AN ABSTRACT OF THE THESIS OF

MARY ELLEN SASS for the degree of MASTER OF SCIENCE

in Agricultural and Resource Economics presented on July 17, 1978

Title: A MONTE CARLO ANALYSIS OF PRINCIPAL COMPONENTS AND RIDGE REGRESSION WITH APPLICATION TO A FISHERIES EVALUATION MODEL

Abstract approved: Redacted for Privacy
William G. Brown

Essential to the testing of propositions in economic theory is the estimation of the parameters involved in relationships among economic variables. Probably the most widely used method of estimation is ordinary least squares (OLS). However, severe multicollinearity or near linear dependence of the explanatory variables can cause the OLS estimates to be unrealistic and imprecise due to large sampling variances. One common solution to the multicollinearity problem is to drop highly intercorrelated variables from the regression model. Unfortunately, variable deletion may result in serious specification bias and thus may be a poor method for alleviating multicollinearity.

This paper investigates the use of two methods of biased linear estimation, principal components analysis

and ridge regression, as alternatives to OLS estimation of the full model in the presence of multicollinearity. A biased linear estimator may be an attractive alternative if its mean square error (MSE) compares favorably with OLS. In this paper, three ridge estimators and two types of principal components estimators are compared to OLS in a series of Monte Carlo experiments and in an application to an empirical problem. The three ridge estimators are: (1) an estimator proposed by Lawless and Wang; (2) a fixed k-value estimator (k = 0.1); (3) RIDGM, an estimator proposed by Dempster, Wermuth, et al. The two types of principal components estimators are: (1) the traditional t-criteria for deleting principal components; (2) a proposed loss-function-related criterion for deleting principal components.

In the Monte Carlo experiments, OLS and the biased estimators are applied to four data sets, each characterized by a different level of multicollinearity and various information-to-noise ratios. The Monte Carlo results indicate that all the biased estimators can be more effective than OLS (considering MSE) in estimating the parameters of the full model under conditions of high multicollinearity at low and moderate information-to-noise ratios. (The RIDGM estimator, however, produced lower MSE than OLS at all information-to-noise ratios in the data sets where multicollinearity was present.) For principal

components analysis, the proposed loss-function-related criterion produced generally lower MSE than the traditional t-criteria. For ridge regression, the Lawless-Wang estimator, which is shown to minimize estimated MSE, produces generally lower MSE than the other ridge estimators in the Monte Carlo experiments. Also, the Lawless-Wang estimator was somewhat more effective overall than the proposed loss-function-related criterion for deleting components.

Another comparison of the estimators is made in their application to an empirical problem, a recreation demand model specified by the travel cost method. The comparison of the estimators is based on estimated MSE and on prior information about the coefficients. In this particular case, the Lawless-Wang estimator appears to produce the best improvement over OLS. However, this empirical problem is merely an example of the application of the biased estimators rather than a crucial test of their effectiveness.

The inability to judge the reliability of biased estimates, due to the unknown bias squared component of MSE, has been a serious limitation in the application of biased linear estimation to empirical problems. Brown, however, has proposed a method for estimating the MSE of ridge coefficients. His method is applied in the empirical example and in the Monte Carlo experiments to the ridge estimators, and in principle, to the proposed loss-

function-related criterion for deleting principal components. For the Lawless-Wang ridge estimator and the proposed loss-function-related criterion, the suggested method for estimating MSE appears to produce good estimates of MSE under conditions of high multicollinearity at low and moderate information-to-noise ratios. In fact, at all but the highest information-to-noise ratio, the Monte Carlo results indicate that this estimate of MSE can be much more accurate than estimates of OLS variances.

A MONTE CARLO ANALYSIS OF PRINCIPAL COMPONENTS AND RIDGE REGRESSION WITH APPLICATION TO A FISHERIES EVALUATION MODEL

by

Mary Ellen Sass

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

June 1979

APPROVED:

Redacted for Privacy

Professor of Agricultural and Resource Economics
in charge of major

Redacted for Privacy

Head of Department of Agricultural and Resource Economics

Redacted for Privacy

Dean of Graduate School

Date thesis is presented July 17, 1978

Typed by Deanna L. Cramer for Mary Ellen Sass

ACKNOWLEDGEMENT

I wish to express my most sincere gratitude and appreciation to Dr. William Brown, my major professor, for his contribution to my graduate studies and to this thesis. I am very grateful to him for suggesting this research topic and for granting my request to work under his direction. Without his patient guidance and generous contribution of time, this thesis could not have been completed.

My special thanks to Dr. Richard Towey for his tolerant and generous support of my computer habit. Without his help, I would not have developed the programming skills which greatly expedited much of the analysis in this thesis.

Also, I wish to thank Dr. Richard Johnston, Dr. Charles Warren, and Dr. Donald Pierce for serving on my graduate committee.

TABLE OF CONTENTS

Table of Contents--continued

LIST OF TABLES

List of Tables -- continued

List of Tables -- continued

A MONTE CARLO ANALYSIS OF PRINCIPAL COMPONENTS AND RIDGE REGRESSION WITH APPLICATION TO A FISHERIES EVALUATION MODEL

I. INTRODUCTION

Perhaps the greatest handicap under which an economist must work is the lack of opportunity to experiment. Economists obtain data by observing processes which cannot be controlled or repeated at will. As a result, when the sample data have undesirable characteristics, the economist is unable to choose an experimental design to remedy the situation.

One undesirable characteristic commonly found in economic data is the intercorrelation of explanatory variables. Intercorrelation among economic variables is common since many variables tend to move together due to trends and cycles inherent in economic processes. Moreover, much of economic data is aggregated or averaged which further increases the intercorrelation among variables (Brown and Nawas, 1973). Consequently, at least some intercorrelation among economic variables is expected due to the nature of the data.

Multicollinearity

Multicollinearity arises from the existence of intercorrelation among the explanatory variables in the

regression model. It exists in varying degrees, from extremes which are well defined and easily recognized to intermediate stages which are more subtle, more complex and, unfortunately, more frequent. It is possible in economic data, though highly unlikely, for multicollinearity to be totally absent. Much more frequent, however, is the other extreme, perfect multicollinearity.

Multicollinearity is said to be "perfect" when one of the assumptions of the classical linear regression model is violated. The assumption is that each of the explanatory variables in the model is linearly independent of the other variables and of any linear combination of the other variables. In other words, it is assumed that the X'X matrix has full rank so that its determinant and inverse exist. If multicollinearity is perfect, this assumption does not hold, and the ordinary least squares (OLS) regression coefficients cannot be estimated since $\hat{\beta} = (X'X)^{-1}X'Y$ requires the inversion of X'X. Thus, it is impossible to overlook the presence of perfect multicollinearity.

Often, perfect multicollinearity is the result of including too many dummy variables in the model. For example, if the model includes an intercept term or one is automatically generated in a regression program, then n-1 dummy variables should be used to represent n classifications. Otherwise, the columns of X'X will be linearly dependent and X'X will be singular. Unfortunately, the cause of perfect

multicollinearity is not always so obvious or easily corrected. In fact, the pattern of intercorrelation may be such that perfect multicollinearity exists despite low pairwise correlations of the variables. (An example is presented later in this chapter.)

Suppose that multicollinearity is not perfect but is severe. Although the existence of $(X'X)^{-1}$ guarantees that the OLS coefficients can be calculated, it does not guarantee their accuracy. For example, suppose that the explanatory variables are standardized, making X'X a correlation matrix. When there is no multicollinearity, X'X is an identity matrix and its determinant equals one. As multicollinearity increases in severity, the determinant approaches zero and the inverse increases in size since

$$(X'X)^{-1} = \frac{1}{\det(X'X)} \text{ adj } (X'X) \ .$$

A small determinant resulting from severe multicollinearity may cause serious roundoff errors in the regression calculations. As a result, the regression coefficients may not be computationally accurate.

More importantly, though, the regression coefficients tend to be imprecise because of large sampling variances. Consider that the variance-covariance of the OLS estimator is defined as

$$\text{var } (\hat{\beta}) = \sigma^2 \ (X'X)^{-1}$$

where σ^2 is the error term variance. In the two-explanatory-variable case,

$$\text{var } (\hat{\beta}) = \frac{\sigma^2}{1-r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}$$

where r_{12} is the pairwise correlation between the explanatory variables. As r_{12} increases, the variance-covariance of $\hat{\beta}$ increases. If the variables are perfectly correlated, r_{12} = 1.0 and var $(\hat{\beta})$ becomes infinite.

In addition to being imprecise, both the estimated coefficients and their variances are very sensitive to changes in the data and in the specification of the model. In the presence of multicollinearity, even small changes in X'X and its determinant can cause large changes in its inverse.

Another related and perhaps more serious consequence of severe multicollinearity is that the separate effects of the explanatory variables cannot be isolated. The regression coefficients of any explanatory variable must include the effects of the other correlated variables in the model. The entangled effects of the variables plus the wide variety and range of possible estimates make interpretation of the regression coefficients difficult if not impossible.

Detection of Multicollinearity

Merely detecting the presence of multicolinearity is not so much the problem since some intercorrelation of

variables almost always exists in economic data. Of greater concern, is the detection of the pattern and the degree of multicollinearity.

When multicollinearity is severe, the variances of the regression coefficients may be so greatly increased that no coefficient is significant even though the overall regression may be highly significant. Thus, the paradoxical situation of low t-statistics and a high F-statistic for the overall regression indicates that the multicollinearity is severe.

Of course, as previously mentioned, a low value of the determinant of the X'X correlation matrix also indicates the presence of severe multicollinearity. However, determinants do not provide an index for measuring the degree of multicollinearity.

Measurement of the degree of multicollinearity is quite straightforward for a model with two explanatory variables. In that case, r_{12}, the zero order correlation between the explanatory variables indicates the severity of the intercorrelation. (When there is no multicollinearity, $r_{12} = 0$. When multicollinearity is perfect, $r_{12} = \pm 1.0$.) However, in models with more than two explanatory variables, the zero order correlations are not a reliable index. In fact, perfect multicollinearity can occur even though the zero order correlations are low. To illustrate this, consider the five-explanatory-variable model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + e$$

where the X matrix is

$$X = \begin{bmatrix} 0 & -1 & -1 & 0 & 0 \\ -1 & 0 & 0 & -1 & 0 \\ 0 & -1 & 1 & 0 & -2 \\ -1 & 0 & 0 & 1 & -2 \\ 0 & 1 & -1 & 0 & 2 \\ 1 & 0 & 0 & -1 & 2 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

The corresponding correlation matrix of the explanatory variables is

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 & 0.5 \\ 0 & 1 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0 & -0.5 \\ 0 & 0 & 0 & 1 & -0.5 \\ 0.5 & 0.5 & -0.5 & -0.5 & 1 \end{bmatrix}$$

The X'X matrix is

$$X'X = \begin{bmatrix} 4 & 0 & 0 & 0 & 4 \\ 0 & 4 & 0 & 0 & 4 \\ 0 & 0 & 4 & 0 & -4 \\ 0 & 0 & 0 & 4 & -4 \\ 4 & 4 & -4 & -4 & 16 \end{bmatrix}$$

The column vectors of X'X are linearly dependent since

$$\text{Col (1)} + \text{Col (2)} - \text{Col (3)} - \text{Col (4)} - \text{Col (5)} = 0.$$

Thus, det (X'X) = 0 and X'X cannot be inverted. Perfect multicollinearity exists in this case even though the highest zero order correlation is ± 0.5.

The most thorough approach to measuring multicollinearity and to determining its pattern would be to calculate all the possible regressions among the explanatory variables. Although thorough, this can be costly and inconvenient for large models.

A much more convenient method was suggested by Farrar and Glauber (1967). Their method involves only the calculation of the correlation matrix X'X and its inverse. The diagonal elements of the inverse are then inspected for indications of the location and the degree of correlation among the explanatory variables.

A main diagonal element of $(X'X)^{-1}$ corresponding to the ith explanatory variable is defined as

$$r^{ii} = \frac{\det (X'X)_{ii}}{\det (X'X)}$$

where $(X'X)_{ii}$ is a matrix obtained by deleting the row and the column of X'X in which the ith variable appears. When there is no multicollinearity, X'X and $(X'X)_{ii}$ are identity matrices and $\det(X'X) = \det(X'X)_{ii} = 1.0$, so that $r^{ii} = 1.0$. When an explanatory variable X_i, is linearly dependent on the other variables in the set, $\det(X'X) = 0$ and r^{ii} becomes infinite. The greater the correlation of X_i with the other variables, the closer is $\det(X'X)$ to zero and the larger is r^{ii}.

The magnitude of r^{ii} not only indicates the degree of multicollinearity but also indicates how much the variances of the OLS estimates will be inflated. Recall that the variance-covariance matrix of $\hat{\beta}$ is

$$\text{var } (\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

where the variance of $\hat{\beta}_i$ is the ith diagonal element of $(X'X)^{-1}$ times σ^2, that is,

$$v(\hat{\beta}_i) = \sigma^2 r^{ii} \ .$$

Consequently, the main diagonal element, r^{ii}, can be considered as a variance inflation factor. For example, if $r^{22} = 25$, then $v(\hat{\beta}_2)$ would be 25 times larger than if X_2 were orthogonal to the other explanatory variables.

Corrective Procedures for Multicollinearity

There are a number of procedures available which may reduce multicollinearity. Just which procedure, if any, will be effective depends upon the characteristics of the particular data set. Unfortunately, there is no complete "remedy" nor one particular procedure which is appropriate in all situations. There are, however, several widely used approaches to the problem.

Disaggregation of the Data

As pointed out by Brown and Nawas (1973), when the data are in aggregated form, returning to the individual observations will reduce the degree of multicollinearity. Of course, the individual observations may not always be available. This is usually the case when the data are subject to confidentiality restrictions or are secondary data (e.g. time series, census data).

Additional Observations

Increasing the sample size often reduces the degree of multicollinearity. However, there is no guarantee that some pattern of intercorrelation will not also exist in the additional observations. More importantly, however, it is often impossible to obtain additional observations when using published data. Even though this suggestion is often not practicable, the possibility should not be overlooked.

A Priori Information

Additional information about the coefficients of the explanatory variables can be used to break the pattern of multicollinearity. Such information can be obtained from similar studies on different data, and other sources extraneous to the sample data.

For example, assume the original regression model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e \qquad (1.1)$$

where X_2 and X_3 are highly correlated. If we have some prior estimate of β_3, perhaps $\hat{\beta}_3$, then we can estimate the model

$$Y - \hat{\beta}_3 X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e \ .$$

Another possibility is that we may have a priori information about a relationship between the regression coefficients, perhaps that two parameters are of equal magnitude or their sum is a fixed value. As an example, suppose that $\beta_2/\beta_3 = K$, then $\beta_2 = K\ \beta_3$. Substituting for β_2 in (1.1), we have

$$Y = \beta_0 + \beta_1 X_1 + K\ \beta_3 X_2 + \beta_3 X_3 + e$$

$$= \beta_0 + \beta_1 X_1 + \beta_3 (K X_2 + X_3) + e \ .$$

The coefficients β_1 and β_3 can then be estimated under reduced multicollinearity and β_2 can be obtained from its relationship with β_3.

Deletion of Variables

Deletion of variables is often an unavoidable necessity when multicollinearity produces nonsensical regression results and no other solution to the problem is available. It should be kept in mind, however, that deleting a relevant

variable may be undesirable because of the resulting specification bias (Johnston, p. 169).

When a relevant but intercorrelated variable is excluded from the model, much of its effect on the dependent variable is attributed to the remaining variables. The resulting bias of the coefficients of the remaining variables can be derived for the general case. Consider the full model,

$$Y_{nl} = X_{nk}\,\beta_{kl} + Z_{nr}\,\Gamma_{rl} + \varepsilon_{nl}$$

where X and Z are matrices of explanatory variables and β and Γ are vectors of their respective coefficients. Suppose that the Z variables are omitted and that the reduced model $Y = X\beta + \varepsilon$ is fitted. The OLS estimator in this case is

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

Substituting for Y from the full model, we have

$$\hat{\beta} = (X'X)^{-1}\,X'(X\beta + Z\Gamma + \varepsilon).$$

the expected value of $\hat{\beta}$ is

$$\begin{aligned} E(\hat{\beta}) &= E[(X'X)^{-1}\,X'(X\beta + Z\Gamma + \varepsilon)] \\ &= E[I\beta + (X'X)^{-1}\,X'Z\Gamma + (X'X)^{-1}\,X'\varepsilon] \\ &= \beta + (X'X)^{-1}\,X'Z\Gamma \end{aligned}$$

assuming that X and Z are fixed and that $E(\varepsilon) = 0$. The bias of $\hat{\beta}$ is, of course,

$$E(\hat{\beta}) - \beta = (X'X)^{-1} X'Z\Gamma \quad .$$

Notice that the term $(X'X)^{-1} X'Z$ defines an OLS estimator, $\hat{\alpha}$, in a regression of Z on X:

$$\hat{\alpha} = (X'X)^{-1} X'Z \quad .$$

The bias of $\hat{\beta}$ becomes

$$E(\hat{\beta}) - \beta = \hat{\alpha}\Gamma \quad . \tag{1.2}$$

Obviously, $\hat{\beta}$ will be unbiased if $\hat{\alpha}$ is a null vector, in which case, X and Z are uncorrelated. $\hat{\beta}$ will also be unbiased if $\Gamma = 0$, that is, if the omitted variables have no effect on Y. Of course, the direction and magnitude of the bias of an individual estimate, $\hat{\beta}_i$, depends upon the magnitudes and the signs of $\hat{\alpha}_i$ and γ_i. The rationalization for deleting variables is that the resulting specification bias will be more than offset by the improvement in the estimation of the coefficients of the remaining variables.

II. BIASED LINEAR ESTIMATION

Even in the presence of multicollinearity, the OLS estimator is unbiased and has minimum variance among linear unbiased estimators (if the basic assumptions of the classical linear model are satisfied). However, multicollinearity may cause the OLS variance to be very large. This raises the possibility of using a biased estimator with smaller variance than OLS. Of course, not only the variance but also the mean square error (the variance plus the bias squared) of the biased estimator must compare favorably with OLS.

Ridge Regression

Ridge regression is a method of biased linear estimation which has been shown to be effective in coping with multicollinearity. Essentially, ridge regression reduces multicollinearity by adding small positive amounts, k, to the main diagonal elements of X'X, the correlation matrix. Consider the model

$$Y = X\beta + \varepsilon \qquad (2\text{-}1)$$

where

Y_{n1} is the dependent variable vector;

X_{np} is a matrix of the standardized independent variables which has full rank;

β_{p1} is a vector of the true parameters;

ε_{n1} is a vector of the random error terms such that $E(\varepsilon) = 0$ and $E(\varepsilon\varepsilon') = \sigma^2 I_n$.

The ridge estimator is defined as

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

The variance-covariance of $\hat{\beta}^*$ is

$$\text{var } (\hat{\beta}^*) = \sigma^2 (X'X + kI)^{-1} X'X(X'X + kI)^{-1} \ .$$

When $k = 0$, the ridge estimator and its variance-covariance are equal to the OLS estimator and its variance-covariance. For some values of k, however, the variance of $\hat{\beta}^*$ will be smaller than the OLS variance. The values of k for which this is true can be determined by considering an orthogonal transformation of the regression model.

If Q is an orthogonal matrix such that $Q'Q = QQ' = I$ and $Q'X'XQ = \Lambda$ where Λ is a diagonal matrix of the eigenvalues of X'X, then Q will always exist if X'X is a positive definite matrix (Hadley, pp. 247-249). Since $QQ' = I$, the regression model in 2.1) is equivalently,

$$Y = XQQ'\beta + \varepsilon$$

which becomes

$$Y = Z\alpha + \varepsilon$$

where $Z_{np} = X_{np}Q_{pp}$ and $\alpha_{p1} = Q'_{pp}\beta_{p1}$.

For this transformed model, the OLS estimator is

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

where $Z'Z = \Lambda$. The inverse of $Z'Z$ is a diagonal matrix of the reciprocals of the eigenvalues of $X'X$. Thus, the OLS variance-covariance for the transformed model is

$$\text{var } (\hat{\alpha}) = \sigma^2 (Z'Z)^{-1} . \qquad (2.2)$$

The variance of the OLS estimator is the trace of the variance-covariance matrix:

$$V(\hat{\alpha}) = \sigma^2 \text{ trace } (Z'Z)^{-1}$$

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

The ridge estimator for the transformed model is

$$\hat{\alpha}^* = (Z'Z + kI)^{-1} Z'Y$$

and the variance-covariance of $\hat{\alpha}^*$ is

$$\text{var } (\hat{\alpha}^*) = \sigma^2 (Z'Z + kI)^{-1} Z'Z(Z'Z + kI)^{-1} .$$

the variance of $\hat{\alpha}^*$ is

$$V(\hat{\alpha}^*) = \sigma^2 \text{ trace } [(Z'Z + kI)^{-1} Z'Z(Z'Z + kI)^{-1}]$$

$$= \sigma^2 \text{ trace } [Z'Z(Z'Z + kI)^{-1} (Z'Z + kI)^{-1}]$$

(since tr (ABC) = tr (BAC))

$$= \sigma^2 \text{ trace } [Z'Z (Z'Z + kI)^{-2}]$$

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

The variance of the ridge estimator can be shown to be smaller than the variance of the OLS estimator for positive values of k. Consider the following comparison of the variances:

$$v(\hat{\alpha}) - v(\hat{\alpha}^*) \stackrel{?}{>} 0$$

$$\sigma^2 \sum_{i=1}^{p} 1/\lambda_i - \sigma^2 \sum_{i=1}^{p} \lambda_i/(\lambda_i+k)^2 \stackrel{?}{>} 0$$

$$\sigma^2 \sum_{i=1}^{p} [(\lambda_i+k)^2 - \lambda_i^2] \stackrel{?}{>} 0$$

$$\sigma^2 \sum_{i=1}^{p} (\lambda_i^2 + 2\lambda_i k + k^2 - \lambda_i^2) \stackrel{?}{>} 0$$

$$\sigma^2 \sum_{i=1}^{p} (2\lambda_i k + k^2) \stackrel{?}{>} 0$$

Assuming that X'X is positive definite (i.e., $\lambda_i > 0$ for all i), and that $k > 0$, then the ridge estimator will have a variance smaller than the OLS estimator. It can also be shown that as k increases, the variance of the ridge estimator decreases (Schmidt, p. 52):

$$\frac{dv(\hat{\alpha}^*)}{dk} = \frac{d}{dk} \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2} = -2 \sigma^2 \sum_{i=1}^{p} \lambda_i/(\lambda_i+k)^3$$

$\frac{dV(\hat{\alpha}^*)}{dk}$ is negative since σ^2, λ_i and $k > 0$.

Thus, the variance of the ridge estimator can be made as small as we like by increasing the value of k. However, the accuracy of the ridge estimator depends not only on the magnitude of its variance but also on the magnitude of its

bias. The bias of the ridge estimator can be derived as follows:

$$E(\hat{\alpha}^*) = E[(Z'Z + k\ I)^{-1}\ Z'Y]$$

$$= E[(Z'Z + k\ I)^{-1}\ Z'(Z\alpha + \varepsilon)]$$

$$= E[(Z'Z + k\ I)^{-1}\ (Z'Z\alpha + Z'\varepsilon)]$$

$$= E[(Z'Z + k\ I)^{-1}\ (Z'Z\alpha + k\ I\alpha - kI\alpha + Z'\varepsilon)]$$

$$= E\ \{[Z'Z + k\ I]^{-1}\ [(Z'Z + k\ I)\alpha - k\ I\alpha + Z'\varepsilon]\}$$

$$= E[I\alpha - (Z'Z + k\ I)^{-1}\ k\ I\alpha + (Z'Z + k\ I)^{-1}\ Z'\varepsilon]$$

$$= \alpha - (Z'Z + k\ I)^{-1}\ k\alpha\ .$$

Thus, the bias is

$$E(\hat{\alpha}^*) - \alpha = -k\ (Z'Z + k\ I)^{-1}\ \alpha.$$

In order to give both the positive and negative values of $[E(\hat{\alpha}^*) - \alpha]$ equal weight, let us consider the square of the bias,

$$[E(\hat{\alpha}^*)-\alpha]'[E(\hat{\alpha}^*)-\alpha] = [-k(Z'Z+k\ I)^{-1}\alpha]'[-k(Z'Z+k\ I)^{-1}\alpha]$$

$$= [\alpha'(Z'Z+k\ I)^{-1}-k][-k(Z'Z+k\ I)^{-1}\alpha]$$

$$= k^2\ \alpha'(Z'Z+k\ I)^{-2}\alpha$$

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

It can be shown that the square of the bias increases with

larger values of k (Schmidt, p. 52):

$$\frac{d\ \text{Bias}^2\ (\hat{\alpha}^*)}{dk} = \frac{d}{dk}\ k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i+k)^2}$$

$$= 2k \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i+k)^2} - 2k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i+k)^3}$$

$$= 2k \sum_{i=1}^{p} \frac{\alpha_i^2\lambda_i}{(\lambda_i+k)^3} \geq 0$$

Although we can decrease the variance of $\hat{\alpha}^*$ by increasing k, we will cause the other component of mean square error, the bias squared of $\hat{\alpha}^*$, to increase.

Of course, the magnitude of the bias depends not only on the magnitude of k but also on the magnitudes of the α_i and λ_i. (The values of the λ_i will range from slightly greater than zero to p, the number of standardized explanatory variables, if X'X is a positive definite matrix. When $\lambda_i = 0$, then X'X is singular. If all the $\lambda_i = 1$, then all the X_i are orthogonal. The closer to zero are some λ_i, the greater is the intercorrelation of the X_i.) The smaller the value of λ_i, the larger is the bias of $\hat{\alpha}_i^*$ for given values of α_i and k, since

$$\text{Bias}\ (\hat{\alpha}_i^*) = \frac{-k\alpha_i}{\lambda_i+k} \quad .$$

Naturally, the bias will be lessened if the α_i corresponding to the low λ_i is not too large. (Recall that the α_i are linear combinations of the true β parameters.

$$\alpha_i = q_{1i}\beta_1 + q_{2i}\beta_2 + \ldots + q_{pi}\beta_p).$$

For empirical applications, it is more convenient to consider how the magnitude of the bias is affected by the values of the true β parameters. A priori information about the true parameters provides some indication whether conditions will be favorable for ridge regression. Consider the simple case of a two-explanatory-variable model,

$$Y = \beta_1 X_1 + \beta_2 X_2 + e.$$

If X_1 and X_2 are standardized so that X'X is a correlation matrix, then the bias of the ridge estimator, $\hat{\beta}^*$ is

$$E(\hat{\beta}^*) - \beta = -k(x'x + k\ I)^{-1}\beta$$

$$= \frac{-k}{(1+k)^2 - r_{12}^2} \begin{bmatrix} 1+k & -r_{12} \\ -r_{12} & 1+k \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

Individually, the bias of $\hat{\beta}_1^*$ and the bias of $\hat{\beta}_2^*$ are

$$E(\hat{\beta}_1^* - \beta_1) = \frac{-k}{(1+k)^2 - r_{12}^2} [\beta_1(1+k) - r_{12}\beta_2]$$

$$E(\hat{\beta}_2^* - \beta_2) = \frac{-k}{(1+k)^2 - r_{12}^2} [-r_{12}\beta_1 + \beta_2(1+k)] .$$

The bias will be lessened if: (1) The true parameters are of about equal magnitude and have the same sign and the explanatory variables are positively correlated; (2) The true parameters are of about equal magnitude and opposite signs but the explanatory variables are negatively correlated. Ridge regression will be more effective for models with these characteristics (Brown, 1973).

The interesting question is, of course, how does the accuracy of ridge regression compare with OLS? Hoerl and Kennard (1970) showed that there will always be a ridge estimate more accurate than OLS when accuracy is measured by the mean square error. The difficulty is to determine the value of k which will provide the more accurate ridge estimate.

The optimal value of k can be found by setting the derivative of the mean square error (with respect to k) equal to zero and solving for k:

$$\frac{d\ MSE\ (\hat{\beta}^*)}{dk} = \frac{d}{dk}\ [var\ (\hat{\beta}^*) + Bias^2\ (\hat{\beta}^*)]$$

$$= \frac{d}{dk}\ [\sigma^2 (x'x + k\ I)^{-1}\ x'x(x'x + k\ I)^{-1} + k^2\beta^2$$

$$(x'x + k\ I)^{-2}\beta]\ .$$

However, in most empirical analysis, σ^2 and the values of β are unknown. As a result, the value of k must usually be chosen based on estimates of the true values, or some other criterion for selecting k.

Ridge estimators are differentiated by the method used to compute k. Three ridge estimators were used in the Monte Carlo experiments described in this paper: a fixed k-value estimator; an estimator developed by Lawless and Wang (1976); and an estimator proposed by Dempster, Schatzoff and Wermuth (1977).

Unfortunately, ridge estimators have sampling distributions which are complex and not well understood. Consequently, it has not been possible to test hypotheses. This has been the most serious limitation of applying ridge regression in economic analysis.

Principal Components Analysis

Principal components analysis is a method of biased linear estimation which can be effective in mitigating the effects of multicollinearity. Basically, the procedure is to reduce the information demands on the sample data by deleting some principal components. In this respect, principal components analysis is similar to variable deletion; however, only components of variables, rather than whole variables, are deleted.

Principal components analysis begins by transforming the explanatory variables into another set of variables which are pairwise uncorrelated. These variables are the principal components. The transformation used is the

orthogonal transformation presented earlier where the model, $Y = X\beta + \varepsilon$ is transformed to the orthogonal form,

$$Y = Z\alpha + \varepsilon$$

where $Z_{np} = X_{np}Q_{pp}$ and $\alpha_{pl} = Q'_{pp}\beta_{pl}$. Recall that X'X is the correlation matrix, Q is an orthogonal matrix of the eigenvectors of X'X, and Z'Z is a diagonal matrix of the eigenvalues of X'X. The principal components of the explanatory variables are the Z_i which are linear combinations of the X variables, e.g.,

$$\overleftarrow{Z}_1 = X\overleftarrow{q}_1$$

where the ith element of the Z_1 vector is

$$Z_{i1} = q_{11}X_{i1} + q_{21}X_{i2} + \cdots + q_{p1}X_{ip}$$

(Johnston, pp. 322-331).

An OLS regression of the dependent variable on the principal components yields estimated coefficients, $\hat{\alpha}_i$, which can be transformed to the OLS coefficients of the X variables since $\hat{\alpha} = Q'\hat{\beta}$ which implies that $\hat{\beta} = Q\hat{\alpha}$. Thus, when all the principal components are retained, the principal components results are equivalent to OLS. Principal components analysis, then, is essentially a process of selecting a subset of principal components to be used in the regression model, $Y = Z\alpha + \varepsilon$.

Reducing the number of principal components can result in more stable estimates than OLS. This can be seen by considering that the total variation in the explanatory variables is $\sum_{i=1}^{p} X_i^2$. This is the trace of the correlation matrix, X'X, which equals p, the number of standardized explanatory variables. In terms of the principal components, the total variation in the X variables is

$$\begin{aligned} tr\ (X'X) &= tr(X'XI) \\ &= tr(X'XQQ') \\ &= tr(Q'X'XQ) \\ &= tr(Z'Z) \\ &= \sum_{i=1}^{p} Z_i^2 \\ &= \sum_{i=1}^{p} \lambda_i \quad . \end{aligned}$$

Assume that multicollinearity is not perfect, so that X'X is positive definite and all $\lambda_i > 0$. If multicollinearity is severe, some of the λ_i will be close to zero. In that case, a smaller number of the λ_i will account for most of the variation in the explanatory variables. Equivalently, a corresponding subset of the Z_i will also account for most of the variation in the explanatory variables.

The deletion of principal components can be accomplished by replacing the corresponding eigenvectors of matrix Q with null vectors. This new matrix, Q*, can be

used in a transformation of the $\hat{\alpha}_i$ coefficients into estimated coefficients, β_i^*, of the X variables. The β_i^* are the principal components estimates (McCallum, 1970). Thus, the principal components estimator can be defined as

$$\beta^* = Q^*\hat{\alpha}.$$

The variance-covariance of β^* can be easily determined if we recall from (2.2) that

$$\begin{aligned} \text{var}(\hat{\alpha}) &= \sigma^2(Z'Z)^{-1} \\ &= \sigma^2 \Lambda^{-1} \end{aligned}$$

where Λ^{-1} is a diagonal matrix of the reciprocals of the eigenvalues of X'X. When some components are deleted, $Z=XQ^*$ and $\beta^* = Q^*\hat{\alpha}$. Consequently, the variance-covariance of β^* is

$$\begin{aligned} \text{var}(\beta^*) &= \sigma^2[(XQ^*)'(XQ^*)] \\ &= \sigma^2(Q^{*\prime}X'XQ^*) \\ &= \sigma^2 \Lambda^{*-1} \end{aligned}$$

where Λ^{*-1} is a diagonal matrix with some diagonal elements equal to zero but the rest identical to the elements of Λ^{-1}.

The variance of β^* is

$$v(\beta^*) = \sigma^2 \text{ tr}(\Lambda^{*-1}) \ .$$

A comparison of the variance of β^* can be made with the

variance of the OLS estimator, $\hat{\beta}$. Recall that

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

This can be expressed in terms of Λ^{-1} since $Z = XQ$, which implies $X = ZQ'$:

$$\begin{aligned}
\text{var}(\hat{\beta}) &= \sigma^2[(ZQ')'(ZQ')]^{-1} \\
&= \sigma^2\ (QZ'ZQ')^{-1} \\
&= \sigma^2\ Q(Z'Z)^{-1}Q' \text{ since} \\
&\quad (ABC)^{-1} = C^{-1}B^{-1}A^{-1} \\
&= \sigma^2\ Q\Lambda^{-1}Q'.
\end{aligned}$$

The variance of $\hat{\beta}$ is

$$\begin{aligned}
v(\hat{\beta}) &= \sigma^2\ \text{tr}\ (Q\Lambda^{-1}Q') \\
&= \sigma^2\ \text{tr}\ (Q'Q\Lambda^{-1}) \text{ since} \\
&\quad \text{tr}\ (ABC) = \text{tr}(CAB) \\
&= \sigma^2\ \text{tr}(\Lambda^{-1}) \\
&= v(\hat{\alpha}) .
\end{aligned}$$

Clearly, the variance of $\hat{\beta}$ is greater than the variance of β^*, that is,

$$\sigma^2\ \text{tr}(\Lambda^{-1}) > \sigma^2\ \text{tr}(\Lambda^{*-1}) ,$$

since one or more of the diagonal elements of Λ^{-1} associated with the smallest λ_i values have been set to zero in Λ^{*-1}.

Although the principal components estimator may be more stable than the OLS estimator, lower variance does not

necessarily imply that β^* will be more accurate than $\hat{\beta}$. It must not be overlooked that β^* is a biased estimator and that this will be reflected in the magnitude of the mean square error of β^* through the biased squared component. As shown by McCallum (1970), the bias of β^* can be derived as

$$\begin{aligned} E(\beta^*) - \beta &= E(Q^*\hat{\alpha}) - Q\alpha \\ &= Q^* E(\hat{\alpha}) - Q\alpha \\ &= Q^*\alpha - Q\alpha \\ &= (Q^*-Q)\alpha \ . \end{aligned} \tag{2.3}$$

Of course, the magnitude of the bias of β^* is affected by the magnitudes of the true β parameters. As an illustration, consider the two-explanatory-variable model,

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where the matrix of eigenvectors is

$$Q = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix}$$

Suppose that Z_2, one of the two principal components is deleted, then the eigenvector matrix is

$$Q^* = \begin{bmatrix} q_{11} & 0 \\ q_{21} & 0 \end{bmatrix} \ .$$

The difference term in the bias in (2.3) is

$$Q^*-Q = \begin{bmatrix} 0 & -q_{12} \\ 0 & -q_{22} \end{bmatrix}$$

For this two variable model, the bias of β^* is

$$E(\beta^*)-\beta = \begin{bmatrix} 0 & -q_{12} \\ 0 & -q_{22} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$

$$= \begin{bmatrix} -q_{12}\alpha_2 \\ -q_{22}\alpha_2 \end{bmatrix} .$$

Of course, α_2 can be expressed as a linear combination of the true β parameters,

$$\alpha_2 = q_{12}\beta_1 + q_{22}\beta_2$$

since $\alpha = Q'\beta$. By substituting for α_2 in the bias of β^*, we have

$$\begin{aligned} \text{Bias } (\beta_1^*) &= -q_{12}\ (q_{12}\beta_1 + q_{22}\beta_2) \\ &= -q_{12}^2\beta_1 - q_{12}q_{22}\beta_2 \end{aligned}$$

and

$$\begin{aligned} \text{Bias } (\beta_2^*) &= -q_{22}(q_{12}\beta_1 + q_{22}\beta_2) \\ &= -q_{22}q_{12}\beta_1 - q_{22}^2\beta_2 \ . \end{aligned}$$

These expressions for the bias of β^* can be further refined

since in the two-variable case, the correlation matrix is

$$X'X = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$$

The eigenvalues of X'X are $\lambda_1 = 1 + r$ and $\lambda_2 = 1 - r$. The matrix of eigenvectors is

$$Q = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

Thus, the bias of β_1^* and the bias of β_2^* become

$$\begin{aligned} \text{Bias } (\beta_1^*) &= -(\tfrac{1}{\sqrt{2}})^2 \beta_1 - (\tfrac{1}{\sqrt{2}})(-\tfrac{1}{\sqrt{2}}) \beta_2 \\ &= -\tfrac{1}{2} \beta_1 + \tfrac{1}{2} \beta_2 \\ &= \tfrac{1}{2} (\beta_2 - \beta_1) \end{aligned}$$

and

$$\begin{aligned} \text{Bias } (\beta_2^*) &= -(\tfrac{1}{\sqrt{2}})(-\tfrac{1}{\sqrt{2}}) \beta_1 - (-\tfrac{1}{\sqrt{2}})^2 \beta_2 \\ &= \tfrac{1}{2} \beta_1 - \tfrac{1}{2} \beta_2 \\ &= \tfrac{1}{2} (\beta_1 - \beta_2) \ . \end{aligned}$$

Obviously, the bias of β^* will be lessened if β_1 and β_2 are the same sign and about the same magnitude.

Thus, as in the case of ridge regression, a priori information about the true parameters is helpful in

determining whether principal components analysis is appropriate to the model.

There are many principal components estimators, each differentiated by the method used to select principal components. One method is to delete the principal components which correspond to the lowest eigenvalues. However, "lowest" is a subjective classification. A more objective method is to test the statistical significance of the coefficient, $\hat{\alpha}_i$, of each principal component. The traditional t-test can be used to determine whether $\hat{\alpha}_i$ is significantly different from zero and, consequently, whether the corresponding component should be deleted.

Unfortunately, neither of these methods relates the deletion of a principal component to its effect on mean square error. The objective, of course, is to reduce the mean square error of estimating the true parameters. With that in mind, it is desirable to delete components only if this reduces mean square error. Such a criterion, of course, has not been operational since mean square error depends upon the true parameters, σ^2 and β_i, which are usually unknown.

An operational, mean-square-error-related criterion for deleting principal components is proposed in the next chapter. The effectiveness of this proposed criterion versus the traditional t-criteria is investigated in the Monte Carlo experiments described later in this paper.

III. THE LOSS FUNCTION

The loss function is used in judging the accuracy of the estimates of parameters in a regression model. Probably the most widely used loss function is the squared error loss function,

$$\sum_{i=1}^{p} (b_i - \beta_i)^2$$

where the deviation squared $(b_i - \beta_i)^2$ measures the loss incurred when the true parameter, β_i, is estimated by b_i.

Naturally, to minimize the loss for a particular sample, we would want the contribution of each estimate to the loss function to be as small as possible. Of course, if b_i is very unstable, the expression $(b_i - \beta_i)^2$ would have a smaller value if b_i were set equal to zero. In that case, the contribution to the loss function would be only β_i^2. Consequently, when $(b_i - \beta_i)^2 > \beta_i^2$, there would be a net gain, in terms of minimizing the loss function, by setting b_i equal to zero and deleting the corresponding variable from the model. Thus, a comparison of $(b_i - \beta_i)^2$ with β_i^2 becomes a possible criterion for deleting variables from the regression model.

Estimation of the Loss Function for Ridge Estimators

Unfortunately, as they stand, the preceding loss function and deletion criterion are of little help in empirical problems where, typically, the true parameters are not known. However, it is conceivable to estimate the loss function when its expected value, the mean square error, is expressed as the sum of the variance and the square of the bias:

$$E[(b - \beta)^2] = E[(b - Eb)^2] + E[(Eb - \beta)^2]$$

The loss function can then be estimated through its separate components.

Of course, in the case of unbiased estimators, the estimation of the loss function is greatly simplified since the mean square error equals the variance. In contrast, estimation is more difficult for biased estimators due to the bias squared component which depends on the true β parameters which are usually unknown.

A suggested method for estimating the loss function for ridge estimators was recently proposed by Brown (1978). Recall from Chapter II that the ridge estimator and its mean square error for the orthogonalized model, $Y = Z\alpha + \varepsilon$, are

$$\hat{\alpha}^* = (Z'Z + kI)^{-1} Z'Y$$

and

$$MSE(\hat{\alpha}^*) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2} + k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i+k)^2}$$

where $Z = XQ$ and $\alpha = Q'\beta$ and $Z'Z$ is a diagonal matrix of the eigenvalues of $X'X$, the correlation matrix.

Before presenting Brown's method for estimating the mean square error of the ridge estimator, let us derive an equivalent expression for $MSE(\hat{\alpha}^*)$ which will make the exposition easier. As a first step, consider that the ridge estimator can be expressed as a function of the OLS estimator. That is,

$$\begin{aligned}\hat{\alpha}^* &= (Z'Z + kI)^{-1} Z'Y \\ &= (Z'Z + kI)^{-1} (Z'Z)(Z'Z)^{-1} Z'Y \\ &= M\hat{\alpha}\end{aligned}$$

since $\hat{\alpha} = (Z'Z)^{-1} Z'Y$ and defining $M = (Z'Z + kI)^{-1}(Z'Z)$. Of course, an individual ridge estimate can also be expressed as a function of the OLS estimate, i.e.,

$$\hat{\alpha}_i^* = m_i \hat{\alpha}_i$$

where $m_i = \lambda_i/(\lambda_i + k)$ since the diagonal elements of the matrices $(Z'Z + kI)^{-1}$ and $Z'Z$ are $1/(\lambda_i + k)$ and λ_i respectively. Of course, if $k = 0$, then $m_i = 1$ and $\hat{\alpha}_i^* = \hat{\alpha}_i$. In other words, if $k = 0$, then the ridge estimate and the OLS estimate are equal.

Using $\hat{\alpha}_i^* = m_i \hat{\alpha}_i$, the MSE of the ridge estimator can be expressed as

$$\sum_{i=1}^{p} (\hat{\alpha}_i^* - \alpha_i)^2 = \sum_{i=1}^{p} (m_i\hat{\alpha}_i - \alpha_i)^2$$

$$= \sum_{i=1}^{p} (m_i^2\hat{\alpha}_i^2 - 2m_i\hat{\alpha}_i\alpha_i + \alpha_i^2)$$

$$= \sum_{i=1}^{p} [m_i^2(\alpha_i^2 + \sigma^2/\lambda_i) - 2m_i\alpha_i^2 + \alpha_i^2]$$

since $E\hat{\alpha}_i^2 = \alpha_i^2 + \sigma^2/\lambda_i$ and $E(\hat{\alpha}_i\alpha_i) = \alpha_i^2$

$$= \sum_{i=1}^{p} [m_i^2(\sigma^2/\lambda_i) + m_i\alpha_i^2 - 2m_i\alpha_i^2 + \alpha_i^2]$$

$$= \sum_{i=1}^{p} [m_i^2(\sigma^2/\lambda_i) + (m_i - 1)^2\alpha_i^2] .$$

As we have seen, when $k = 0$, then the ridge estimator is equal to the OLS estimator. Thus, if $k = 0$, then $m_i = 1$ and $v(\hat{\alpha}_i^*) = v(\hat{\alpha}_i) = \sigma^2/\lambda_i$. As we would expect for OLS, the bias squared component of MSE vanishes. Consequently, when $k = 0$, then $MSE(\hat{\alpha}^*) = MSE(\hat{\alpha})$.

Although the variance component of $MSE(\hat{\alpha}^*)$ can be estimated using the OLS estimate of σ^2, the bias squared component depends upon the unknown α parameters. Brown points out, however, that the bias squared component can also be estimated if we make certain assumptions about the distributions of the α_i. Since we have no idea about the magnitudes and signs of the α_i, it may not be unreasonable to assume that the α_i are distributed about zero and that their variances are equal and finite. Since we are assuming that

$E(\alpha_i) = 0$, the variance of α_i becomes

$$E(\alpha_i - E\alpha_i)^2 = E(\alpha_i^2) \quad .$$

Further, we are assuming that the variances of the α_i are equal to a constant, c^2, so that,

$$E(\alpha_1^2) = E(\alpha_2^2) = \ldots = E(\alpha_p^2) = c^2 \quad .$$

It will be useful to derive an equivalent expression for c^2 in terms of the eigenvalues of the X'X correlation matrix. Recall that Z'Z is a diagonal matrix of the eigenvalues and that its trace is

$$\begin{aligned} tr(Z'Z) &= tr[(XQ)'XQ] \\ &= tr(Q'X'XQ) \\ &= tr(QQ'X'X) \\ &= tr(X'X) \\ &= p \end{aligned}$$

since there are p standardized explanatory variables in the model. This implies that

$$\sum_{i=1}^{p} \lambda_i = p$$

which also implies

$$\frac{1}{p} \sum_{i=1}^{p} \lambda_i = 1 \quad .$$

We can now state that

$$c^2 = E(\alpha_i^2) = \frac{1}{p} \sum_{i=1}^{p} \lambda_i E(\alpha_i^2) .$$

We can use c^2 and the OLS coefficients, $\hat{\alpha}_i$ to estimate the bias squared component of MSE($\hat{\alpha}^*$),

$$\text{Bias}^2(\hat{\alpha}^*) = \sum_{i=1}^{p} (m_i - 1)^2 \alpha_i^2 .$$

If we assume that the expected values of the squared parameters are equal to c^2, then we can substitute c^2 for α_i^2. Of course, c^2 itself must be estimated since $c^2 = \frac{1}{p} \sum_{i=1}^{p} \lambda_i \alpha_i^2$. This can be accomplished by using the OLS estimates, $\hat{\alpha}_i$, to estimate c^2:

$$\hat{c}^2 = \frac{1}{p} \sum_{i=1}^{p} \lambda_i \hat{\alpha}_i^2 .$$

Now we have a formula for estimating the mean square error of $\hat{\alpha}^*$:

$$\text{est MSE}(\hat{\alpha}^*) = \sum_{i=1}^{p} [m_i^2 \hat{\sigma}^2 / \lambda_i + (m_i - 1)^2 \hat{c}^2] .$$

But how is MSE($\hat{\alpha}^*$) related to MSE($\hat{\beta}^*$), the mean square error of the ridge estimator for original, untransformed model, $Y = X\beta + \varepsilon$? It can be shown that the MSE of the orthogonalized estimator, $\hat{\alpha}^*$, is equal to MSE($\hat{\beta}^*$). Recall that $\alpha = Q'\beta$ which implies that $\hat{\alpha}^* = Q'\hat{\beta}^*$. It follows that

$$\begin{aligned} \text{MSE}(\hat{\alpha}^*) &= E[(\hat{\alpha}^* - \alpha)'(\hat{\alpha}^* - \alpha)] \\ &= E[(Q'\hat{\beta}^* - Q'\beta)'(Q'\hat{\beta}^* - Q'\beta)] \end{aligned}$$

$$= E\,[(Q'\hat{\beta}^* - Q'\beta)'(Q'\hat{\beta}^* - Q'\beta)]$$

$$= E[(\hat{\beta}^* - \beta)'\,QQ'(\hat{\beta}^* - \beta)]$$

$$= E[(\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta)]$$

$$= MSE(\hat{\beta}^*) \ .$$

A Proposed Loss-Function-Related Criterion for Deleting Principal Components

The assumptions underlying the suggested method for estimating the mean square error of ridge estimators can be used to form a proposed loss-function-related criterion for deleting principal components.

As discussed in Chapter II, it is desirable to delete principal components only if this reduces the mean square error of estimating the true parameters. Ideally, we would want to compare $(\hat{\alpha}_i - \alpha_i)^2$, the contribution of an estimate to $MSE(\hat{\alpha})$ when a component is retained, with the contribution, α_i^2, when the component is deleted.

If $(\hat{\alpha}_i - \alpha_i)^2 > \alpha_i^2$, then it would be advantageous to set $\hat{\alpha}_i$ equal to zero and to delete the corresponding component. Unfortunately, this criterion is not operational, since the α_i are usually unknown.

Of course, since $\hat{\alpha}_i$ is unbiased $(\hat{\alpha}_i - \alpha_i)^2$ can be estimated by $\hat{v}(\hat{\alpha}_i) = \hat{\sigma}^2/\lambda_i$. The difficulty is with the other part of the criterion, α_i^2, which is unknown. However, by employing Brown's suggestion and assuming that

$$E(\alpha_1^2) = E(\alpha_2^2) = \ldots = E(\alpha_p^2) = c^2$$

where c^2 is estimated by

$$\hat{c}^2 = \frac{1}{p} \sum_{i=1}^{p} \lambda_i \hat{\alpha}_i^2 ,$$

we can propose the following operational criterion:

$$\text{if } \hat{v}(\hat{\alpha}_i) \leq \hat{c}^2 \quad \text{then retain } Z_i ;$$
$$\text{if } \hat{v}(\hat{\alpha}_i) > \hat{c}^2 \quad \text{then delete } Z_i .$$

Note that in computing $\hat{c}^2$, each OLS estimate is weighted by the corresponding eigenvalue. The lower the eigenvalue, the less weight is given the estimate. Of course, the lower the eigenvalue, the larger the variance of $\hat{\alpha}_i$ since $v(\hat{\alpha}_i) = \sigma^2/\lambda_i$. Thus, the more unstable the estimate, the more likely is the deletion of the corresponding principal component.

IV. THREE RIDGE ESTIMATORS

The accuracy of the ridge estimator,

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

depends on the value of k chosen. The optimal value of k can be determined if the true β parameters and the error term variance, σ^2, are known. Otherwise, k must be chosen based on some other criterion.

Ridge estimators can be classified according to the method used to select k. Three ridge estimators were used in the Monte Carlo experiments described in this paper: a fixed-value estimator; an estimator proposed by Lawless and Wang (1976); and an estimator proposed by Dempster, Schatzoff and Wermuth (1977).

The Lawless-Wang Estimator

The Lawless and Wang value of k is defined as

$$K_B = \frac{p\ \hat{\sigma}^2}{\sum_{i=1}^{p} \lambda_i \hat{\alpha}_i^2}$$

where p is the number of explanatory variables, $\hat{\sigma}^2$ is the OLS estimate of σ^2, λ_i are the eigenvalues of the X'X correlation matrix, and $\hat{\alpha}_i$ are the OLS estimates for the orthogonalized model, $Y = Z\alpha + \varepsilon$. Brown has shown that the

Lawless-Wang value of k minimizes the estimated mean square error of the ridge estimator, $\hat{\alpha}^*$. Recall that

$$\widehat{MSE}(\hat{\alpha}^*) = \sum_{i=1}^{p} [m_i^2(\hat{\sigma}^2/\lambda_i) + (m_i - 1)^2\hat{c}^2]$$

$$= \sum_{i=1}^{p} [m_i^2(\hat{\sigma}^2/\lambda_i) + m_i^2\hat{c}^2 - 2m_i\hat{c}^2 + \hat{c}^2] .$$

If we take the first partial derivative with respect to m_i, we have

$$\frac{\partial \widehat{MSE}(\hat{\alpha}_i^*)}{\partial m_i} = 2[m_i(\hat{\sigma}^2/\lambda_i) + m_i\hat{c}^2 - \hat{c}^2] .$$

Setting this derivative equal to zero, and solving for m_i,

$$m_i(\hat{\sigma}^2/\lambda_i) + m_i\hat{c}^2 - \hat{c}^2 = 0$$

$$m_i[(\hat{\sigma}^2/\lambda_i) + \hat{c}^2] = \hat{c}^2$$

$$m_i = \hat{c}^2/[\hat{\sigma}^2/\lambda_i) + \hat{c}^2].$$

If we substitute $m_i = \lambda_i/(\lambda_i + k)$, we can solve for k:

$$\lambda_i/(\lambda_i + k) = \hat{c}^2/[(\hat{\sigma}^2/\lambda_i) + \hat{c}^2]$$

$$\lambda_i + k = [1 + (\hat{\sigma}^2/\lambda_i\hat{c}^2)]\lambda_i$$

$$k = \lambda_i + (\hat{\sigma}^2/\hat{c}^2) - \lambda_i$$

$$k = \hat{\sigma}^2/\hat{c}^2$$

which is the value of k that minimizes $\widehat{MSE}(\hat{\alpha}^*)$. Since $\hat{c}^2 = (1/p) \sum_{i=1}^{p} \lambda_i\hat{\alpha}_i^2$, then the optimal value of k can also be written as

$$k = \frac{p\,\hat{\sigma}^2}{\sum_{i=1}^{p} \lambda_i \hat{\alpha}_i^2}$$

which is the Lawless-Wang estimator. Notice that the OLS estimates of the α_i are weighted by their eigenvalues. Consequently, the more unstable the estimate, the less weight it will be given in computing K_B.

RIDGM Estimator

Dempster, Schatzoff, and Wermuth (1977) developed RIDGM which is motivated by a Bayesian interpretation of ridge regression. The following formula for the RIDGM estimator was used by Rahuma (1977):

$$K_M = \frac{p\,\hat{\sigma}^2}{\sum_{i=1}^{p} \hat{\alpha}_i^2 - p\hat{\sigma}^2} .$$

Note the similarity to the Lawless-Wang estimator. Also note, however, that in contrast to Lawless-Wang. RIDGM does not weight the $\hat{\alpha}_i^2$ by the corresponding eigenvalues. Consequently, unstable estimates receive as much weight as other

estimates in computing K_M. This implies that RIDGM may produce less stable estimates of k than the Lawless-Wang estimator.

Fixed Value Estimator

As we have seen, both K_B and K_M depend upon the true parameters or their estimates, $\hat{\alpha}_i$ and $\hat{\sigma}^2$. For the sake of comparison, a fixed value of k, unrelated to $\hat{\alpha}_i$ and $\hat{\sigma}^2$, was chosen as the third ridge estimator for use in the Monte Carlo experiments. The value of k was arbitrarily set as 0.1.

V. THE MONTE CARLO EXPERIMENTS

Description

The main objective of the Monte Carlo experiments is to compare the performance of the OLS estimator with seven biased linear estimators, that is, with four principal components estimators and three ridge estimators. Three of the principal components estimators use traditional t-criteria to delete principal components. The fourth principal components estimator employs the proposed loss-function-related criterion. The three ridge estimators are the fixed value ($k = 0.1$) estimator, the Lawless-Wang estimator, and RIDGM.

The Monte Carlo experiments were based on the four-explanatory-variable model used by Hoerl, Kennard, and Baldwin (1975), and Lawless and Wang (1976) in Monte Carlo studies of ridge regression. The Monte Carlo experiments in this paper involved four sets of data, each with 13 observations but characterized by a different level of multicollinearity. One of the data sets was that used by Hoerl, Kennard, and Baldwin, and Lawless and Wang in their Monte Carlo studies. (It was originally published by Hald in 1952.) That data set was characterized by severe multicollinearity. Two other data sets were created by manipulating the observations of the original Hald data to produce

lower levels of multicollinearity. A fourth data set was created to represent perfect orthogonality. (The original data and the orthogonal data set are presented in Appendix Tables 1 and 2 of Appendix A. The creation of the other two data sets is briefly explained in Appendix A.)

The following table shows the diagonal elements, r^{ii}, of the inverted correlation matrices of the four data sets.

Table 1. The r^{ii} and sum of the r^{ii} of the four data sets.

Data Set	r^{11}	r^{22}	r^{33}	r^{44}	Sum of the r^{ii}
1	1.0	1.0	1.0	1.0	4.0
2	5.1	13.7	4.8	15.4	39.0
3	8.3	89.9	9.9	95.9	204.0
4	38.5	254.4	46.9	282.5	622.3

The degree of multicollinearity increases from perfect orthogcnality in Data Set 1 to high multicollinearity in Data Set 4, the Hald data. Since the higher the value of an r^{ii}, the more intercorrelated the corresponding explanatory variable, we can see that X_2 and X_4 are the most intercorrelated of the variables in Data Sets 2, 3 and 4.

For each of the four data sets, twelve groups of Monte Carlo experiments were run. Each group consisted of 400 experiments and was differentiated by the information-to-noise ratio, $(\alpha'\Lambda\alpha)/\sigma^2$, where α is the vector of the true parameters of the orthogonalized model, $Y = Z\alpha + \varepsilon$, and

σ^2 is the error term variance. The noise component of the desired information-to-noise ratio was held constant at $\sigma^2 = 1.0$. The desired information level, $\alpha'\Lambda\alpha$, assumed 12 values selected in a range from 4 to 10,000.[1/]

In each Monte Carlo experiment, the values of the true parameters, α_i, were generated. In this procedure, the first step was to pseudo-randomly generate four uniform variates, u_i, ranging from -0.5 to 0.5. These u_i were adjusted by a factor, M, to give the α_i values for the desired information-to-noise ratio. The adjustment factor, M, was calculated as

$$M = \sqrt{\frac{\text{desired } \alpha'\Lambda\alpha}{\sum_{i=1}^{p} \lambda_i u_i^2}}$$

The true parameters were then computed as

$$\alpha_i = M\, u_i \ .$$

The distribution of the α_i, over all experiments, implies that

$$E(\alpha_i) = 0 \quad \text{for all } i \tag{5.1}$$

and that

[1/] The twelve values selected for the information-to-noise ratio were: 4; 9; 16; 25; 64; 100; 200; 400; 900; 1600; 2500; and 10,000.

$$E(\alpha_1^2) \doteq E(\alpha_2^2) \doteq \ldots \doteq E(\alpha_p^2) \quad . \tag{5.2}$$

Recall that these are the assumptions underlying the proposed method of estimating the mean square error of ridge estimators and underlying the proposed loss-function-related criterion for deleting principal components. Note that although (5.1) and (5.2) are true when considering all experiments, for any one experiment, the α_i^2 values vary greatly, ranging from small positive values to large positive values.

The error terms for each Monte Carlo experiment were pseudo-randomly generated from an approximately normal distribution with mean zero and variance equal to one. Each error term was actually the sum of 12 uniform random variates. Consequently, as pointed out by Newman and Odell (1971), the distribution of the error terms is only approximately normal with the tails not perfectly approximated over large samples. Given the error terms and the parameters, the values of the dependent variable were then computed from $Y = Z\alpha + \varepsilon$.

The α parameters of each experiment were estimated using three methods of estimation: OLS; principal components analysis; and ridge regression. The true mean square error (MSE) of each estimate, $\hat{\alpha}_i^*$, was computed as

$$\text{True MSE}(\hat{\alpha}_i^*) = (\hat{\alpha}_i^* - \alpha_i)^2 \quad .$$

These were accumulated over the 400 experiments for each estimation method, including OLS. The average value of true MSE for the 400 experiments was the basis for comparing the estimation methods.

In addition to true MSE, estimated MSE was also accumulated and averaged. The method of estimating MSE was that suggested by Brown and explained in Chapter III, where

$$\text{est MSE}(\hat{\alpha}_i^*) = [m_i^2 \frac{\hat{\sigma}^2}{\lambda_i} + (m_i - 1)^2 \hat{c}^2]$$

and

$$m_i = \lambda_i/(\lambda_i + k)$$

and

$$\hat{c}^2 = (1/p) \sum_{i=1}^{p} \lambda_i \hat{\alpha}_i^2 .$$

Ordinary Least Squares Results

The effect of multicollinearity on the variances of OLS estimates can be seen from the four sets of data. Recall that the variance of an OLS estimate in an orthogonal regression is $v(\hat{\alpha}_i) = \sigma^2/\lambda_i$ where $1/\lambda_i$ can be considered the variance inflation factor (VIF_i). The eigenvalues and the sum of the VIF_i for the four data sets are shown in Table 2. Note that the $\sum 1/\lambda_i$ is equal to the $\sum r^{ii}$ (shown in Table 1), for the original, untransformed data. (There are slight differences between the two sums due to rounding errors.) The explanatory variables of Data Set 1 are perfectly orthogonal while the explanatory variables of Data

Table 2. The eigenvalues and the sum of the VIF_i of the four data sets.

Data Set	λ_1	λ_2	λ_3	λ_4	Sum of the VIF_i
1	1.0	1.0	1.0	1.0	4.0
2	2.16909	1.61615	0.18404	0.03070	39.07
3	2.17764	1.52105	0.29630	0.00501	203.97
4	2.23570	1.57607	0.18661	0.00162	622.20

Set 4, the Hald data are nearly collinear. In fact, the variance of $\hat{\alpha}$ in Data Set 4 is almost 156 times what it would be if the explanatory variables were orthogonal.

Since $E(\hat{\sigma}^2) = 1$ in the Monte Carlo experiments, the sum of the VIF_i above are the expected values of the OLS variances for the data sets. (In Chapter II, it was shown that $v(\hat{\alpha}) = v(\hat{\beta})$.)

For the purpose of comparison with the other estimators, the OLS results are presented with the principal components and ridge results in Tables 3 through 10.

Principal Components Analysis Results

Two types of criteria were used to delete principal components in the Monte Carlo experiments:

(1) The traditional t-criterion: A component is deleted if

$$\frac{|\hat{\alpha}_i|}{\sqrt{v(\hat{\alpha}_i)}} < t_{\alpha/2,\ n-p-1}$$

where α is the level of significance and $n - p - 1$ is the degrees of freedom. In the Monte Carlo experiments, three critical t-values were used: 1.8; 2.3; and 2.896. These correspond approximately to significance levels 0.11, 0.05, and 0.02, respectively, with 8 degrees of freedom.

(2) The proposed loss-function-related criterion: As explained in Chapter III, a component is deleted if

$$v(\hat{\alpha}_i) > \hat{c}^2$$

where

$$\hat{c}^2 = (1/p) \sum_{i=1}^{p} \lambda_i \hat{\alpha}_i^2 \quad .$$

In order to evaluate the performance of the t-criteria and the proposed loss-function-related criterion, the true MSE (α^*) for each criterion was accumulated and averaged over 400 experiments per information-to-noise ratio. In addition, the true MSE(α^*) for the proposed loss-function-related criterion was estimated using the method suggested by Brown. Although the method was intended to estimate the MSE of ridge estimators, it seemed reasonable to apply it also to the proposed principal components estimator since the proposed loss-function-related criterion is based on the same assumptions regarding the distributions of the α_i.

The results of the Monte Carlo experiments for principal components analysis are as follows:

Data Set 1--No Multicollinearity

In this Data Set, the explanatory variables are perfectly orthogonal. Consequently, the OLS estimates, $\hat{\beta}_i$, are stable. There is nothing to be gained by deleting principal components (assuming, of course, that each explanatory variable is relevant). In this case, the question is not whether the criteria for deleting components can reduce MSE, but whether their use will increase MSE.

The results of the Monte Carlo experiments for Data Set 1 are shown in Table 3. These results are briefly summarized in the following table:

Table 3a. The effect on MSE by the criteria for deleting components, categorized by ranges of information-to-noise ratios.

Deletion Criteria	MSE Reduced (Information-to-Noise Ratios)	MSE Increased (Information-to-Noise Ratios)	MSE Unchanged (No Components Deleted) (Information-to-Noise Ratios)
Proposed Loss-Function-Related Criterion	None	4-16	25-10,000
Traditional t-criteria	None	all	None

The proposed loss-function-related criterion increased MSE only at the low information-to-noise ratios while the t-criteria increased MSE at all ratios. Moreover, the increases in MSE by the proposed criterion were not as great as those resulting from the t-criteria. (Note the results

Table 3. DATA SET 1. True MSE comparisons for a four-explanatory-variable, orthogonal model estimated by OLS and by principal components analysis with the proposed loss-function-related criterion and the traditional t-criteria for deleting components.

Information-to-Noise Ratio	OLS		Principal Components Analysis				
	Full Model		Proposed Criterion[a]		t = 1.8[b]	t = 2.3[b]	t = 2.896[b]
	Average True Variance	Average Estimated Variance	Average True MSE	Average Estimated MSE	Average True MSE	Average True MSE	Average True MSE
4	4.16	4.04	4.36	3.66	4.60	4.57	4.43
9	3.86	3.91	4.13	3.84	5.74	6.94	7.89
16	4.16	4.09	4.24	4.07	6.68	8.68	10.76
25	3.81	3.96	3.81	3.96	6.13	8.30	11.51
64	3.99	4.13	3.99	4.13	5.31	6.98	9.67
100	4.16	4.10	4.16	4.10	5.23	6.56	9.07
200	3.96	3.99	3.96	3.99	4.53	5.14	6.76
400	3.87	3.99	3.87	3.99	4.28	4.72	5.78
900	4.00	3.83	4.00	3.83	4.22	4.42	4.95
1600	4.08	4.00	4.08	4.00	4.24	4.42	4.77
2500	4.05	3.97	4.05	3.97	4.24	4.41	4.71
10000	4.09	3.91	4.09	3.91	4.13	4.23	4.34
Average	4.02	3.99	4.06	3.95	4.78	5.78	7.05

[a] Component i was deleted if Est MSE($\hat{\alpha}_i$) > $\hat{c}^2$.

[b] Component i was deleted if the computed t value was less than t_α.

in Table 3.)

The t-criteria performed somewhat better at the higher information-to-noise ratios. In this range, the t-criteria caused fewer components to be deleted. Consequently, the effect of the t-criteria was less detrimental for these ratios.

The average estimated MSE for the proposed loss-function-related criterion was very close to the average true MSE.

Data Set 2--Low-to-Moderate Multicollinearity

Table 4 shows the results of the Monte Carlo experiments for Data Set 2. These results are summarized in the following table:

Table 4a. The effect on MSE by the criteria for deleting components, categorized by ranges of information-to-noise ratios.

Deletion Criteria	MSE Reduced (Information-to Noise Ratios	MSE Increased (Information-to Noise Ratios	MSE Unchanged (No Components Deleted) (Information-to Noise Ratios
Proposed Loss-Function-Related Criterion	4-64	100-1600	2500 & 10,000
Traditional t-criteria	4-64	100-10,000	None

Each of the criteria reduced MSE at low information-to-noise-ratios. However, the reductions by the proposed criterion were generally greater than those by the t-

Table 4. DATA SET 2. True MSE comparisons for a four-explanatory-variable model with low-to-moderate multicollinearity estimated by OLS and by principal components analysis with the proposed loss-function-related criterion and the traditional t-criteria for deleting components.

Information-to-Noise Ratio	OLS Full Model		Principal Components Analysis				
			Proposed Criterion[a]		t = 1.8[b]	t = 2.3[b]	t = 2.896[b]
	Average True Variance	Average Estimated Variance	Average True MSE	Average Estimated MSE	Average True MSE	Average True MSE	Average True MSE
4	41.9	37.7	3.9	4.8	22.5	15.3	9.5
9	41.5	37.7	7.4	7.2	24.0	20.2	14.2
16	46.0	36.4	11.4	9.6	30.1	25.3	21.9
25	34.8	39.4	14.8	12.7	23.2	20.9	19.1
64	39.4	39.1	27.7	21.9	37.3	34.3	31.7
100	38.0	37.6	41.4	26.7	40.8	42.7	44.1
200	38.7	39.9	54.5	34.4	51.9	59.0	65.6
400	42.2	38.3	52.3	36.9	65.9	77.3	94.7
900	41.2	38.7	47.2	38.6	58.7	67.8	91.9
1600	39.1	38.9	49.0	38.9	55.0	67.3	98.5
2500	39.1	39.1	39.1	39.1	50.0	63.1	84.9
10000	38.9	40.3	38.9	40.3	41.3	46.3	56.5
Average	40.1	38.6	32.3	25.9	41.7	45.0	52.7

[a]Component i was deleted if Est MSE($\hat{\alpha}_i$) > $\hat{c}^2$.

[b]Component i was deleted if the computed t-value was less than t_α.

criteria. (Refer to the results in Table 4.) Also, the increases in MSE at higher ratios were generally smaller by the proposed criterion. Moreover, at very high information-to-noise ratios, the proposed criterion did not delete components and thus, did not increase MSE.

Referring to Table 4, the values or average estimated MSE tend to underestimate true MSE. The underestimation is substantial at information-to-noise ratios from 100 to 1600 which is also the range in which MSE was increased by the use of the proposed loss-function-related criterion.

Data Set 3--Moderate-to-High Multicollinearity

Table 5 shows the Monte Carlo results for Data Set 3. The following table is a brief summary of those results.

Table 5a. The effect on MSE by the criteria for deleting components, categorized by ranges of information-to-noise ratios.

Deletion Criteria	MSE Reduced (Information-to-Noise Ratios)	MSE Increased (Information-to-Noise Ratios)	MSE Unchanged (No Components Deleted) (Information-to-Noise Ratios)
Proposed Loss-Function-Related Criterion	4-400	900-10,000	None
Traditional t-criteria	4-400	900-10,000	None

For this data set, all the deletion criteria had much the same pattern of effect on MSE. In general, however, reductions in MSE were greater and increases in MSE smaller

Table 5. DATA SET 3. True MSE comparisons for a four-explanatory-variable model with moderate-to-high multicollinearity estimated by OLS and by principal components analysis with the proposed loss-function-related criterion and the traditional t-criteria for deleting components.

Information-to-Noise Ratio	Full Model		Principal Components Analysis				
			Proposed Criterion[a]		t = 1.8[b]	t = 2.3[b]	t = 2.896[b]
	Average True Variance	Average Estimated Variance	Average True MSE	Average Estimated MSE	Average True MSE	Average True MSE	Average True MSE
4	209.7	206.3	3.9	4.8	87.6	46.7	27.3
9	197.2	203.7	6.5	6.8	72.6	49.7	38.8
16	184.7	209.0	9.4	8.8	78.5	41.1	27.5
25	193.7	204.8	12.0	11.4	77.5	46.1	36.5
64	181.5	203.2	20.7	21.1	83.0	64.4	44.1
100	174.7	197.5	30.7	30.5	94.0	70.4	55.5
200	205.1	210.8	59.9	54.8	135.4	117.2	89.8
400	226.3	222.8	130.7	97.5	187.7	173.3	142.6
900	223.9	207.4	244.0	163.0	232.4	239.8	240.6
1600	182.9	201.3	262.7	188.3	285.5	322.6	350.6
2500	223.9	207.4	308.6	200.1	339.1	400.2	464.7
10000	220.4	209.3	282.8	207.9	274.5	360.3	525.4
Average	202.0	207.0	114.3	82.9	162.3	161.0	170.3

[a] Component i was deleted if Est MSE(α_i) > $\hat{c}^2$.

[b] Component i was deleted if computed t-value was less than t_α.

by the proposed loss-function-related criterion. (Refer to the results in Table 5).

The values of average estimated MSE for the proposed loss-function-related criterion substantially underestimate true MSE for information-to-noise ratios above 400. This was the range of ratios in which MSE was increased by the proposed-loss-function-related criterion.

Data Set 4--High Multicollinearity

Table 6 shows the results of the Monte Carlo experiments for Data Set 4. The results are briefly summarized in the following table:

Table 6a. The effect on MSE by the criteria for deleting components, categorized by ranges of information-to-noise ratios.

Deletion Criteria	MSE Reduced (Information-to-Noise Ratios)	MSE Increased (Information-to-Noise Ratios)	MSE Unchanged (No Components Deleted) (Information-to-Noise Ratios
Proposed Loss-Function-Related Criterion	4-1600	2500 & 10,000	None
Traditional t-criteria	4-1600	2500 & 10,000	None

Here, under conditions of high multicollinearity, the deletion criteria had much the same pattern of effect on MSE. In general, however, the reductions in MSE were larger by the proposed loss-function-related criterion. (Note the results in Table 6.)

Table 6. DATA SET 4. True MSE comparisons for a four-explanatory variable model with high multicollinearity estimated by OLS and by principal components analysis with the proposed loss-function-related criterion and the traditional t-criteria for deleting components.

Information-to-Noise Ratio	Full Model		Principal Components Analysis				
			Proposed Criterion[a]		t = 1.8[b]	t = 2.3[b]	t = 2.896[b]
	Average True Variance	Average Estimated Variance	Average True MSE	Average Estimated MSE	Average True MSE	Average True MSE	Average True MSE
4	651.1	597.9	4.4	5.1	291.4	190.1	134.2
9	673.6	645.3	6.8	7.4	290.7	180.5	60.9
16	607.4	603.5	9.5	9.5	234.6	110.0	54.2
25	632.5	609.1	14.1	12.6	272.9	118.4	37.9
64	602.9	597.8	24.1	22.6	279.0	189.9	73.3
100	650.6	599.5	35.0	31.9	312.7	191.5	102.7
200	624.1	625.9	54.2	57.1	322.6	198.9	166.2
400	639.6	619.0	105.4	107.4	370.9	261.2	201.5
900	577.7	610.2	258.6	229.9	443.1	364.6	309.5
1600	601.1	632.2	483.6	364.1	542.9	491.4	445.7
2500	648.3	622.7	730.5	465.4	705.6	747.6	687.6
10000	661.9	596.8	835.9	585.6	982.8	1249.7	1592.6
Average	630.9	613.3	213.5	158.2	420.8	358.3	322.2

[a]Component i was deleted if Est MSE($\hat{\alpha}_i$) > $\hat{c}^2$.

[b]Component i was deleted if the computed t-value was less than t_α.

Referring to Table 6, the average estimated MSE for the proposed loss-function-related criterion was substantially underestimated at high information-to-noise ratios (over 900).

A Comparison of the Deletion Criteria

Both the proposed loss-function-related criterion and the traditional t-criteria had similar patterns of effect on MSE, increasing or decreasing MSE at the same information-to-noise ratios. The main difference between the criteria was in the magnitudes of their effects on MSE. Generally, the proposed criterion produced larger reductions in MSE than did the t-criteria over all four data sets. Also, when both types of criteria increased MSE at the higher information-to-noise ratios, the increases were generally smaller for the proposed criterion than for the t-criteria. Moreover, there were ranges of information-to-noise ratios, particularly in Data Set 1 (no multicollinearity), where no components were deleted by the proposed criterion. As a result, at those ratios, the proposed criterion did not change MSE from the OLS value. In contrast, at the same ratios, the t-criteria deleted components and caused MSE to increase. (Actually, the t-criteria caused components to be deleted at all information-to-noise ratios in all data sets.)

The results for both the t-criteria and the proposed criterion were particularly good for Data Set 4, high multicollinearity. Substantial reductions in MSE as compared to OLS were achieved by both criteria at low-to-moderate information-to-noise ratios. (In fact, a reduction of MSE to 0.7% of the OLS variance was achieved at the lowest ratio by the proposed criterion.) However, at high information-to-noise ratios (over 2500), both types of criteria produced higher MSE than OLS. This result implies that the greatest potential of both types of deletion criteria exists under conditions of high multicollinearity at low-to-moderate information-to-noise ratios.

The main implication of the Monte Carlo experiments is that the proposed loss-function-related criterion is more likely to produce good results than the traditional t-criteria. No doubt the t-criteria were less effective than the proposed criterion due to their rigidity. For one thing, the "optimal" t-value which minimizes the cost of Type I and Type II errors is different for each α_i in an experiment. Consider the cost of Type I and Type II errors in terms of mean square error. The effect of a Type I error (i.e., erroneously rejecting the null hypothesis, $\alpha_i = 0$, when it is true) is that $(\hat{\alpha}_i - 0)^2 = \hat{\alpha}_i^2$ is added to MSE($\hat{\alpha}$) instead of zero. Similarly, a Type II error (i.e., not rejecting the null hypothesis that $\alpha_i = 0$, when it is false) contributes $(0 - \alpha_i)^2 = \alpha_i^2$ rather

than zero to MSE($\hat{\alpha}$).

Of course, even when the conclusion about the null hypothesis is correct, the effect on MSE may be detrimental since the t-criteria do not take into account the effect on MSE of deleting or retaining components. For example, for an unstable estimate, it could be that $(\hat{\alpha}_1 - \alpha_1)^2 > \alpha_1^2$. In the interest of minimizing MSE, it will pay, in this case, to set $\hat{\alpha}_1$ equal to zero and delete Z_1, even though, in truth, $\alpha_1 \neq 0$.

The good results obtained by the proposed loss-function-related criterion in deleting components of variables suggest that the proposed criterion might also be effective if applied to the deletion of whole variables. Extending the proposed criterion, a variable would be deleted from the regression model if $v(\hat{\beta}_i) > \hat{c}^2$ where $\hat{\beta}_i$ is the OLS estimate of the true β_i and $\hat{c}^2 = (1/p)\hat{\beta}'X'X\hat{\beta}$. Similar to the assumptions explained earlier about the distributions of the α_i, underlying this criterion, are two assumptions about the distributions of the β_i:

$$(1) \quad E(\beta_i) = 0 \text{ for all } i;$$

$$(2) \quad E(\beta_1^2) = E(\beta_2^2) = \ldots = E(\beta_p^2) \ .$$

A Monte Carlo study to evaluate the effectiveness of the application of the proposed criterion to variable deletion might be an interesting area for further research.

Ridge Regression Results

Three ridge estimators were used in the Monte Carlo experiments:

(1) Fixed Value Estimator, $k = 0.1$

(2) RIDGM

(3) Lawless-Wang Estimator

The basis for comparing the performance of the estimators is true MSE($\hat{\alpha}^*$), accumulated and averaged over 400 experiments for each estimator. Also recorded over 400 experiments, was the average value of estimated MSE($\hat{\alpha}^*$) which was computed using the method suggested by Brown.

Data Set 1--No Multicollinearity

The explanatory variables in Data Set 1 are orthogonal. Consequently, there is nothing to be gained by using a biased estimator instead of OLS. Still, it is interesting

to compare the mean square errors of the ridge estimators with OLS.

Table 7 shows the results of the Monte Carlo experiments for Data Set 1. The expected value of the OLS variance is 4.0. Both RIDGM and the Lawless-Wang estimator were close to 4.0 at all information-to-noise ratios. The fixed value estimator (k = 0.1), however, produced substantially increased MSE at information-to-noise ratios above 200.

The values of averaged estimated MSE were close to the true values except in one instance (RIDGM, information-to-noise ratio, 4).

Data Set 2--Low-to-Moderate Multicollinearity

Table 8 shows the results of the Monte Carlo experiments for Data Set 2. For this data set, the expected value of the OLS variance is 39.07. All three ridge estimators produced substantial improvement in MSE compared to OLS at low-to-moderate information-to-noise ratios (4 to 200). In comparing the three ridge estimators, k = 0.1 and Lawless-Wang did somewhat better than RIDGM at low information-to-noise ratios. At high information ratios, RIDGM and the Lawless-Wang estimator produced results similar to OLS but k = 0.1 caused large increases in MSE.

The values of average estimated MSE were generally quite close to the true values, although there was a

Table 7. DATA SET 1. Average true and average estimated mean square error (MSE) of the ridge estimator for k = 0.0 (OLS), for k = 0.1, and for the RIDGM and Lawless-Wang ridge estimators for 400 experiments per information-to-noise ratio for the four-explanatory-variable orthogonal model.

Information-to-Noise Ratio	OLS k = 0.0		k = 0.1		RIDGM		Lawless-Wang	
	Average True MSE	Average Estimated MSE	Average True MSE	Average Estimated MSE	Average True MSE	Average Estimated MSE	Average True MSE	Average Estimated MSE
4	4.16	4.04	3.45	3.41	4.23	9.68	2.71	2.43
9	3.86	3.91	3.22	3.34	3.99	3.56	3.19	2.77
16	4.16	4.09	3.56	3.55	4.36	3.66	3.81	3.20
25	3.81	3.96	3.35	3.51	3.93	3.41	3.70	3.35
64	3.99	4.13	3.85	3.97	4.08	3.83	4.03	3.82
100	4.16	4.10	4.28	4.25	4.23	3.90	4.20	3.90
200	3.96	3.99	4.89	4.98	3.97	3.89	3.96	3.89
400	3.87	3.99	6.71	6.62	3.90	3.94	3.90	3.94
900	4.00	3.83	10.85	10.63	4.01	3.81	4.01	3.81
1600	4.08	4.00	16.05	16.62	4.06	3.99	4.06	3.99
2500	4.05	3.97	24.12	23.96	4.06	3.96	4.06	3.96
10000	4.09	3.91	86.85	85.83	4.10	3.91	4.10	3.91
Average	4.02	3.99	14.27	14.22	4.08	4.30	3.81	3.58

Table 8. DATA SET 2. Average true and average estimated mean square error (MSE) of the ridge estimator for k = 0.0 (OLS), for k = 0.1, and for the RIDGM and Lawless-Wang ridge estimators for 400 experiments per information-to-noise ratio for the four-explanatory-variable model with low-to-moderate multicollinearity.

Information-to-Noise Ratio	OLS k = 0.0		k = 0.1		RIDGM		Lawless-Wang	
	Average True MSE	Average Estimated MSE	Average True MSE	Average Estimated MSE	Average True MSE	Average Estimated MSE	Average True MSE	Average Estimated MSE
4	41.9	37.7	4.2	4.4	14.0	9.1	3.4	3.8
9	41.5	37.7	6.7	6.1	17.6	12.6	6.0	5.5
16	46.0	36.4	9.0	7.7	20.9	27.5	8.7	7.1
25	34.8	39.4	10.2	9.9	14.4	10.9	10.7	9.2
64	39.4	39.1	17.1	17.0	21.7	16.3	18.8	14.8
100	38.0	37.6	23.5	23.4	25.5	18.9	24.7	17.7
200	38.7	39.9	41.7	41.1	31.0	25.1	32.7	23.4
400	42.2	38.3	78.3	76.9	39.4	29.4	40.7	28.1
900	41.2	38.7	156.4	168.9	38.2	33.9	41.3	32.8
1600	39.1	38.9	288.4	291.4	38.9	35.8	44.4	34.9
2500	39.1	39.1	436.2	457.9	39.6	37.0	44.1	36.4
10000	38.9	40.3	1791.6	1831.5	38.6	39.8	42.0	39.5
Average	40.1	38.6	238.6	244.7	28.3	24.7	27.1	21.1

tendency to underestimate true MSE.

Data Set 3--Moderate-to-High Multicollinearity

The results of the Monte Carlo experiments for Data Set 3 are shown in Table 9. The expected value of the OLS variance for this data set is 204. All three ridge estimators did much better than OLS at information-to-noise ratios below 1600. Of the three, however, RIDGM was the least effective, particularly at low information-to-noise ratios. However, RIDGM did somewhat better than the other estimators at higher ratios. In contrast, at these high ratios, k = 0.1 did poorly, even compared with OLS.

Overall, the values of average estimated MSE were fairly close to the true values. However, there was a tendency to underestimate the true MSE($\hat{\alpha}^*$) for RIDGM and Lawless-Wang.

Data Set 4--High Multicollinearity

Table 10 shows the results of the Monte Carlo experiments for Data Set 4. The expected value of the OLS variance for this data set is 622. Both RIDGM and Lawless-Wang produced substantial improvement over OLS at all information-to-noise ratios. However, the improvement in MSE became less dramatic at high ratios. Of the two, the Lawless-Wang estimator was the more effective, especially at low-to-moderate information-to-noise ratios. Similarly,

Table 9. DATA SET 3. Average true and average estimated mean square error (MSE) of the ridge estimator for k = 0.0 (OLS), for k = 0.1, and for the RIDGM and Lawless-Wang ridge estimators for 400 experiments per information-to-noise ratio for the four-explanatory-variable model with moderate-to-high multicollinearity.

Information-to-Noise Ratio	OLS k = 0.0		k = 0.1		RIDGM		Lawless-Wang	
	Average True MSE	Average Estimated MSE	Average True MSE	Average Estimated MSE	Average True MSE	Average Estimated MSE	Average True MSE	Average Estimated MSE
4	209.7	206.3	4.3	5.3	44.0	15.1	3.0	3.8
9	197.2	203.7	5.7	6.6	44.7	15.9	5.2	5.6
16	184.7	209.0	7.5	8.2	53.0	38.9	7.7	7.4
25	193.7	204.8	9.4	10.4	43.9	19.1	10.6	9.8
64	181.5	203.2	19.2	19.6	47.4	25.7	20.4	18.5
100	174.7	197.5	27.3	28.9	54.1	32.5	29.6	25.7
200	205.1	210.8	51.3	52.9	80.9	49.3	51.1	41.9
400	226.3	222.8	106.8	99.3	111.3	74.8	92.7	66.1
900	223.9	207.4	216.7	224.6	141.8	109.9	136.7	100.9
1600	182.9	201.3	381.7	404.3	153.8	133.1	165.8	125.9
2500	223.9	207.4	597.0	613.7	192.4	153.7	207.0	145.2
10000	220.4	209.3	2331.7	2557.0	220.0	191.1	272.4	185.6
Average	202.0	207.0	313.2	335.9	98.9	71.6	83.5	61.4

Table 10. DATA SET 4. Average true and average estimated mean square error (MSE) of the ridge estimator for k = 0.0 (OLS), for k = 0.1, and for the RIDGM and Lawless-Wang ridge estimators for 400 experiments per information-to-noise ratio for the four-explanatory-variable model with high multicollinearity.

Information-to-Noise Ratio	OLS k = 0.0		k = 0.1		RIDGM		Lawless-Wang	
	Average True MSE	Average Estimated MSE	Average True MSE	Average Estimated MSE	Average True MSE	Average Estimated MSE	Average True MSE	Average Estimated MSE
4	651.1	597.9	4.2	4.6	158.9	38.2	3.6	4.1
9	673.6	645.3	6.0	6.6	227.4	114.4	5.3	6.1
16	607.4	603.5	8.0	8.4	116.3	39.4	7.6	7.9
25	632.5	609.1	11.2	11.1	126.0	43.3	11.4	10.8
64	602.9	597.8	21.3	21.4	141.8	49.1	22.6	20.4
100	650.6	599.5	32.7	31.5	159.2	59.5	33.8	29.0
200	624.1	625.9	55.8	58.9	167.3	77.4	54.4	50.8
400	639.6	619.0	110.0	114.5	211.5	112.7	102.0	87.8
900	577.7	610.2	247.9	256.1	257.0	182.1	190.9	159.9
1600	601.1	632.2	430.3	447.8	321.1	248.5	293.6	228.1
2500	648.3	622.7	670.3	705.5	413.2	309.7	390.6	284.5
10000	661.9	596.8	2706.6	2771.9	599.7	470.6	641.6	448.2
Average	630.9	613.3	358.7	369.9	241.6	145.4	146.5	111.5

k = 0.1 was also more effective than RIDGM except at the very high information ratios, 2500 and 10,000, where k = 0.1 did poorly, even when compared to OLS.

The values of average estimated MSE were very close to the true values for k = 0.1. Average estimated MSE was also close for the Lawless-Wang estimator except at high information-to-noise ratios where true MSE($\hat{\alpha}^*$) was underestimated. As for RIDGM, overall, the average estimated values of MSE($\hat{\alpha}^*$) were not very close to the true values.

A Comparison of the Ridge Estimators

The Monte Carlo experiments imply that the potential of ridge regression increases as the degree of multicollinearity increases. In fact, the ridge estimators were most effective for Data Set 4, high multicollinearity. Here the Lawless-Wang estimator reduced true MSE($\hat{\alpha}^*$) to as little as 0.5% of OLS.

Of the three ridge estimators, the fixed k-value estimator has the least to recommend it since its effectiveness depends upon its chance proximity to the optimal k-value. In the Monte Carlo experiments, k = 0.1 did quite well at low and moderate information-to-noise ratios. However, at higher ratios, k = 0.1 apparently was not close to the optimal value of k. As a result, at these ratios, the fixed value estimator produced values of true MSE higher than OLS.

In comparing RIDGM with Lawless-Wang, RIDGM produces similar and in some cases, better results than Lawless-Wang at high information-to-noise ratios. This pattern was more often evident at low and moderate levels of multicollinearity (Data Sets 2 and 3). This similarity in MSE at low levels of multicollinearity, is understandable since the difference between the Lawless-Wang estimator and RIDGM is the weighting of the Lawless-Wang k-value by the eigenvalues of the X'X correlation matrix. At lower levels of multicollinearity, very low eigenvalues do not occur, thus, the importance of the weighting by the eigenvalues is diminished. Consequently, when multicollinearity is low, the difference between RIDGM and Lawless-Wang is also diminished. The results of the Monte Carlo experiments imply that RIDGM is more likely to do better than Lawless-Wang at low levels of multicollinearity and high information-to-noise ratios.

However, considering the remarkable reduction in MSE at all information-to-noise ratios under conditions of high multicollinearity, the Lawless-Wang estimator emerges as the best of the three ridge estimators as a possible remedy for multicollinearity.

Summary of the Results

In the Monte Carlo experiments, both principal components analysis and ridge regression produced their greatest improvement over OLS under conditions of high multicollinearity (Data Set 4). This improvement is understandable

since the variance of the OLS estimator can be greatly inflated under such conditions, giving a biased estimator a good chance of offsetting its bias with a reduction in variance. Of the principal components estimators used in the Monte Carlo experiments, the proposed loss-function-related criterion produced the best results, and of the ridge estimators, the Lawless-Wang estimator produced the best results. Since both of these estimators are oriented toward minimizing mean square error, it is not surprising that in their groups, they would produce the greatest improvement over OLS. Interestingly, the Monte Carlo results for the two estimators were close in magnitude, with Lawless-Wang doing slightly better overall than the proposed loss-function-related criterion, particularly at high information-to-noise ratios.

The Accuracy of Estimated Mean Square Error

Although the true mean square errors of the coefficients were known in the Monte Carlo experiments, the true mean square errors were also estimated in order to simulate empirical problems in which the true parameters are unknown. The mean square error was estimated for OLS, for the proposed loss-function-related criterion for deleting principal components, and for the three ridge estimators. To review, the mean square error of an OLS coefficient was estimated by

$$MSE(\hat{\alpha}_i) = \hat{v}(\hat{\alpha}_i) = \hat{\sigma}^2/\lambda_i$$

since the OLS estimates are unbiased. In principal components analysis, estimated MSE is equivalent to OLS for the coefficients of components retained in the model. However, when components are deleted, some bias must be included in MSE. If a component is deleted, then $\hat{\alpha}_i$ is set equal to zero and true MSE $(\hat{\alpha}_i)$ becomes

$$MSE(\hat{\alpha}_i) = (0 - \alpha_i)^2 = \alpha_i^2 .$$

In the Monte Carlo experiments, α_i^2 and thus $MSE(\hat{\alpha}_i)$ were estimated by

$$\hat{c}^2 = (1/p) \sum_{i=1}^{p} \lambda_i \hat{\alpha}_i^2$$

when a component was deleted by the proposed loss-function-related criterion.

For the ridge estimators, mean square error was estimated by

$$MSE(\hat{\alpha}_i^*) = [m_i^2(\hat{\sigma}^2/\lambda_i) + (m_i - 1)^2\hat{c}^2]$$

where $m_i = \lambda_i/(\lambda_i + k)$.

Although some indication of the accuracy of the estimates of MSE can be gained by comparing the average estimated MSE with the average true value at each information-to-noise ratio, this does not give any indication of the stability of these estimates of MSE.

In order to judge the reliability of estimated MSE, the average mean square error of the estimated MSE was computed for each information-to-noise ratio of Data Set 4. This was computed as

$$\begin{matrix}\text{Average} \\ \text{MSE of} \\ \widehat{\text{MSE}}(\hat{\alpha}^*)\end{matrix} = \sum_{i=1}^{400} [\widehat{\text{MSE}}(\hat{\alpha}^*) - \text{MSE}(\hat{\alpha}^*)]^2/400 \ .$$

Data Set 4 was chosen because principal components analysis and ridge regression would most often be used under conditions of high multicollinearity where they are most likely to give good results.

Table 11 shows the average mean square error of estimated $\text{MSE}(\hat{\alpha}_i^*)$. For the sake of comparison, the average mean square error of the estimated variance of the OLS estimates was also computed. Note that the estimated $\text{MSE}(\hat{\alpha}_i^*)$ for the biased estimators is more accurate than $\hat{v}(\hat{\alpha}_i)$ for OLS except at very high information-to-noise ratios. These results imply that Brown's suggestion that α_i^2 be estimated by $\hat{c}^2$ can produce reliable estimates of MSE.

Table 11. The mean square error of estimated MSE($\hat{\alpha}^*$) of various estimators for 400 experiments per information-to-noise ratio for Data Set 4--high multicollinearity.

Information-to-Noise Ratio	OLS	Ridge Regression			Principal Components Analysis
			Lawless-Wang	RIDGM	Proposed Loss-Function-Related Criterion
	k = 0.0	k = 0.1			
4	950,504	23	13	185,160	38
9	873,243	26	23	228,646	37
16	875,580	49	47	180,903	74
25	703,861	92	101	66,047	142
64	926,549	419	444	99,493	537
100	810,782	1,033	1,053	74,620	1,202
200	836,884	4,086	4,171	146,921	4,075
400	902,272	15,659	13,713	146,353	14,876
900	774,146	71,598	50,614	149,758	89,272
1600	762,274	217,228	112,290	204,873	265,997
2500	1,036,953	571,134	295,346	353,719	900,822
10000	899,306	8,260,684	938,423	697,868	1,951,428

VI. APPLICATION TO A FISHERIES EVALUATION MODEL

The Monte Carlo experiments presented earlier indicated that ridge regression and principal components analysis can provide much more accurate results than OLS under conditions of high multicollinearity at low and moderate information-to-noise ratios. It would be interesting to apply these methods to an empirical problem where the true parameters are unknown but some a priori information exists about their values. For this application, a recreation demand model was chosen and specified using the travel cost method.

The purpose of this chapter is to demonstrate the mechanics involved in applying the biased methods of estimation to an empirical problem. Although the discussion in this paper has been mostly in terms of an orthogonalized regression model, it is not necessary to orthogonalize the explanatory variables in order to apply ridge regression and the method for estimating MSE.

In this empirical example, a comparison of the biased estimators versus OLS is made, based on estimated MSE and on prior information about the coefficients. However, the conclusions drawn from this comparison represent "best guesses" about the reliability of the regression results

and do not represent a crucial test of the effectiveness of the estimators, as do the Monte Carlo experiments presented earlier.

The Travel Cost Method

One simple and familiar demand model relates the quantity of a good demanded per unit of time, ceteris paribus, to its own price (Henderson and Quandt, p. 27):

$$q_1 = D(p_1) \quad . \tag{6.1}$$

It would seem straightforward to specify the demand for a recreation good in the same way. That is, to specify the rate of participation in the recreation activity as a function of its own price. However, prices charged for recreation activities are either non-existent or show too little variation to allow the estimation of a conventional

demand function such as (6.1) (Dwyer, Kelly, and Bowes, 1977).

A widely used approach to the estimation of demand for recreation goods is a two-stage procedure known as the travel cost method. The first stage of the procedure is to estimate the demand for the total recreation experience, i.e., the on-site recreation experience plus the travel necessary to reach the site. The rate of participation in the total recreation experience is considered to be a function of the travel costs, i.e., the dollar and time cost incurred both at the site and while travelling to the site (Clawson and Knetsch, p. 62; Dwyer, Kelly, and Bowes, 1977). The assumption is that people who face higher travel costs will have a lower participation rate than those who are similar in other respects. Of course, travel costs increase with the distance from the recreation site. Thus, people living far from the site would be expected to participate at a lower rate than those living near the site (Clawson and Knetsch, p. 64-70; Dwyer, Kelly, and Bowes, 1977).

The second stage of the travel cost method is the estimation of the demand for the recreation activity itself. This estimation is based on the assumption that if travel costs were increased for people living near the site, their participation rate would fall to the rate of people farther away who face the same level of travel costs. Such hypothetical increases in travel costs for groups at varying

distances from the site simulate increases in fees at the site. In that sense, travel costs are surrogate prices. Thus, the range of travel costs facing people at varying distances from the site allows a demand relationship to be estimated.

One problem hindering the estimation of travel cost models is the intercorrelation of the explanatory variables. As pointed out by Clawson and Knetsch (p. 62), time costs and dollar costs are usually highly correlated since both increase with distance. As a result, the estimated coefficients of these variables often have variances which are greatly inflated. Moreover, severe multicollinearity may cause the coefficients to have incorrect signs and to be statistically insignificant. Of course, in such cases, the model should be estimated with the individual observations in order to reduce the degree of multicollinearity. In lieu of this solution, it has been common to delete the time cost variable from the regression model (Brown and Nawas, 1973). The rationalization is that the precision of the dollar cost coefficient is increased by removing the highly correlated time cost variable. Although deletion of a variable is an attractive alternative to imprecise and nonsensical regression results, it should be kept in mind that this alternative may incur specification bias.

To realize how serious the specification bias can be in this case, suppose that the first stage demand model is

specified as the linear statistical function,

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad (6.2)$$

where Y is the rate of participation in the total recreation experience;

X_1 is the dollar cost of the total recreation experience;

X_2 is the time cost of the total recreation experience;

ε is the error term.

Recalling the discussion of specification bias in Chapter I, we know from (1.2) that if X_2 is omitted from the model, then the bias of $\hat{\beta}_1$ will be

$$E(\hat{\beta}_1) - \beta_1 = \alpha \beta_2$$

where α is the estimated coefficient in the regression of X_2 on X_1,

$$X_2 = \alpha X_1 + \varepsilon .$$

Of course, the magnitude and the direction of the bias of $\hat{\beta}_1$ depends on the signs and magnitudes of α and β_2. We know that α is positive and increases in magnitude with r_{12}, the simple correlation between X_1 and X_2, since

$$\alpha = r_{12} \left[\frac{\sum x_2^2}{\sum x_1^2} \right] 1/2$$

(Neter and Wasserman, p. 91). We can assume that β_2 is negative since, according to economic theory, the quantity demanded of a normal good is negatively related to the price of the good. As for the magnitude of β_2, there is some prior information available. A study by Brown and Nawas (1973) of participation in big game hunting in Oregon, indicated that time cost and dollar cost had about equal statistical effect on the participation rate. (The t-values of the coefficients were about equal. Disaggregated data was used which reduced the multicollinearity and allowed the separate effects of the variables to be estimated.) This information implies that a substantial negative bias of the coefficient of dollar cost will exist if the time cost variable is omitted from the model

$$(\alpha \beta_2 < 0, \text{ since } \alpha > 0 \text{ and } \beta_2 < 0).$$

If this is the case, the first stage demand relationship will be underestimated. In turn, this underestimation will cause the participation rates predicted in the second stage to be underestimated. For example, if dollar cost is increased for participants near the site, their participation rate will fall. However, it will not fall to the level of the more distant participants who face the same dollar cost. Even though the dollar costs for the two groups of participants are equal, the participants near the site will have lower time costs and, thus, lower travel costs than

participants farther away. Thus, attributing the lower participation rates of farther groups to the nearer groups results in underestimates of the rates of participation.

Although it is a common practice, deletion of the time cost variable is clearly unacceptable. Of course, the more correlated the time cost variable, the more likely is it to be deleted and, ironically, the greater will be the resulting specification bias. As a result, the problem has been viewed as a poor choice between multicollinearity and specification bias. The Monte Carlo experiments, however, imply that there are alternatives to this dilemma. Both principal components analysis and ridge regression can be effective in coping with multicollinearity. Conditions will be favorable for both methods under conditions of high multicollinearity at low and moderate information-to-noise ratios.

The Data and the Model

The data to which the travel cost method will be applied are from a survey taken in 1962 of salmon and steelhead anglers in Oregon (Brown, Singh, and Castle, 1964). The data were aggregated into 35 subzones with observations on fishing days per subzone, variable costs per subzone, average family income, trip miles per subzone, average miles traveled per trip by subzone, and population per subzone. (These data are presented by Brown,

Singh, and Castle, p. 43). From these data, Brown, Singh, and Castle estimated the following function using OLS:

$$Y_j = 2.4730 - 0.17456X_{1j} - 0.00320X_{2k} + 0.00993X_{3j} \quad (6.3)$$
$$R^2 = 0.512 \qquad (.05380) \qquad (.002995) \qquad (.003004)$$

where Y_j is salmon and steelhead fishing days taken per unit of population of subzone j;

X_{1j} is average variable cost per salmon and steelhead fishing day of subzone j;

X_{2k} is average miles per salmon and steelhead fishing trip for the main distance zone in which the jth subzone falls;

X_{3j} is average family income;

with the standard errors in parentheses. In this specification, distance is a proxy variable for time cost. The inclusion of the income and distance variables was an important contribution by Brown, Singh, and Castle over earlier travel cost models. For these data, the simple correlation between the distance and cost variables was 0.627. Although the distance coefficient had the expected sign, it was not statistically significant due to the intercorrelation.

Another well-known form of the travel cost model would have been the specification of per capita fishing trips, rather than days, as the dependent variable (Dwyer, Kelly, and Bowes, 1977):

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2k} + \beta_3 X_{3j} + \beta_4 X_{3j}^2 + \varepsilon \quad (6.4)$$

where Y_j is salmon and steelhead fishing trips taken per capita;

X_{1j} is average variable cost for salmon and steelhead fishing trip of subzone j;

Y_{2k} is average miles per salmon and steelhead fishing trip for the main distance zone in which the jth subzone falls;

X_{3j} is average family income.

The square of the income variable was included because further analysis by Brown (1976) on model (6.3) indicated that income should have a significant positive effect and that income squared should have a negative effect on days per capita. However, including income squared increases the degree of multicollinearity due to its high correlation with the income variable.

Model (6.4) was estimated using OLS with the following result:

$$\begin{aligned} Y_j = {} & 6.5831\text{E-}06X_{1j} - 6.4383\text{E-}06X_{2k} \\ & (1.4538\text{E-}05) \quad (1.5962\text{E-}06) \\ & + 9.2976\text{E-}06X_{3j} - 3.1673\text{E-}08X_{3j}^2 \\ & (3.6570\text{E-}06) \quad (1.8379\text{E-}08) \quad R^2 = 0.674 \end{aligned} \tag{6.5}$$

Here, the distance and dollar cost variables were highly correlated since $r_{12} = 0.835$. Due to the instability caused by this intercorrelation, the dollar cost coefficient is not significant and takes the wrong sign.

The sum of the VIF_i was 39.2 which was considered as low-to-moderate multicollinearity in the Monte Carlo experiments. In order to increase the degree of multicollinearity,

the data were further aggregated by averaging 8 subsets of the 35 observations, resulting in 8 observations. (This smaller set of data, the standard deviations, the correlation matrix, the error term variance, the diagonal elements of the inverted correlation matrix, and the eigenvalues of the correlation matrix are shown in Appendix B.) The sum of the VIF_i for these data was 350, indicating a high degree of multicollinearity. The information-to-noise ratio was 53, considered as moderate in the Monte Carlo experiments.

Using these highly aggregated data, the model in (6.4) was estimated using OLS with the following results:

$$\begin{aligned} Y_j &= \underset{(7.4419E-05)}{6.5593E-05X_{ij}} - \underset{(6.5431E-06)}{1.1341E-05X_{2k}} \\ &+ \underset{(1.3746E-05)}{6.2538E-06X_{3j}} - \underset{(7.7541E-08)}{3.9344E-08X_{3j}^2} \qquad R^2 = 0.946 \end{aligned} \tag{6.5}$$

None of the coefficients are significant at the 5% level, even though the overall regression is significant. This result is a symptom of multicollinearity. Note that in addition to being insignificant, the dollar cost variable takes the wrong sign. However, if the distance variable is dropped from the model, then the dollar cost coefficient becomes significant and takes the expected sign:

$$Y_j = -6.2470E\text{-}05X_{1j} + 2.1244E\text{-}05X_{3j} \quad (6.6)$$
$$(1.0903E\text{-}05) \qquad (1.3091E\text{-}05)$$
$$- 1.1028E\text{-}07X_{3j}^2 .$$
$$(8.0690E\text{-}08) \qquad R^2 = 0.892$$

Faced with a choice between the nonsensical results in (6.5) and the specification bias in (6.6), the researcher might consider principal components analysis and ridge regression as possible alternatives. The high multicollinearity and the moderate information-to-noise ratio in this case suggest (based on the Monte Carlo results) two biased estimation methods are likely to produce more accurate results than OLS estimation of the full model.

Principal Components Analysis

As the first step in principal components analysis, the travel cost model in (6.4) was transformed to the orthogonal form, $Y = Z\alpha + \varepsilon$, where $Z = XQ$, $\alpha = Q'\beta$, and $Z'Z = \Lambda$. The Z variables are the principal components. The transformed model was estimated using OLS on the highly aggregated data with this result:

$$Y = -3.3178E\text{-}01Z_1 - 6.1004E\text{-}01Z_2 \quad (6.7)$$
$$(8.7854E\text{-}02) \qquad (1.0387E\text{-}01)$$
$$-2.2074Z_3 + 6.6021E\text{-}01Z_4$$
$$(1.1075) \qquad (2.2191)$$

As explained earlier in Chapter II, the $\hat{\alpha}$ coefficients in this regression equation are linear combinations of the $\hat{\beta}$ coefficients in (6.5).

The Traditional t-criteria

Three critical t-values for deleting principal components were chosen. These values of t were 2.353, 3.182, and 4.541, corresponding to significance levels of 0.10, 0.05, and 0.02, respectively, with 3 degrees of freedom. (Approximately the same levels of significance were used in the Monte Carlo experiments of Chapter V.)

The calculated t-values for the principal components were:

Z_i	t^*
Z_1	3.777
Z_2	5.873
Z_3	1.993
Z_4	0.298

where $t^* = |\hat{\alpha}_i|/\sqrt{v(\hat{\alpha}_i)}$. A component is deleted if $t^* < t_{\alpha/2}$. Consequently, all three t-criteria require the deletion of Z_3 and Z_4, while Z_1 should also be deleted if $t_{\alpha/2} = 4.541$. In order to delete a principal component, the corresponding $\hat{\alpha}_i$ can be set equal to zero. Let α^* denote the resulting vector of $\hat{\alpha}$ coefficients when some coefficients are set equal to zero. The α_i^* can be

transformed into the principal components estimates, β_i^*, the coefficients of the standardized X variables, since $\beta^* = Q\alpha^*$.

Deleting both Z_3 and Z_4, and transforming to the β^* values, we have

$$Y = -0.4556X_1 - 0.4922X_2 + 0.1203X_3 + 0.1338X_3^2 \quad (6.8)$$

Note that the dollar cost coefficient has the expected sign but the coefficient of income squared is contrary to expectation. If Z_1 is also deleted and α^* is transformed to β^*, we get

$$Y = 0.4608X_1 + 0.5619X_2 - 0.4778X_3 - 0.4937X_3^2 .$$

Note that neither the dollar cost coefficient nor the income coefficient have the expected signs in this case.

The Proposed Loss-Function-Related Criterion

The proposed loss-function-related criterion for deleting principal components requires a comparison of $v(\hat{\alpha}_i)$ with $\hat{c}^2 = (1/p) \sum \lambda_i \hat{\alpha}_i^2$. If $v(\hat{\alpha}_i) > \hat{c}^2$, then Z_i is deleted. In this case, $\hat{c}^2 = 0.2366$ and the estimated variances of the $\hat{\alpha}_i$ are as follows:

$\hat{\alpha}_i$	$\hat{v}(\hat{\alpha}_i)$
$\hat{\alpha}_1$	0.0077
$\hat{\alpha}_2$	0.0108
$\hat{\alpha}_3$	0.12267
$\hat{\alpha}_4$	4.9245

Consequently, the proposed loss-function-related criterion indicates that Z_3 and Z_4 should be deleted. This conclusion coincides with the results of the two t-criteria, $t_{\alpha/2} = 2.353$ and $t_{\alpha/2} = 3.182$. Thus, the principal components estimates using this criterion would be those in (6.8).

Using the proposed estimate of mean square error discussed in Chapter III, we obtain the following result for the principal components estimates in (6.8):

$$\begin{aligned} MSE(\hat{\alpha}_1) &= \hat{v}(\hat{\alpha}_1) = 0.0077 \\ MSE(\hat{\alpha}_2) &= \hat{v}(\hat{\alpha}_2) = 0.0108 \\ MSE(\alpha_3^*) &= \hat{c}^2 = 0.2366 \\ MSE(\alpha_4^*) &= \hat{c}^2 = 0.2366 \\ \hline MSE(\alpha^*) &= MSE(\beta^*) = 0.4917 \end{aligned}$$

Recall that $\hat{c}^2$ is an estimate of the true bias squared component which is added to MSE when Z_i is deleted.

In this case, $MSE(\beta^*)$ is estimated to be 0.4917 which is much less than 6.1691, the estimated variance of the OLS estimates. The application of principal components analysis to this highly aggregated and highly intercorrelated set of data has apparently produced more accurate estimates than OLS

Ridge Regression

Three Ridge Estimators

The following three ridge estimators used in the Monte Carlo experiments were used to estimate model (6.4) with the highly aggregated set of data.

(1) RIDGM. The RIDGM value of k for this sample was calculated as

$$K_M = \frac{p\hat{\sigma}^2}{\sum\hat{\alpha}_i^2 - p\hat{\sigma}^2} \equiv \frac{p\hat{\sigma}^2}{\sum\hat{\beta}_i^2 - p\hat{\sigma}^2} = 0.0125,$$

using the $\hat{\alpha}$ values from the orthogonalized regression equation in (6.7), or equivalently, using the standardized $\hat{\beta}$ estimates from the following OLS regression on the standardized X variables:

$$Y = 1.0520X_1 - 1.9727X_2 + 0.6257X_3 - 0.6334X_3^2.$$

(2) Lawless-Wang. The Lawless-Wang value of k was computed as

$$K_B = \frac{p\hat{\sigma}^2}{\hat{\alpha}'Z'Z\hat{\alpha}} \equiv \frac{p\hat{\sigma}^2}{\hat{\beta}'X'X\hat{\beta}} = \frac{\hat{\sigma}^2}{\hat{c}^2} = 0.0757 .$$

(3) Fixed Value Estimator. k = 0.1

A ridge regression computer program written by David Fawcett in 1973 calculated the ridge coefficients for the three estimators above. The results of the ridge

regressions are shown in Table 12. For the sake of comparison, the OLS results are also shown.

Regression Results

Using OLS, none of the estimated coefficients were statistically significant. Moreover, the dollar cost coefficient had the wrong sign. Both results are symptomatic of the severe multicollinearity present in the highly aggregated data.

In contrast, the ridge estimators were effective in coping with the multicollinearity since all three produced lower estimated MSE($\hat{\beta}$*) than OLS. Naturally, the Lawless-Wang estimator had the lowest estimated MSE($\hat{\beta}$*) because it minimizes the estimated mean square error as shown earlier in Chapter IV. The MSE($\hat{\beta}$*) of the fixed($k = 0.1$) value estimator was very close to the Lawless-Wang value because, merely by chance, $k = 0.1$ happened to be close to the Lawless-Wang k-value of 0.0757. As a result, the coefficients estimated by $k = 0.1$ were also close to the Lawless-Wang estimated coefficients. In both cases, the dollar cost coefficient had the expected sign, an important improvement over the OLS results. Moreover, the variances of the estimates were greatly reduced by both ridge estimators as compared to OLS.

Of the three ridge estimators, RIDGM was the least effective. Not only did RIDGM have the largest estimated

Table 12. Estimated standardized coefficients, estimated variances, and estimated MSE of the ridge estimator for k = 0.0 (OLS), for k = 0.1, and for the RIDGM and Lawless-Wang ridge estimators for the four-explanatory variable travel cost model.

	OLS	Ridge Regression		
	k = 0	RIDGM	Lawless-Wang	k = 0.1
$\hat{\beta}_1^*$	1.05200	0.26392	-0.22947	-0.26856
$\hat{v}(\hat{\beta}_1^*)$	1.42457	0.16363	0.01583	0.01121
$\widehat{MSE}(\hat{\beta}_1^*)$	1.42457	0.20899	0.11147	0.11212
$\hat{\beta}_2^*$	-1.97270	-1.19656	-0.67869	-0.62871
$\hat{v}(\hat{\beta}_2^*)$	1.29534	0.15634	0.01592	0.01144
$\widehat{MSE}(\hat{\beta}_2^*)$	1.29534	0.19790	0.10544	0.10607
$\hat{\beta}_3^*$	0.62573	0.51251	0.23956	0.21094
$\hat{v}(\hat{\beta}_3^*)$	1.89172	0.14831	0.01296	0.00941
$\widehat{MSE}(\hat{\beta}_3^*)$	1.89172	0.20515	0.11404	0.11454
$\hat{\beta}_4^*$	-0.63334	-0.38514	-0.03667	-0.00489
$\hat{v}(\hat{\beta}_4^*)$	1.55806	0.15638	0.01484	0.01063
$\widehat{MSE}(\hat{\beta}_4^*)$	1.55806	0.20487	0.11088	0.11147
$\widehat{MSE}(\hat{\beta}^*)$	6.16969	0.81692	0.44183	0.44421

MSE($\hat{\beta}$*), but the sign of the dollar cost coefficient was not changed from OLS. No doubt RIDGM was less effective than Lawless-Wang because the $\hat{\alpha}$ parameter estimates were not weighted by the corresponding eigenvalues. As a result, the RIDGM value of k was probably much lower than the optimal value.

Summary

Both methods of biased linear estimation compared favorably with OLS. Of the principal components estimates, the proposed loss-function-related criterion and two of the t-criteria produced the same result. Of the ridge estimators, the Lawless-Wang estimator produced the greatest reduction in estimated MSE($\hat{\beta}$*).

Both principal components analysis and ridge regression achieved remarkable improvement in estimated mean square error, almost a 93% improvement over OLS. However, there were some differences between the regression results of the two methods of estimation. First of all, the Lawless-Wang value of estimated MSE($\hat{\beta}$*) was slightly lower than the value for principal components analysis (0.4418 versus 0.4917). Furthermore, although for both methods the cost and distance coefficients had the expected signs, the income squared coefficient did not have the expected sign under principal components analysis. It seems that the Lawless-Wang ridge estimator has produced slightly

more accurate and reasonable results than principal components analysis for this particular empirical problem.

What the Lawless-Wang ridge estimator achieved in this case is quite remarkable: low estimated MSE; variances reduced to a fraction of the OLS variances; and all coefficients with the expected signs. In this case, ridge regression appears to offer a viable alternative to the full model estimated by OLS.

While we can greatly appreciate the improvements made by ridge regression in this situation, we should not overlook two important points. The first is that specification of the model is extremely important. Compare the OLS regressions in (6.3) and (6.5). In the Brown, Singh, and Castle specification with days rather than trips as the dependent variable, the travel cost coefficients had the expected signs. The second point is that aggregation of the data is a major cause of multicollinearity. The example in this chapter illustrates all too well the consequences of aggregation.

VII. SUMMARY AND CONCLUSIONS

In the presence of severe multicollinearity, the OLS estimates will be unreliable and imprecise due to greatly inflated variances. As a result, individual coefficients may be statistically insignificant and may take the wrong signs even though the overall regression may be highly significant. Since multicollinearity is a problem of the data set related to the underlying experimental design, the first step should be to attempt solutions oriented to the data itself, such as including additional observations or disaggregating the data. Of course, another approach to the problem is to focus on the model specification. Often, redefinition of the variables or the use of a priori information about the values of the parameters can improve the OLS estimation. When such solutions are not viable, it has been common practice to delete variables in order to reduce the degree of multicollinearity. However, the more relevant and intercorrelated the deleted variable, the greater the specification bias that is incurred. When the specification bias is large, variable deletion may be a poor remedy for multicollinearity. A more attractive alternative in such cases may be the use of a biased linear estimator, if the added bias is more than offset by reduced variance.

Principal components analysis is a method of biased linear estimation which is less extreme than variable deletion since only components of variables rather than whole variables are deleted. Essential to the effectiveness of principal components analysis is the choice of the method used to delete components. In the Monte Carlo experiments, the traditional t-criteria were compared with a proposed loss-function-related criterion. Both types of criteria performed better than OLS under conditions of high multicollinearity at low and moderate information-to-noise ratios. In general, however, the proposed loss-function-related criterion produced more accurate results, in terms of true mean square error, than did the t-criteria. No doubt the t-criteria were less effective than the proposed criterion because they do not take into account the effect on mean square error of deleting components.

Ridge regression is another method of biased linear estimation. Like principal components analysis, ridge regression was found to be more effective than OLS in the Monte Carlo experiments of Chapter V, under conditions of high multicollinearity at low and moderate information-to-noise ratios. Three ridge estimators were compared in the Monte Carlo experiments: RIDGM; Lawless-Wang; and $k = 0.1$. It was shown in Chapter IV that the Lawless-Wang estimator minimizes the estimated mean square error. Thus, it is not surprising that of the three ridge estimators, Lawless-Wang

produced the most accurate results in terms of true mean square error. A comparison in the Monte Carlo experiments, of the Lawless-Wang estimator with the proposed loss-function-related criterion for deleting principal components, indicated the Lawless-Wang estimator to be slightly more accurate in terms of true mean square error. Moreover, the Lawless-Wang estimator also appeared to be more accurate in the empirical problem, considering the a priori information about the coefficients.

Perhaps the most serious limitation in the use of biased linear estimators in empirical problems has been the inability to evaluate the reliability of the estimates. Of course, a priori information offers some basis for judging the quality of the estimates. However, such information is not always available. In any case, an estimate of mean square error may sometimes be more useful and objective.

The method suggested by Brown for estimating the mean square error of ridge estimates appears to be helpful in the use of ridge regression and other biased estimators. The Monte Carlo results indicate that the suggested method can produce good estimates under conditions of high multicollinearity at low and moderate information-to-noise ratios.

Limitations and Additional Research Needed

Hindsight reveals a few things that could have been done differently in the Monte Carlo experiments. First of all, an average mean square error of $\hat{c}^2$ could have been recorded. This would have provided an indication of the accuracy of $\hat{c}^2$ as an estimate of the true $\hat{\alpha}_i^2$. (Of course, favorable results by the suggested method for estimating MSE($\hat{\alpha}^*$) and by the proposed loss-function-related criterion for deleting principal components suggest that $\hat{c}^2$ is a good estimate under conditions of high multicollinearity at low and moderate information-to-noise ratios.) Second, it would have been interesting to have recorded the frequency distributions of the true α_i parameters generated in the Monte Carlo experiments. (Assumptions about the α_i distributions form the basis for both the proposed loss-function-related criterion for deleting principal components and the suggested method for estimating the mean square error of ridge estimators.) Third, the possibility of using a random number generator which better approximates the normal distribution should be considered in future Monte Carlo experiments.

Certainly, it would have been interesting to have run a larger number of experiments per information-to-noise ratio and to have considered models with more than four explanatory variables and thirteen observations. However,

there is reason to believe that incorporating these features would not have changed the conclusions drawn from the Monte Carlo experiments. This statement is based on the results of a Monte Carlo study done by Lawless and Wang (1976) in which 5000 experiments were run per information-to-noise ration on 4 models (one model with 17 explanatory variables and 50 observations). Among other estimators, their study included two principal components estimators and the Lawless-Wang ridge estimator. The results of their experiments agreed with the conclusions in this paper about the principal components and ridge estimators.

A direction for further research suggested in Chapter V, is a Monte Carlo study of the effectiveness of the application of the proposed loss-function-related criterion beyond the deletion of principal components to the deletion of whole variables. Due to limited time, such a study was not attempted in this paper. Also, of great value, would be further research to evaluate the performance of biased linear estimators when other econometric problems, such as measurement error or heteroscedasticity, exist along with multicollinearity in the data set. Monte Carlo experiments would be one approach to such a study. Of course, further research is also needed in the application of biased linear estimation methods to empirical problems and in the interpretation of the regression results.

BIBLIOGRAPHY

Brown, W. G. "Economic Implications of Allocations." Chapter in Marine Recreational Fisheries, Sport Fishing Institute, Washington, D.C., 1976.

_____. "Effect of Omitting Relevant Variables Versus Use of Ridge Regression in Economic Research." Agricultural Experiment Station, Special Report #394, Oregon State University, Corvallis, 1973.

_____. "Estimation of Mean Square Error of Ridge Regression Coefficients from the Sample Data." Submitted for publication, March, 1978.

Brown, W.G., Farid Nawas, and Joe B. Stevens. "The Oregon Big Game Resources: An Economic Evaluation." Oregon Agricultural Experiment Station Special Report #379, Corvallis, 1973.

Brown, W. G., Ajmer Singh, and Emery Castle. "An Economic Evaluation of the Oregon Salmon and Steelhead Sport Fishery." Oregon Agricultural Experiment Station Technical Bulletin #78, Corvallis, 1964.

Clawson, Marion and Jack Knetsch. Economics of Outdoor Recreation, John Hopkins University, Baltimore, Md., 1966.

Dempster, A. P., M. Schatzoff and N. Wermuth. "A simulation Study of Alternatives to Ordinary Least Squares." Journal of the American Statistical Association, Vol. 72, pp. 78-81, 1977.

Dwyer, John F., John R. Kelly, and Michael D. Bowes. Improved Procedures for Valuation of the Contribution of Recreation to National Economic Development, Research Report No. 128. University of Illinois Water Resources Center, Urbana-Champaign, 1977.

Farrar, D. E., and R. R. Glauber. "Multicollinearity in Regression Analysis: The Problem Revisited." Review of Economics and Statistics, Vol. 49, pp. 92-107, 1967.

Hadley, G. Linear Algebra, Addison-Wesley, Reading, Mass., 1961.

Hald, A. Statistical Theory with Engineering Applications, John Wiley and Sons, New York, 1952.

Henderson, J. M. and R. E. Quandt. Microeconomic Theory, 2nd Ed. McGraw-Hill Book Co., Inc., New York, 1971.

Hoerl, A. E., and R. W. Kennard. "Ridge Regression: Biased Estimation for Non-Orthogonal Problems." Technometrics, Vol. 12, No. 1, pp. 55-67, 1970.

Johnston, J. Econometric Methods, 2nd Ed. McGraw-Hill Book Co., Inc., New York, 1972.

Lawless, J. F. and P. Wang. "A Simulation Study of Ridge and Other Regression Estimators." Communications in Statistics, A5(4), 307-323, 1976.

McCallum, B. T. "Artificial Orthogonalization in Regression Analysis." Review of Economic Statistics, 52, 110-113, 1970.

Neter, John and William Wasserman. Applied Linear Statistical Models, Richard D. Irwin, Inc., Homewood, Illinois, 1974.

Newman, Thomas G. and Patrick L. Odell. The Generation of Random Variates, Hafner Publishing Co., New York, 1971.

Rahuma, Ali Ahmed. Application and Evaluation of Ridge Regression to Selected Empirical Economic Models, M.S. Thesis, Oregon State University, Corvallis, 1978.

Schmidt, Peter. Econometrics, Marcel Dekker, Inc., New York, 1976.

APPENDICES

APPENDIX A

Appendix Table 1. The original Hald data or Data Set 4 in the Monte Carlo experiments.

Observation Number	X_1	X_2	X_3	X_4	Y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

Appendix Table 2. The orthogonal data or Data Set 1 in the Monte Carlo Experiments.

Observation Number	X_1	X_2	X_3	X_4
1	-0.5	0	0	0
2	0	0.5	0	0
3	0	0	$-1/\sqrt{1.5}$	0
4	0	0	0	$-1/\sqrt{2}$
5	-0.5	0	0	0
6	0	0.5	0	0
7	0	0	$1/\sqrt{6}$	0
8	0	0	0	$1/\sqrt{2}$
9	0.5	0	0	0
10	0	-0.5	0	0
11	0	0	$1/\sqrt{6}$	0
12	0	-0.5	0	0
13	0.5	0	0	0

Data Sets 2 and 3

Data Set 2 was created from the original Hald data by changing observations 12 and 13 of variable X_2 from 66 and 68 to 56 and 53, respectively. Data Set 3 was created from the original data by changing observation 11 of variable X_1 from 1 to 7.

APPENDIX B

Appendix Table 3. The highly aggregated set of data from the salmon and steelhead survey data published by Brown, Singh, and Castle.

Y_j	X_{1j}	X_{2k}	X_{3j}	X_{3j}^2
0.001303	7.623	38	59.0	3481.0
0.000686	11.902	104	41.8	1747.2
0.000852	15.059	104	92.2	8500.8
0.000505	15.853	140	43.8	1918.4
0.000605	16.617	140	102.8	10567.8
0.000310	24.250	221	69.0	4761.0
0.000690	14.377	120	53.6	2873.0
0.000833	16.176	120	121.3	14713.7

The data set above was created by averaging the following observations of the original data published by Brown, Singh, and Castle:

Original Data Set Observations	Observation in the Highly Aggregated Data Set
1- 5	1
6-10	2
11-15	3
16-20	4
21-24	5
25-27	6
28-32	7
33-35	8

Selected Statistics for the Highly Aggregated Set of Salmon and Steelhead Data

The Standard Deviations:

Variable	Standard Deviation
Y_j	2.92707E-04
X_{1j}	4.69454E 00
X_{2k}	5.09143E 01
X_{3j}	2.92867E 01
X_{3j}^2	4.71187E 03

The Error Term Variance: 0.0179151

The Correlation Matrix:

	Y_j	X_{1j}	X_{2k}	X_{3j}	X_{3j}^2
Y_j	1.0	-0.8532341	-0.9345133	0.0984555	0.0967206
X_{1j}	-0.8532341	1.0	0.9747886	0.2663543	0.2351646
X_{2k}	-0.9345133	0.9747886	1.0	0.0909855	0.0698352
X_{3j}	0.0984555	0.2663543	0.0909855	1.0	0.9915443
X_{3j}^2	0.0967206	0.2351646	0.0698352	0.9915433	1.0

The Diagonal Elements of the Inverted X'X Correlation Matrix:

r^{11}	79.5179	r^{33}	105.5940
r^{22}	72.3045	r^{44}	86.9688

The Eigenvalues of the X'X Correlation Matrix:

λ_1	2.3211277	λ_3	0.01460489
λ_2	1.6606295	λ_4	0.0036379318