)</td>
<td>25.7671</td>
<td>62.0455</td>
</tr>
<tr>
<td>E.dist.sq.</td>
<td>8.1165</td>
<td>12.5183</td>
</tr>
<tr>
<td>Res.S.sq.</td>
<td>2.7528</td>
<td>6.94725</td>
</tr>
<tr>
<td>Estimated β₁</td>
<td>-2.59918</td>
<td>-3.85885</td>
</tr>
<tr>
<td>β₂</td>
<td>5.72552</td>
<td>4.43064</td>
</tr>
<tr>
<td>β₃</td>
<td>3.61112</td>
<td>4.54606</td>
</tr>
<tr>
<td>β₄</td>
<td>0.70566</td>
<td>1.70672</td>
</tr>
<tr>
<td>β₅</td>
<td>3.35435</td>
<td>4.02831</td>
</tr>
</tbody>
</table>
Table 7.11. Four points infinitely weighted. A-optimality of the bias matrix, then A-optimality of the variance-covariance matrix.

<table>
<thead>
<tr>
<th>Point Number</th>
<th>Ridge Regression</th>
<th>Iteration 0</th>
<th>Iteration 1</th>
<th>Iteration 2</th>
<th>Iteration 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>-0.270553</td>
<td>-0.284363</td>
<td>-0.376099</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>-0.141624</td>
<td>-0.236988</td>
<td>-0.316607</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>∞</td>
<td>∞</td>
<td>∞</td>
<td>∞</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>-0.203555</td>
<td>-0.275134</td>
<td>-0.348391</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>∞</td>
<td>∞</td>
<td>∞</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>∞</td>
<td>∞</td>
<td>∞</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>-0.234249</td>
<td>-0.284786</td>
<td>-0.361115</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>∞</td>
<td>∞</td>
<td>∞</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>-0.175140</td>
<td>-0.262986</td>
<td>-0.335639</td>
<td></td>
</tr>
<tr>
<td>tr(Υ)</td>
<td>25.5034</td>
<td>61.3360</td>
<td>25.2058</td>
<td>24.6712</td>
<td>23.7734</td>
</tr>
<tr>
<td>tr(8)</td>
<td>0.263</td>
<td>0.0073</td>
<td>0.2640</td>
<td>0.2646</td>
<td>0.2659</td>
</tr>
<tr>
<td>tr(M.S.E.)</td>
<td>25.7671</td>
<td>61.3433</td>
<td>25.4698</td>
<td>24.9358</td>
<td>24.0393</td>
</tr>
<tr>
<td>E.dist.sq.</td>
<td>8.1165</td>
<td>19.7697</td>
<td>7.9567</td>
<td>7.8743</td>
<td>7.7323</td>
</tr>
<tr>
<td>Res.S.q.</td>
<td>2.7528</td>
<td>9.2856</td>
<td>2.7558</td>
<td>2.7581</td>
<td>2.7638</td>
</tr>
<tr>
<td>Estimated</td>
<td>β₁</td>
<td>-2.59918</td>
<td>-4.82746</td>
<td>-2.61123</td>
<td>-2.61810</td>
</tr>
<tr>
<td></td>
<td>β₂</td>
<td>5.72552</td>
<td>4.81672</td>
<td>5.66950</td>
<td>5.63809</td>
</tr>
<tr>
<td></td>
<td>β₃</td>
<td>3.61112</td>
<td>4.55607</td>
<td>3.66810</td>
<td>3.70046</td>
</tr>
<tr>
<td></td>
<td>β₄</td>
<td>0.70566</td>
<td>1.09132</td>
<td>0.67335</td>
<td>0.65499</td>
</tr>
<tr>
<td></td>
<td>β₅</td>
<td>3.35435</td>
<td>2.90991</td>
<td>3.35339</td>
<td>3.35292</td>
</tr>
</tbody>
</table>
8. SUMMARY

Ordinary least squares, weighted estimation, and ridge regression are special cases of the more general class, weighted ridge regression. The investigation of some of the properties of this general class has been the subject of this dissertation.

The basic properties of weighted estimation were discussed in Chapter 3. The relationship between the artificial assigning of a variance-covariance structure to the error terms of the assumed linear model, and experimental design was investigated. In general, weighted estimation is a poor competitor to O.L.S. except for the case when a change of sign of an estimate is desired.

The application of the ridge technique to weighted estimation was covered in Chapter 4. For a given k-value, positive n-weightings caused a decrease in bias-squared, but an increase in variance. Negative n-weightings have the effect of decreasing the variance, but often increase the bias-squared. It is shown, and demonstrated by the example of Chapter 7, that there exist negative n-weightings such that both variance and bias-squared are reduced. Comparison of bias-matrices for two estimates is seldom straight-forward when the true parameter values are not known. A technique for comparing the bias-matrices of two estimates, when the true parameter values are unknown, was developed and named equivalent bias-squared.

Chapter 5 dealt with the actual determination of the n-weightings used in weighted ridge estimation. The criteria considered represent a considerable cross-section of those criteria commonly used. With each criterion, a sequential method of determining the n-weighting was used. This method examined each point in the data set in a sequential
order, and the point yielding the maximum improvement (with respect to the criterion being used) over the preceding stage, chosen to be weighted.

The extension of the concept of \( n \)-weighting to more complex situations was investigated in Chapter 6, and the basic properties of each of the three situations discussed were examined.

An example of weighted ridge regression was given in Chapter 7, and was chosen after some deliberation. Examples and counter-examples, each showing the relative merits and drawbacks with biased estimation, are easy to construct. The example represented a "typical" weighted ridge regression problem with parameter values frequently met with by data analysts. When information concerning parameter values is available to the data analyst, biased estimation in general, and weighted ridge regression in particular, have estimators superior (for a specified criterion) to O.L.S.

Further work in weighted ridge regression might be directed toward the following three areas:

(i) Development of an efficient algorithm for the selection of the \( n \)-weightings. The sequential technique has the same kind of drawbacks as does the technique of forward selection in building a model in multiple regression. Possibly suitable algorithms may exist in the areas of non-linear programming or simplex-type search.

(ii) Investigation of how the assumption of a model that is different from the true model is affected by both the ridge-technique and the weighting of points.
(iii) Finally, a full extension from ridge regression technique of adding an equal positive quantity to the main diagonal of $X'X$ prior to inversion to the general technique of unequal positive quantities should be investigated.
REFERENCES


