A COMPARISON OF THREE PREDICTION BASED METHODS OF CHOOSING THE RIDGE REGRESSION PARAMETER K

by

Philip L. Gatz, Jr.

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
Master of Science
in
Statistics

APPROVED:

Raymond H. Myers, Chairman	Marion R. Reynolds
Robert S. Schulman	Gary J. Ulrich

July 22, 1985

Blacksburg, Virginia

A COMPARISON OF THREE PREDICTION BASED METHODS OF CHOOSING THE RIDGE REGRESSION PARAMETER K

by

Philip L. Gatz, Jr.

Raymond H. Myers, Chairman

Statistics

(ABSTRACT)

A solution to the regression model $\underline{y} = x\underline{\beta}+\underline{\varepsilon}$ is usually obtained using ordinary least squares. However, when the condition of multicollinearity exists among the regressor variables, then many qualities of this solution deterioriate. The qualities include the variances, the length, the stability, and the prediction capabilities of the solution.

An analysis called ridge regression introduced a solution to combat this deterioration (Hoerl and Kennard, 1970a). The method uses a solution biased by a parameter k. Many methods have been developed to determine an optimal value of k. This study chose to investigate three little-used methods of determining k: the PRESS statistic, Mallows' C_k statistic, and DF-trace. The study compared the prediction capabilities of the three methods using data that contained various levels of both collinearity and leverage. This was completed by using a Monte Carlo experiment.

ACKNOWLEDGEMENTS

This thesis is dedicated to my loving wife , whose patience, toil, and support has made all this possible.

The author is deeply indebted to Dr. Raymond Myers for his time, ideas, and encouragement given to make this project a success.

The author also wishes to acknowledge the prayerful guidance of his mentor, .

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

1.0 INTRODUCTION

Multiple linear regression is a very popular statistical tool. The analysis involves the investigation and modeling of the relationships between variables and some type of response. The choice of a set of variables will depend heavily on the experiment and experimenter. However, the use of two variables that produce the same information in the problem will render the analysis useless. An example is an experiment monitoring the human heart's response to a new drug that uses as separate variables the amount of drug administered per hour and the amount of drug administered per day. This problem is easily solved by deleting one of the duplicate variables. A more important, but subtler, problem occurs when two, three or several variables are bringing similar, yet not exactly the same, information into the analysis. This characteristic among the regressor variables is called multicollinearity.

For a more detailed look into the causes of and problems caused by multicollinearity, consider the standard multiple linear regression model

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon},$$

where $\underline{y}$ is an nx1 response vector, X is an nxp data matrix, $\underline{\beta}$ is a px1 vector of model coefficients and ε is an nx1 vector of random disturbances. Also, it is assumed that $E(\underline{\varepsilon}) = \underline{0}$

and $E(\underline{\varepsilon}\underline{\varepsilon}') = \sigma^2 I_n$. From the Gauss-Markov theorem a linearly unbiased estimator of $\underline{\beta}$ can be found. This estimator is given by:

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

This procedure produces the best results, in terms of estimation and prediction, when, after appropriate centering and scaling of the columns of the data matrix, X'X is nearly an identity matrix. However, when a near-linear dependency exists among the columns of X, i.e., multicollinearity, the estimation and prediction capabilities of $\hat{\underline{\beta}}$ deteriorate.

When the data matrix contains near dependencies, it can be shown that some biased estimators of β have a smaller mean square error than the Gauss-Markov estimate. One of the first of these estimators, introduced by Hoerl and Kennard (1970a), is called the ridge estimator and is given by

$$\underline{\beta}_R = (X'X + kI)^{-1}X'\underline{y},$$

where $k \geq 0$. Since its introduction, numerous methods have been developed to choose an optimal value of k. The bases for these methods can be categorized into two groups: prediction oriented and coefficient estimation oriented. Examples in the latter group would be the Ridge Trace (Hoerl and Kennard, 1970a), the iterative method using the generalized ridge regression procedure (Hoerl and Kennard, 1970a), the harmonic mean method (Hoerl, Kennard and Baldwin, 1975), and the iterative harmonic mean method (Hoerl and Kennard, 1976). These procedures have been considered in numerous papers and

Monte Carlo studies. The purpose of this paper is to review the lesser-known, prediction-based techniques: the PRESS statistic, the C_k statistic and the DF-trace procedure, and compare their prediction capabilities with a limited Monte Carlo study.

2.0 EFFECTS OF MULTICOLLINEARITY ON THE PROPERTIES OF $\hat{\beta}$

Multicollinearity can be described as an ill-conditioning in the data matrix or the existence of near-linear dependencies among the columns of the X matrix. In more technical terms, multicollinearity is said to exist if there exists a set of constants $(a_1, a_2, \ldots, a_p)$ such that:

$$\sum_{j=1}^{p} a_j \underline{x}_j \approx \underline{0},$$

where $\underline{x}_j$ is the jth column of X. An eigenvalue decomposition can be used to explicitly determine the values of the constants $(a_1, a_2, \ldots, a_p)$. The eigenvalue decomposition of X'X is defined as:

$$V'(X'X)V = \Lambda = \text{diag}(\lambda_i),$$

where the λ_i's are the eigenvalues of X'X and $V = [\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_p]$ is an orthogonal matrix of eigenvectors. Multicollinearity is now defined as the presence of any eigenvector that produces $X\underline{v}_j \approx \underline{0}$. This dot product can be rewritten as $\sum_{i=1}^{p} \underline{x}_i v_{ji} \approx \underline{0}$, where v_{ji} is the ith entry in the jth eigenvector. Therefore, the set of constants $(a_1, a_2, \ldots, a_p)$ can be better explained as the weights of a particular eigenvector. That eigenvector will be one that corresponds to an $\lambda_i \approx 0$.

2.1 EFFECT OF MULTICOLLINEARITY ON THE SUM OF THE VARIANCES OF THE REGRESSION COEFFICIENTS

Multicollinearity can severely affect many properties of the estimator, $\hat{\underline{\beta}}$. The first of these that will be studied is multicollinearity's effect on the sum of the variances of the regression coefficients. The variance of the coefficient, $\hat{\beta}_j$, apart from σ^2, is found on the jth diagonal of $(X'X)^{-1}$. Employing the eigenvalue decomposition of X'X, then

$$(X'X)^{-1} = V\Lambda^{-1}V'$$
$$= V[\text{diag}(1/\lambda_i)]V'.$$

The sum of the variances can then be found by taking the trace of $(X'X)^{-1}$:

$$\sum_{i=1}^{p} \text{var}(\hat{\beta}_j) = \sigma^2 \text{tr}(X'X)^{-1}$$
$$= \sigma^2 \text{tr}(V\Lambda^{-1}V')$$
$$= \sigma^2 \sum_{j=1}^{p} 1/\lambda_j.$$

Thus, when multicollinearity is present, i.e., at least one $\lambda_j \approx 0$, the sum of the variances will be inflated (Montgomery and Peck, 1982).

2.2 EFFECT OF MULTICOLLINEARITY ON LENGTH OF $\hat{\underline{\beta}}$

A second property of $\hat{\underline{\beta}}$ that is damaged by multicollinearity is the expected length or norm squared of $\hat{\underline{\beta}}$. The expected length of $\hat{\underline{\beta}}$ is derived by:

$$
\begin{aligned}
E(\hat{\underline{\beta}}'\hat{\underline{\beta}}) &= E[\underline{y}'X(X'X)^{-1}(X'X)^{-1}X'\underline{y}] \\
&= \sigma^2 tr[X(X'X)^{-1}(X'X)^{-1}X'] + \underline{\beta}'\underline{\beta} \\
&= \sigma^2 tr[(X'X)^{-1}] + \underline{\beta}'\underline{\beta} \\
&= \sigma^2 tr[V\Lambda^{-1}V'] + \underline{\beta}'\underline{\beta} \\
&= \sigma^2 \sum_{i=1}^{p} 1/\lambda_i + \underline{\beta}'\underline{\beta} \ .
\end{aligned}
$$

Therefore, the expected length is positively biased and considerably so, in the presence of multicollinearity.

2.3 EFFECT OF MULTICOLLINEARITY ON STABILITY OF $\hat{\underline{\beta}}$

A third property that is distorted by multicollinearity is the stability of the estimator, $\hat{\underline{\beta}}=(X'X)^{-1}X'\underline{y}$. To see this, $\hat{\underline{\beta}}$ is rewritten using the eigenvalue decomposition:

$$
\begin{aligned}
\hat{\underline{\beta}} &= V\Lambda^{-1}V'X'\underline{y} \\
&= \sum_{j=1}^{p} \underline{v}_j(1/\lambda_j)c_j,
\end{aligned}
$$

where $c_j=\underline{v}_j'X'\underline{y}$, is a constant. This form implies that small changes in c_j, i.e., small perturbations in y, could cause severe changes in $\hat{\underline{\beta}}$ (Myers, 1985). This again would be caused by an $\lambda_j \approx 0$ due to the presence of multicollinearity.

2.4 EFFECT OF MULTICOLLINEARITY ON PREDICTION

A final effect of multicollinearity that needs to be explored is the effect on the prediction capabilities of the

model, i.e., $\hat{\underline{\beta}}$. Consider the variance of prediction of a data point, when the X matrix is centered and scaled:

$$\text{var } \hat{y}(\underline{x}_i) = \underline{x}_i'(X'X)^{-1}\underline{x}_i = h_{ii}.$$

It can be shown that h_{ii}, the ith diagonal of the HAT matrix is bounded between 1/n and 1 in spite of the presence of collinearity (Hoaglin and Welsch, 1978).

However, now consider a point $\underline{x}_0$ which is not necessarily a data point. To show multicollinearity's effect, an orthogonal transformation will be used so that the model will be rewritten as:

$$\begin{aligned} \underline{y} &= X\underline{\beta}+\underline{\varepsilon} \\ &= XVV'\underline{\beta} + \underline{\varepsilon} \\ &= Z\underline{\alpha} + \underline{\varepsilon}, \end{aligned}$$

where $V = [\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_p]$, the matrix of eigenvectors. This transformation implies that

$$\begin{aligned} \text{var } \hat{y}(\underline{x}_0) &= \sigma^2\underline{x}_0'(X'X)^{-1}\underline{x}_0 \\ &= \sigma^2\underline{x}'_0V\Lambda^{-1}V'\underline{x}_0 \\ &= \sigma^2\underline{z}'_0\Lambda^{-1}\underline{z}_0 \\ &= \sum_{i=1}^{p} z_{i,0}^2/\lambda_i . \end{aligned}$$

In the case where $\underline{x}_0$ is a data point and multicollinearity is present, $z_{i,0} = \underline{x}_0'\underline{v}_j$ will be approximately equal to 0 because by definition $X\underline{v}_j \approx \underline{0}$. Yet, when $\underline{x}_0$ is not near a data point, $z_{i,0}$ will not be approximately 0 and the prediction variance can be very large. In other words, if $\lambda_i \approx 0$ and $\underline{x}_0$ is in the mainstream of the collinearity, i.e., the point

is near a data point in location, then $\underline{x}_0$ will be nearly orthogonal to its associated eigenvector $\underline{v}_i$, and $z_{i,0}$ and var $\hat{y}(\underline{x}_0)$ will be small. However, if $\underline{x}_0$ is outside the mainstream of the collinearity, then none of the above holds and the var $\hat{y}(\underline{x}_0)$ will get very large.

3.0 RIDGE REGRESSION

Hoerl and Kennard (1970a) suggest using a least squares estimator with the constraint $\hat{\underline{\beta}}'\hat{\underline{\beta}}=\rho$. This is done because, as seen in the previous chapter, the length of $\hat{\underline{\beta}}$ is unbounded when multicollinearity is present. This constraint is then found by differentiating:

$$L = (\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta})+k(\underline{\beta}'\underline{\beta}-\rho)$$

with respect to $\underline{\beta}$ and equating the derivative to zero. Here k is the LaGrangian multiplier. This results in the following set of normal equations:

$$(X'X+kI)\underline{\beta} = X'\underline{y}$$

This then results in the ridge estimator:

$$\underline{\beta}_R = (X'X+kI)^{-1}X'\underline{y}.$$

It is relatively easy to show that the properties of $\hat{\underline{\beta}}$ that were severely damaged by multicollinearity are now being moderated by k.

3.1 EFFECT OF RIDGE REGRESSION ON SUM OF THE VARIANCES OF THE REGRESSION COEFFICIENTS

The first property that this can be noted in is the sum of the variances of the coefficients. The variance of the coefficients in the ridge regression model can be written as:

$$\text{var } \underline{\beta}_R = (X'X+kI)^{-1}X'X(X'X+kI)^{-1}.$$

The same eigenvalue decomposition utilized in the previous chapter is used to produce:

$$V'(X'X+kI)V = \Lambda_k = \text{diag } (\lambda_i+k),$$

where V is the same orthogonal matrix of eigenvectors. Thus, $(X'X+kI)^{-1}$ can be rewritten as $V\Lambda_k^{-1}V'$ and the sum of the variances as:

$$\begin{aligned}\sum_{i=1}^{p} \text{var } \underline{\beta}_{R,i} &= \sigma^2 tr[V\Lambda_k^{-1}V'V\Lambda V'V\Lambda_k^{-1}V'] \\ &= \sigma^2 tr[V\Lambda_k^{-1}\Lambda\Lambda_k^{-1}V'] \\ &= \sigma^2 \sum_{i=1}^{p} \lambda_i/(\lambda_i+k)^2.\end{aligned}$$

The parameter k moderates the effect of the multicollinearity. Also, note that as $k\to\infty$, the $\sum_{i=1}^{p} \text{var } \beta_{R,i} \to 0$.

3.2 EFFECT OF RIDGE REGRESSION ON LENGTH OF $\hat{\underline{\beta}}$

As noted in the opening paragraph of this chapter, the expected length of the ridge estimate, $\underline{\beta}_R$, is constrained or bounded. This can be shown by utilizing the eigenvalue decomposition as follows:

$$\begin{aligned}E(\underline{\beta}'_R\underline{\beta}_R) &= E[\underline{y}'X(X'X+kI)^{-1}(X'X+kI)^{-1}X'\underline{y}] \\ &= \sigma^2 tr[X(X'X+kI)^{-1}(X'X+kI)^{-1}X'] \\ &\quad +\underline{\beta}'X'X(X'X+kI)^{-1}(X'X+kI)^{-1}X'X\underline{\beta} \\ &= \sigma^2 tr[(X'X+kI)^{-1}X'X(X'X+kI)^{-1}] \\ &\quad +\underline{\beta}'V\Lambda V'V\Lambda_k^{-1}V'V\Lambda_k^{-1}V'V\Lambda V'\underline{\beta}\end{aligned}$$

$$= \sigma^2 tr[V\Lambda_k^{-1}V'V\Lambda V'V\Lambda_k^{-1}V'] + \underline{\alpha}'\Lambda\Lambda_k^{-1}\Lambda_k^{-1}\Lambda\underline{\alpha}$$

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

Again, the parameter k moderates the effect multicollinearity has on the expected length of $\underline{\beta}_R$, such that as $k\to\infty$, the $||\underline{\beta}_R||^2\to 0$. This is the reason that parameter k is sometimes called a shrinkage parameter.

3.3 EFFECT OF RIDGE REGRESSION ON STABILITY OF $\hat{\underline{\beta}}$

The presence of the parameter k will also enhance the stability of the estimator $\underline{\beta}_R$. Using the same notation as previously used, $\underline{\beta}_R$ can be rewritten as:

$$\underline{\beta}_R = (X'X+kI)^{-1}X'\underline{y}$$
$$= V\Lambda_k V'X'\underline{y}$$
$$= \sum_{j=1}^{p} \underline{v}_j[1/(\lambda_j+k)]c_j,$$

where $c_j = \underline{v}_j'X'\underline{y}$ is a constant. The parameter k will now limit the effect a small perturbation in $\underline{y}$ could have on the estimator of the coefficients.

3.4 EFFECT OF RIDGE REGRESSION ON PREDICTION

Finally, the parameter k will also moderate the inflation due to multicollinearity found in the variance of prediction at a point that is outside the mainstram of the collinearity. The prediction variance of such a point in the ridge regression model is given by:

$$\begin{aligned}
\text{var } \hat{y}_R(\underline{x}_0) &= \sigma^2 \underline{x}_0 (X'X+kI)^{-1} X'X (X'X+kI)^{-1} \underline{x}_0 \\
&= \sigma^2 \underline{x}_0 V \Lambda_k^{-1} V' V \Lambda V' V \Lambda_k^{-1} V' \underline{x}_0 \\
&= \underline{z}_0' \Lambda_k^{-1} \Lambda \Lambda_k^{-1} \underline{z}_0 \\
&= \sum_{i=1}^{p} (z_{0,i}^2 \lambda_i)/(\lambda_i+k)^2 .
\end{aligned}$$

Now, even though the point is outside the mainstream of the collinearity, implying that $z_{0,i} \neq 0$, the prediction variance is moderated by the parameter k. Hence, as $k \to \infty$, the var $\hat{y}_R(\underline{x}_0) \to 0$, for every $\underline{x}_0$.

A word of caution is needed concerning the choice of the value for the parameter k. From the previous sections in the chapter, it would appear that a large value would be the most appropriate. However, this ignores the induced bias of the estimator. An investigation into the properties of the sum of the mean sqaure errors of the ridge estimators will show that moderation brought by k will continue only up to a point as k increases. After that point the induced bias will cause

the properties of the ridge estimator to become worse than that of the ordinary least sqaures estimator.

4.0 THE PREDICTION ORIENTED METHODS OF CHOOSING K

As mentioned earlier, there are numerous methods by which the parameter k may be chosen, but the emphasis of this paper is to review only those methods which are prediction based. These are lesser known and sometimes little used methods which were developed specifically to enhance the prediction capabilities of a model laden with multicollinearity.

4.1 MALLOWS' C_K STATISTIC.

The first method to be reviewed is Mallows' C_k statistic (1973). In his paper, Mallows suggested a method, using the C_p statistic, to choose the proper or best subset of regressor variables. It involves minimizing:

$$C_p = \sum_{i=1}^{n} \text{var } \hat{y}(\underline{x}_i) + \sum_{i=1}^{n} \text{bias}^2 \hat{y}(\underline{x}_i),$$

where var $\hat{y}(\underline{x}_i)$ is the prediction variance at a data point and bias $\hat{y}(\underline{x}_i)$ is the bias incurred when the model is under-specified.

He then suggests a similar application for choosing the parameter k by minimizing with respect to k:

$$C_k = 1/\sigma^2[\sum_{i=1}^{n} \text{var } \hat{y}_R(\underline{x}_i) + \sum_{i=1}^{n} \text{bias}^2\hat{y}_R(\underline{x}_i)],$$

where the components are similar to those above, but under the ridge regression model. The computational form is as follows:

$$C_k = 2\text{tr}(H_k) - n + (\text{SSRES}_R)(n-p-1)/(\text{SSRES}_{OLS})$$

where H_k is the HAT-like matrix, $X(X'X+kI)^{-1}X'$, SSRES_R is the residual sum of squares under the ridge regression model and SSRES_{OLS} is the residual sum of squares under the ordinary least squares model. This form is derived as follows (Myers, 1985):

$$1/\sigma^2 \text{var } \hat{y}_R(\underline{x}_i) = \underline{x}_i'(X'X+kI)^{-1}X'X(X'X+kI)^{-1}\underline{x}_i.$$

Therefore,

$$1/\sigma^2 \sum_{i=1}^{n} \text{var } \hat{y}_R(\underline{x}_i) = \text{tr}[X(X'X+kI)^{-1}X'X(X'X+kI)^{-1}X']$$
$$= \text{tr}[H_k]^2$$

Also,

$$1/\sigma^2 \sum_{i=1}^{n} \text{bias}^2\hat{y}_R(\underline{x}_i)$$

$$= 1/\sigma^2 E[X\underline{\beta}-X\underline{\beta}_R]'E[X\underline{\beta}-X\underline{\beta}_R]$$
$$= 1/\sigma^2 (X\underline{\beta}-X(X'X+kI)^{-1}X'X\underline{\beta})'(X\underline{\beta}-X(X'X+kI)^{-1}X'X\underline{\beta})$$
$$= 1/\sigma^2 \underline{\beta}'X'(I-H_k)^2X\underline{\beta}$$

Since $1/\sigma^2 \sum_{i=1}^{n} \text{bias}^2\hat{y}_R(\underline{x}_i)$ contains the unknown parameter vector, $\underline{\beta}$, an unbiased estimator of it needs to be derived. Consider, $\text{SSRES}_R = \underline{y}'(I-H_k)^{-2}\underline{y}$ and its expected value:

$$E(SSRES_R) = \sigma^2 tr[I-H_k]^2 + \underline{\beta}'X'(I-H_k)^2X\underline{\beta}.$$

Therefore, an unbiased estimator of the sum of squared biases is:

$$1/\sigma^2 \sum_{i=1}^{n} bias^2 \hat{y}_R(x_i)$$

$$= 1/\sigma^2[SSRES_R - \sigma^2 tr(I-H_k)^2]$$

$$= 1/\sigma^2 SSRES_R - tr(I-H_k)^2.$$

This implies, then, that

$$C_k = tr[H_k]^2 + 1/\sigma^2 SSRES_R - tr[I-H_k]^2$$

$$= tr\,[H_k]^2 + (SSRES_R)(n-p-1)/(SSRES_{OLS})$$

$$-tr[I_n] + 2tr[H_k] - tr[H_k]^2$$

$$= 2tr[H_k] - n + (SSRES_R)(n-p-1)/(SSRES_{OLS})$$

From this computational form, it can be shown, (Hoerl and Kennard, 1970a), that as k increases $SSRES_R$ will increase and $tr[H_k]$ will decrease. Therefore, if the ridge regression model is appropriate, then the $tr[H_k]$ should decrease more rapidly at the onset than the increase in $SSRES_R$. As shown in Figure 4.1, a graphical interpretation can be easily found by plotting C_k versus k.

It should also be noted that the increase in $SSRES_R$ reflects the induced bias in $\hat{\underline{y}}(\underline{x}_i)$, while the decrease in $tr[H_k]$ reflects the movement toward the effective rank or degrees of freedom of the problem.

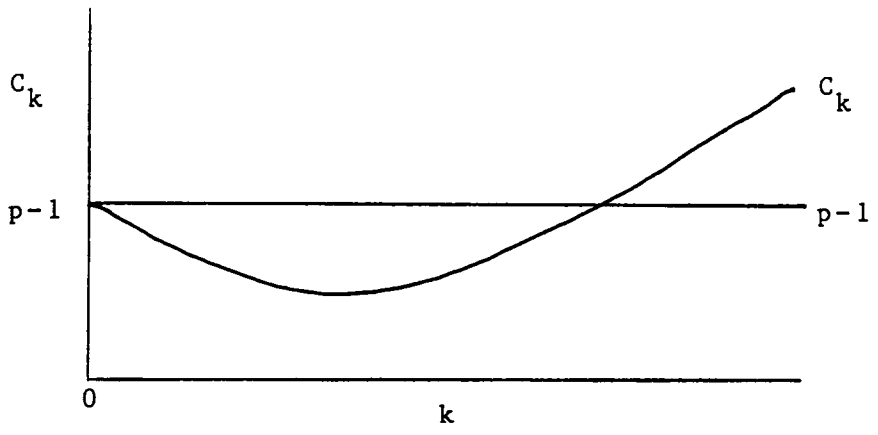

Figure 1. A Plot of C_k versus k: (Note: at k=0 C_k = p-1)

4.2 DF-TRACE METHOD

The DF-trace method (Tripp, 1983) is the most recent of the three methods. It is the only non-stochastic method in this study for choosing the parameter k, i.e., k is truly not considered a random variable. The method is based upon the collinearity structure of X as seen through the "almost-HAT" matrix $X(X'X+kI)^{-1}X'$. This method can be considered a prediction oriented technique because it is essentially the vector of fitted values, $\hat{y}_R$, minus the vector of responses, y.

Under the ordinary least squares regression model, the vector of fitted values is:

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

The matrix $X(X'X)^{-1}X'$ or HAT-matrix is considered a projection matrix of fitted responses. The DF-trace method will attempt to choose the value of k that will enable the almost-HAT matrix, $X(X'X+kI)^{-1}X'$, to best mimic the HAT matrix. The function chosen to monitor this activity is the trace. Thus, when a centered and scaled data matrix, X, is used, then

$$DF = tr[H_k] = tr[X(X'X+kI)^{-1}X'] = \sum_{i=1}^{p} \lambda_i/(\lambda_i+k),$$

where the λ_i's are the eigenvalues of X'X, the correlation matrix.

To determine the appropriate value of k, the values of DF are plotted against k. This is one of the most appealing features of this method as the deflation of DF is clearly seen. The initial value of DF when k=0 will be p. The values of DF will then decrease and eventually stabilize; the steepness of the decline will depend on the severity of the collinearity. The value of k should then be chosen from a range where the rapid decline of the graph of DF diminishes, i.e., somewhere before the stabilization of DF. In Figure 4.2, the range from which k could be chosen would be from .0004 to .0009.

Outside the graphical realm of interpretation, there is another recommendation on choosing an appropriate value of k (Tripp, 1983). This recommendation is in terms of a lower bound for the choice of k. The choice of k should be no smaller than the largest "small" eigenvalue. When multicollinearity is present, the number of small eigenvalues, i.e., $\lambda_i \approx 0$, reflects the number of near-linear dependencies among the data. Therefore, as Tripp states, a choice of k so that it is at least as large as the largest "small" eigenvalue will bring deflation.

Finally, it should be noted that the deflation of DF also represents the movement toward the effective rank of the problem or data set. The rank of the data matrix X is p, the number of independent variables. If a direct dependency between the columns or rows existed, then the rank of X would

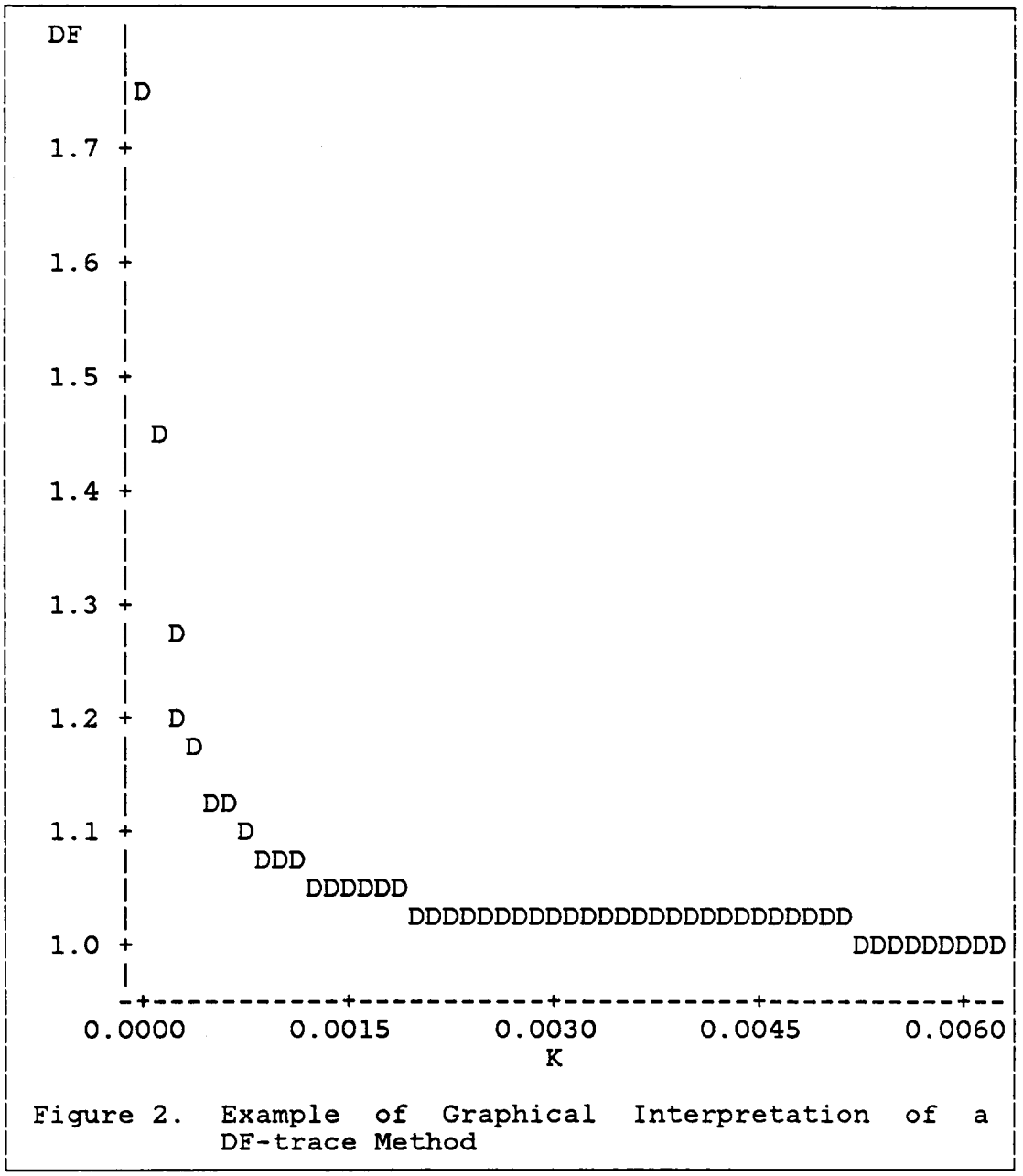

Figure 2. Example of Graphical Interpretation of a DF-trace Method

be less than p. However, when multicollinearity is present, there are near dependencies. Thus, the stabilization of DF represents the movement toward the effective rank of the problem.

4.3 PRESS STATISTIC

The final method of choosing k to be reviewed utilizes the PRESS statistic. In ordinary least squares regression, a residual or fitting error is defined as:

$$e_i = y_i - \hat{y}(\underline{x}_i),$$

the difference between the observed value and the fitted value. Since y_i and $\hat{y}(\underline{x}_i)$ are not independent, the value of attempting to develop a statistical test involving these is limited. A way of alleviating this problem is by "setting aside" the ith data point and then estimating the coefficients using only n-1 observations. The deleted response is then estimated using these estimates of the coefficients. The PRESS residual can then be calculated by:

$$e_{i,-i} = y_i - \hat{y}_{i,-i}.$$

Since y_i and $\hat{y}_{i,-i}$ are independent, the PRESS residuals can be used in a validation criteria for a model. Allen (1971b) proposed such a method by summing the squares of the n PRESS residuals, each having been calculated by the "setting aside"

of the individual point and reestimating the coefficients. The result is the PRESS statistic:

$$\text{PRESS} = \sum_{i=1}^{n} e_{i,-i}^{2} = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2.$$

It should be noted that the calculation of each PRESS residual does not require the "setting aside" of each data point in the regression. Through the use of the Sherman-Morrison-Woodbury Theorem (Rao, 1967), it can be shown that:

$$e_{i,-i} = e_i/1-h_{ii},$$

where h_{ii} is the ith diagonal of $X(X'X)^{-1}X'$.

The same criteria described above can now be used in selecting the appropriate value of the parameter k in the ridge regression model. The ith data point is now "set aside" and coefficients reestimated to form:

$$\hat{y}_{i,-i,k} = \underline{x}_i' \underline{\beta}_{R,-i},$$

which permits the calculation of the PRESS residual,

$$e_{i,-i,k} = y_i - \hat{y}_{i,-i,k}.$$

Thus, the appropriate value of k will be the one that minimizes:

$$\text{PRESS}_k = \sum_{i=1}^{n} e_{i,-i,k}^{2}$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i,k})^2,$$

the sum of the squared PRESS residuals. It is trivial to see that when k=0 the $PRESS_k$ statistic will equal the PRESS statistic.

The time saving feature of calculating the PRESS residual in the ordinary lease squares model, i.e. that $e_{i-1} = e_i/i-h_{ii}$, no longer holds in the ridge regression model. The relationship that does hold is:

$$e_{i,-1,k} \approx e_{i,k}/1-h_{ii,k}$$

where $e_{i,k} = y_i - \underline{x}_i' \underline{\beta}_R$ and $h_{ii,k}$ is the ith diagonal of the almost-HAT matrix, $X(X'X+kI)^{-1}X'$. Therefore, the calculation of the $PRESS_k$ statistic using this form of the PRESS residual will result in an approximate solution. This approximation is due to use of centered and scaled data. The solution would not be approximate if the same centering and scaling constant were used in generating each of the PRESS residuals, but this is rarely the case. Therefore, the calculation of the exact $PRESS_k$ statistic will require the actual "setting aside" of each data point. Though this seems to be more time consuming, the actual computer time is relatively close to that of not "setting aside" each data point for most data sets.

5.0 MULTICOLLINEARITY AND LEVERAGE DIAGNOSTICS

In Chapter 2.0, it was shown that many properties of $\hat{\beta}$ deteriorate rapidly in the presence of multicollinearity. Fortunately, there are many diagnostic tools available for the detection of multicollinearity.

5.1 VARIANCE INFLATION FACTORS

The first of these diagnostics are the variance inflation factors (VIFs). These represent the growth of $var(\hat{\beta})$ above the ideal. It is known that:

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

Also, if the X matrix is centered and scaled, i.e.

$$x_{ij}^{*} = (x_{ij}-\bar{x}_i)/\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_i)^2} ,$$

then $X^{*}'X^{*}$ will be the correlation matrix. If the variables are orthogonal to each other, then the correlation matrix is a pxp identity matrix and its inverse the same. Thus, under these ideal conditions, $var(\hat{\beta}_j)/\sigma^2=1$. The VIFs are a measure of inflation above this ideal.

The VIF for the ith regression coefficient can also be written (Marquardt, 1970):

$$VIF_i = 1/1-R_i^2 ,$$

where R_i^2 is the coefficient of determination found by regressing $\underline{x}_i$ against all other regressor variables. If R_i^2 is large, i.e., near 1, this then indicates a strong linear association between $\underline{x}_i$ and the remaining regressor variables. This will also result in a large VIF.

Though it cannot be specifically determined what value of the VIFs define multicollinearity, a very conservative rule of thumb can be stated. If a VIF exceeds 10, i.e., $R_i^2>.9$, then there is reason to believe there is some ill-conditioning in the data matrix (Montgomery and Peck, 1982).

5.2 EIGENVALUES

A second set of diagnostic tools for detecting multicollinearity are the eigenvalues and associated eigenvectors of the correlation matrix, $X^{*\prime}X^*$. The eigenvalues are easily calculated by the eigenvalue decomposition noted earlier:

$$V'(X^{*\prime}X^*)V = diag(\lambda_1, \lambda_2, \ldots, \lambda_p) ,$$

where the λ_i's are the eigenvalues of $X^{*\prime}X^*$ and V is the orthogonal matrix of eigenvectors. A small eigenvalue, i.e. $\lambda_i \approx 0$, denotes the presence of collinearity. Also, the number of small eigenvalues will denote the number of near-dependencies in the data. An adequate rule for how small an

eigenvalue should be for collinearity to be a problem is $\lambda_i < .01$ (Myers, 1985).

5.3 CONDITION INDICES OF X

A method that also utilizes the eigenvalues of X'X are the condition indices of X (Belsley, Kuh, and Welsch, 1980). The method relies on the singular value decomposition of X (Graybill, 1976):

$$U'XV = D = \text{diag}(\mu_1, \mu_2, \ldots, \mu_p),$$

where U is a matrix of eigenvectors corresponding to the nonzero eigenvalues of XX', V is the orthogonal matrix of eigenvectors of X'X, and μ_i is a singular value. It can be shown that X = UDV' and that X'X = VΛV', thus implying that the singular values of X are the square roots of the eigenvalues of X'X. The condition indices of X are defined as:

$$\eta_j = \mu_{max}/\mu_j = \sqrt{\lambda_{max}}/\sqrt{\lambda_j},$$

for j=1, 2, . . . , p. The extent of the ill-conditioning caused by the multicollinearity will depend on how small an eigenvalue is relative to the largest eigenvalue. A rule of thumb of $\eta_j > 30$ is appropriate for the diagnosing of collinearity.

5.4 VARIANCE PROPORTIONS

The last diagnostic tool dealing with multicollinearity to be evaluated are the variance proportions. A variance proportion is designed to indicate what portion of the variance of each regression coefficient is attributed to each eigenvalue of X'X. To develop this, the eigenvalue decomposition is used on scaled data only, implying:

$$V'(X'X)V = \text{diag}(\lambda_0, \lambda_1, \ldots, \lambda_k),$$

where p=k+1. Also, recall that $\text{var}(\hat{\beta})/\sigma^2=(X'X)^{-1}$, which can be rewritten as:

$$(X'X)^{-1} = V[\text{diag}(\lambda_0, \lambda_1, \ldots, \lambda_k)]^{-1}V'.$$

This implies that the $\text{var}(\hat{\beta}_j)/\sigma^2$, which equals the jth diagonal of $(X'X)^{-1}$, can also be rewritten as:

$$c_{jj} = \sum_{j=1}^{k} v_{ji}^2/\lambda_j,$$

where v_{ji} is the ith element of the jth eigenvector and λ_j is the jth eigenvalue. Therefore, the proportion of the variance of $\hat{\beta}_j$ which can be attributed to λ_i can be defined as:

$$p_{ij} = [v_{ji}^2/\lambda_j]/c_{jj} .$$

The proportions will indicate which variables are involved in the near-linear dependency. If a small eigenvalue, i.e., $\lambda_j<.01$, is accompanied by two or more variables with high

variance proportions, i.e. $p_{ij}>.5$, then those variables are involved in the near-dependency.

5.5 HAT DIAGONAL

Another characteristic, besides multicollinearity, that is important to diagnose in a data set is leverage. Leverage can be described as a condition in which a single observation is "extreme" in the x-direction. This means that it is a large distance away from the center of the data in the x's, even though the point is a very "legitimate" observation. The diagnosis of such a point is important because it could potentially exert undue influence on the estimation of one or more regression coefficients. Several such points in a data set can mask the fact that the data contains severe collinearity.

The diagnostic tool used to detect high leverage points is the HAT diagonal:

$$h_{ii} = \underline{x}_i'(X'X)^{-1}\underline{x}_i \ .$$

The HAT diagonal is a standardized distance measure from $\underline{x}_i$ to $\bar{\underline{x}}$, the centroid in the x's. It can easily be shown (Hoaglin and Welsch, 1978) that $\sum_{i=1}^{n} h_{ii}=p$, the number of model parameters. This implies that p/n would be an average h_{ii}. A rule of thumb for detecting a point that has potential of exerting undue influence would be $h_{ii}>2p/n$ (Belsley, Kuh, and Welsch, 1980).

6.0 EXAMPLE OF COLLINEARITY AND LEVERAGE DIAGNOSTICS

The following example is given to highlight the use of the collinearity and leverage diagnostics. The data in Table 1 reflects information taken from seventeen U.S. Naval hospitals in various parts of the world (Data/Regression Analysis Handbook, 1979). The regressors are workload variables, meaning items that result in the need for manpower in a hospital installation. The variables are described as follows:

y - monthly manhours

x_1 - average daily patient load

x_2 - monthly X-ray exposures

x_3 - monthly occupied bed days

x_4 - eligible population in the area ÷ 1000

x_5 - average length of patient stay, in days .

The goal of the project is to produce an equation that will predict manpower needs for naval hospitals.

The first collinearity diagnostics calculated are the variance inflation factors (VIFs):

$$VIF_1 = 9597.570$$

$$VIF_2 = 7.940$$

$$VIF_3 = 8933.086$$

$$VIF_4 = 23.293$$

Table 1 (Part 1 of 1). Naval Hospital Data

X_1	X_2	X_3	X_4	X_5	Y
15.57	2463	472.92	18.0	4.45	566.52
44.02	2048	1339.75	9.5	6.92	696.82
20.42	3940	620.25	12.8	4.28	1033.15
18.74	6505	568.33	36.7	3.90	1603.62
49.20	5723	1497.60	35.7	5.50	1611.37
44.92	11520	1365.83	24.0	4.60	1613.27
55.48	5779	1687.00	43.3	5.62	1854.17
59.28	5969	1639.92	46.7	5.15	2160.55
94.39	8461	2872.33	78.7	6.18	2305.58
128.02	20106	3655.08	180.5	6.15	3503.93
96.00	13313	2912.00	60.9	5.88	3571.89
131.42	10771	3921.00	103.7	4.88	3741.40
127.21	15543	3865.67	126.8	5.50	4026.52
252.90	36194	7684.10	157.7	7.00	10343.81
409.20	34703	12446.33	169.4	10.78	11732.17
463.70	39204	14098.40	331.4	7.05	15414.94
510.22	86533	15524.00	371.6	6.35	18854.45

$$VIF_5 = \quad 4.279$$

It is clear that at least two of the coefficients, β_1 and β_3 are being poorly estimated as their corresponding VIFs grossly exceed 10. Also, there are two small eigenvalues, $\lambda_5 = .008215$ and $\lambda_6 = .000024$. Thus, there are two near-dependencies among the regressors. According to the variance proportions corresponding to these eigenvalues found in Table 6.2, the near dependencies are between x_1 and x_3 and the intercept and x_5. Collinearity between a variable and intercept is possible, but it is less meaningful if the natural origin is outside the experimental region of the problem. This is the case in this problem. Finally, the condition index μ_6, whose value of 427.326 well exceeds the rule of thumb of 30, points to the severity of the collinearity between x_1 and x_3.

The points that can potentially exert undue influence will be those whose HAT diagonals exceed $2p/n = .7058$. This then includes the following points:

$$h_{10,10} = .8308$$
$$h_{15,15} = .7989$$
$$h_{16,16} = .8321$$
$$h_{17,17} = .8731.$$

Each of these diagnostics need to be given careful attention when building a prediction model.

Eigen-value	Portion Intercept	Portion x_1	Portion x_2	Portion x_3	Portion x_4	Portion x_5
5	.8048	.0004	.1419	.0007	.2537	.7574
6	.1460	.9995	.0031	.9991	.4378	.2001

Figure 3. Fifth and Sixth Variance Proportions for the Naval Hospital Data

7.0 GENERATION OF DATA MATRICES

The Monte Carlo study that follows requires numerous data matrices with varying levels of multicollinearity and leverage. These were generated by the following algorithm (Kennedy and Gentle, 1980):

1. Generate at random an nxp matrix, Z, with n>p, having rank p and the scalar 1 in all positions in its first column.

2. Decompose Z as:

$$Z = Q\begin{bmatrix} R \\ \hline 0 \end{bmatrix}$$

where Q is orthogonal and R is pxp upper-trianglar.

3. Form $U_0 = ZR^{-1}$ and note that $U_0'U_0 = I_p$.

4. Generate at random a p-1 square matrix W and decompose as W=UT, where U is orthogonal. Then form the p square orthogonal matrix:

$$V' = \begin{bmatrix} 1 & \underline{0}' \\ \underline{0} & U \end{bmatrix}.$$

5. Select column means $m_2, m_3, \ldots, m_p$ and form the nxp matrix:

$$E = [\underline{0}\ \underline{m}_2\ \underline{m}_3 \ldots \underline{m}_p] .$$

6. Select a diagonal matrix $D = \text{diag}(\alpha_1, \alpha_2, \ldots, \alpha_p)$ and form the nxp matrix as:

$$X = U_0 DV' + E = A + E .$$

The response vector $\underline{y}$ can then be generated by:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon},$$

where $\underline{\beta}$ is vector of selected coefficients and $\underline{\varepsilon}$ is a vector of computer generated random errors.

When using this algorithm, the random matrices needed in steps 1 and 4 can be obtained by using a uniform random number generator. The decomposition required in step 2 can be accomplished by a Householder transformation (Householder, 1958). The Householder transformation is also used to produce the orthogonal matrix U in step 4.

Since the columns of U_0 are mutually orthogonal and its first column is constant, the mean of each column of the A matrix, except column 1, is zero. Therefore, X has m_j, $j = 2, \ldots, p$, as its column means. The choice of the means and singular values, α_j, in the matrix D regulates the severity of the collinearity. Although the exact level of multicollinearity is random, the bounds developed by Lawson and Hanson (1974) can be used to exercise a certain degree

of control over the condition of the data matrix. It was found that choosing means that were quite similar and singular values that were sizeably different generated severely ill-conditioned data. The computer coding necessary to complete this procedure can be found in Appendix A.

8.0 DESCRIPTION OF THE MONTE CARLO STUDY

As mentioned previously, numerous studies have been performed to ascertain the "best" method of choosing the optimal value of the parameter k among the coefficient estimation based methods. This value of k would presumably give the experimenter the best set of estimates of the regression coefficients, but not necessarily the optimal prediction equation. This is because the coefficient estimation based procedures, on average, poorly estimate the intercept. However, the prediction oriented methods of PRESS, C_k, and DF-trace were developed to remedy this situation. The purpose of the Monte Carlo study is to determine which of these methods, if any, is uniformly best in choosing the value of k that produces the optimal prediction equation for data sets with various levels of multicollinearity and leverage.

This Monte Carlo study began with the generation of data matrices according to the algorithm described in the previous chapter. Because of the limitation of time, only models with two regressor variables were considered. Each data set, X, consisted of twenty-five data points and had built in it various levels, (mild, moderate, or heavy), of both multicollinearity and leverage. Using a predetermined vector of regression coefficients, $\underline{\beta}$, and a vector of normally distributed errors, $\underline{\varepsilon}$, generated by the Kinderman-Ramage algo-

rithm (1976), a vector of responses was computed for the standard regression model:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon} \ .$$

After the completion of the generation of the data matrix and response vector, the optimal k was chosen for each of the prediction oriented methods. This sequence of steps was repeated fifty times per data matrix, i.e., the data matrix, X, remained the same but a new error vector, $\underline{\varepsilon}$, was used. These fifty values of k were then averaged together. It was this average value of k for each method that was then judged by a prediction criteria. This criteria determined which method, on average, produced the "best" prediction of thirty-five new data points randomly chosen from an elliptical region around the original data matrix. The computer code used for this study can be found in Appendix B.

9.0 RESULTS OF THE MONTE CARLO STUDY

The Monte Carlo study described in the previous chapter was completed using fifteen data sets that differed as to the severity of the collinearity and the amount of leverage exerted by one or more data points. For each data set, an average value of k was found for the PRESS, C_k, DF-trace (graphical) and DF-trace (analytic) methods. Also, for each set, thirty-five new data locations were generated from an elliptical region around the original data for the purpose of finding their predicted responses under the ridge regression model.

A criteria was then chosen for the comparison of the average k values produced by the four methods pertaining to the prediction at the thirty-five new x-locations. It should be noted here that this set of new location points involved points which were both near and far in distance from the original data points. This is important because, as Chapter 2.0 stated, even in the presence of multicollinearity new points that are relatively near the original data points are still predicted well. However, points further away in the experimental region are predicted poorly. Therefore, new points at both extremes were used to determine the full prediction capabilities of each method of choosing the parameter k. The criteria chosen for the comparison of the prediction

capabilities was the sum of the mean squared errors of prediction at these thirty-five locations. This can be written as:

$$1/\sigma^2 \sum_{i=1}^{n} MSE(\hat{y}_{R,i})$$

$$= 1/\sigma^2 [\sum_{i=1}^{n} var(\hat{y}_{R,i}) + \sum_{i=1}^{n} bias^2(\hat{y}_{R,i})]$$

$$= tr[X_f(X'X+kI)^{-1}X'X(X'X+kI)^{-1}X_f]$$
$$+1/\sigma^2 [X_f\underline{\beta}-X_f(X'X+kI)^{-1}X'X\underline{\beta}]'[X_f\underline{\beta}-X_f(X'X+kI)^{-1}X'X\underline{\beta}]$$

where $\hat{y}_{R,i}$ is the predicted response under the ridge regression model and X_f is the matrix of new regressor locations. Under the normal experimental conditions, the $\Sigma MSE(\hat{y}_{R,i})$ would need to be estimated because it contains the unknown parameters $\underline{\beta}$ and σ^2. However, under these Monte Carlo conditions of generating the data matrix and response vector, these values are known and $\Sigma MSE(\hat{y}_{R,i})$ need not be estimated. Thus, for each data set, the $\Sigma MSE(\hat{y}_{R,i})$ was calculated for each of the four methods of choosing k.

Table 2 provides a summary of the collinearity and leverage diagnostics for each run of the Monte Carlo study. The first column describes the number and type of leverage points that are found in the data, while the second column notes the severity of the collinearity. The remaining columns are the values of the diagnostics that give credence to the description of the data set.

Table 2 (Part 1 of 1). A Table of Diagnostics for Each Run of the Monte Carlo

NUMBER OF LEVERAGE POINTS	SEVERITY OF COLLINEARITY	LARGEST HAT DIAGONAL	SMALLEST EIGENVALUE	VIF	CONDITION INDEX
None	Mild	.1681	.0001245	697	52.813
None	Moderate	.1681	.0000190	10,000	126.813
None	Heavy	.1681	.0000057	87,693	592.295
1 Mild	Mild	.3457	.0002356	1,976	88.895
1 Mild	Moderate	.3575	.0000449	11,196	210.958
1 Mild	Heavy	.3907	.0000060	82,440	574.247
1 Moderate	Mild	.6084	.0001057	4,726	137.491
1 Moderate	Moderate	.6354	.0000284	17,546	264.919
1 Moderate	Heavy	.6050	.0000058	52,717	459.247
1 Heavy	Mild	.8423	.0001823	2,750	104.878
1 Heavy	Moderate	.8483	.0000304	16,392	256.435
1 Heavy	Heavy	.9284	.0000058	84,992	586.489
1 Mild & 1 Moderate	Mild	.6743 .4307	.0077503	64.76	16.033
2 Mild & 1 Moderate	Mild	.3682 .2309 .5304	.0061992	80.91	17.934
3 Mild	Mild	.3845 .3996 .3739	.0107354	46.82	13.613

Table 3 gives a summary table for the average value of k for each method for every run of the Monte Carlo. Also, listed in the last column is the optimal value of k, i.e., the one that minimizes the $\Sigma MSE(\hat{y}_{R,i})$. This value can be found because as previously noted the values of $\underline{\beta}$ and σ^2 are known.

The first thing to be noted from the table is the wide range of values given by the four methods. In the runs containing one or no leverage points, only the C_k and PRESS methods give similar values of k. However, when more than one leverage point is present, C_k, PRESS, and DF-trace (analytic) provide similar values. Another interesting note is that as collinearity became more severe, each method, with the exception of DF-trace (graphical), provided smaller values of k. This fact was even true among the values of optimal k. This seems somewhat counter-intuitive, for it seems that as collinearity becomes more severe, i.e., at least one λ_i moving closer to 0, that more shrinkage would be required. This result will be fully explained in the following chapter. Finally, it can be noted that the C_k and PRESS methods were the most consistent in being relatively near the optimal value of k.

Table 4 shows describes how each of the values of k shown in the previous table performed in terms of the prediction criteria, $\Sigma MSE(\hat{y}_{R,i})$. The only new column to be included is the value for the sum of the mean square errors of

Table 3 (Part 1 of 1). A Table of Average k Values for Each Method for Each Run of the Monte Carlo

NUMBER OF LEVERAGE POINTS	SEVERITY OF THE COLLINEARITY	C_K	PRESS	DF-TRACE (GRAPH.)	DF-TRACE (ANAL.)	OPTIMAL
None	Mild	.0032	.0088	.0003	.001245	.0050
None	Moderate	.0022	.0039	.0003	.000019	.0033
None	Heavy	.0017	.0009	.0003	.000005	.0022
1 Mild	Mild	.0026	.0056	.0002	.002350	.0048
1 Mild	Moderate	.0019	.0025	.0002	.000045	.0025
1 Mild	Heavy	.0017	.0009	.0003	.000006	.0022
1 Moderate	Mild	.0021	.0036	.0001	.000105	.0035
1 Moderate	Moderate	.0017	.0018	.0001	.000028	.0026
1 Moderate	Heavy	.0017	.0011	.0002	.000009	.0023
1 Heavy	Mild	.0023	.0031	.0002	.000182	.0042
1 Heavy	Moderate	.0016	.0011	.0002	.000030	.0024
1 Heavy	Heavy	.0017	.0006	.0002	.000006	.0021
1 Mild & 1 Moderate	Mild	.0059	.0201	.0003	.007750	.0133
2 Mild & 1 Moderate	Mild	.0062	.0269	.0003	.006199	.0087
3 Mild	Mild	.0071	.0247	.0003	. . .	.0115

Table 4 (Part 1 of 1). A Table of MSE Values for Each Method for Each Run of the Monte Carlo

NUMBER OF LEVERAGE POINTS	SEVERITY OF THE COLLINEARITY	CK	PRESS	DF-TRACE (GRAPH.)	DF-TRACE (ANAL.)	OPTIMAL	OLS
None	Mild	2.115	2.130	4.483	8.634	2.104	433
None	Moderate	2.418	2.414	3.055	24.729	2.413	1700
None	Heavy	3.081	3.087	3.138	66.006	3.080	5401
1 Mild	Mild	3.370	3.302	20.236	14.469	3.298	989
1 Mild	Moderate	2.819	2.818	3.155	8.382	2.818	492
1 Mild	Heavy	2.588	2.594	2.650	118.908	2.587	10310
1 Moderate	Mild	2.905	2.893	12.241	11.360	2.893	746
1 Moderate	Moderate	2.927	2.926	5.232	24.997	2.923	2141
1 Moderate	Heavy	2.655	2.658	2.776	44.749	2.654	3721
1 Heavy	Mild	2.915	2.891	9.565	11.360	2.884	709
1 Heavy	Moderate	3.147	3.155	3.461	23.542	3.145	980
1 Heavy	Heavy	3.018	3.027	3.090	66.319	3.017	5588
1 Mild & 1 Moderate	Mild	3.705	4.147	14.545	3.692	3.312	24.5
2 Mild & 1 Moderate	Mild	3.689	4.762	14.545	3.693	3.632	24.5
3 Mild	Mild	3.379	3.746	14.873	. . .	3.223	20.5

prediction for the ordinary least squares estimate, i.e. the case when k=0. The first observation to be made is that every method considered in this study, no matter how poorly it performs in comparison to the optimal value of the criteria, is a considerable improvement upon the OLS prediction capabilities. The DF-trace (graphical) method appears to perform better as the severity of the collinearity increases, while the DF-trace (analytic) method performed well only in the case of multiple leverage points. It should be noted that this method was expected to perform poorly because the criteria for choosing k was to choose it such that it was as large as the largest "small" eigenvalue, i.e., of those λ_i near zero. In the two variable case, this is very limiting and thus rarely brought about enough shrinkage or deflation. However, it still is a much better method than using ordinary least squares. Finally, the C_k and PRESS methods were uniformly the best in terms of the criteria. In the runs which contained no or one leverage points, the values of the $\Sigma MSE(\hat{y}_{R,i})$ differed only slightly. However, in the runs with multiple leverage points, there seemed to be some separation.

In an effort to determine if the difference between the two was statistically significant, a statistical test was performed on a few runs. These runs were the ones that had values of $\Sigma MSE(\hat{y}_{R,i})$ for C_k and PRESS which differed the most. Since there were no replicates of particular runs, a test of the $\Sigma MSE(\hat{y}_{R,i})$ based on the average k was impossible.

However, as exemplified in Table 5, for each of the fifty values of k computed in a particular run, a $\Sigma MSE(\hat{y}_{R,i})$ was also calculated. Therefore, a paired t-test was performed to determine if there was any significant difference between C_k and PRESS within the run. The three runs that were tested, (1 heavy leverage point-heavy collinearity, 1 mild and 1 moderate leverage point-mild collinearity, and 2 mild and 1 moderate leverage points-mild collinearity), produced the following t-statistics: -.33285, -.36273, and -1.12981. All of these are clearly non-significant. This indicates that though there seems to be some indication of a separation between the performance of the two methods, with C_k performing better, this study in the two variable case does not conclude that. It is believed that if the algorithm used to generate the data could produce heavy collinearity in a data set with multiple high leverage points or that a model with more variables were explored, then this separation might become more significant.

Also, it was attempted to determine whether the four methods of choosing k were consistently ranked the same during the fifty trials within a particular run. Table 5 gives a particular example of these rankings. Kendall's coefficient of concordance (Kendall and Babington-Smith, 1939) was calculated in order to obtain a measure of similarity among the fifty sets of ranks. If W, the coefficient, is large, i.e., near unity, then it indicates the same relative ranks

Table 5 (Part 1 of 2). The ΣMSE and Their Rankings for Run of No Leverage Points & Mild Collinearity

CK	PRESS	DF-TRACE (GRAPH.)	DF-TRACE (ANAL.)	RANKING OF THE MSE FOR EACH ITERATION			
2.10415	2.23107	4.48336	8.63841	1	2	3	4
2.10506	2.19786	4.48336	8.63841	1	2	3	4
2.10707	2.15016	4.48336	8.63841	1	2	3	4
3.12937	2.11237	4.48336	8.63841	2	1	3	4
2.10433	2.20577	4.48336	8.63841	1	2	3	4
2.15076	2.19487	4.48336	8.63841	1	2	3	4
2.10509	2.30691	4.48336	8.63841	1	2	3	4
2.10460	2.28337	4.48336	8.63841	1	2	3	4
12.27114	6.43441	4.48336	8.63841	4	2	1	3
2.10442	2.22238	4.48336	8.63841	1	2	3	4
12.27112	3.12937	4.48336	8.63841	4	1	2	3
2.10433	2.15466	4.48336	8.63841	1	2	3	4
12.27114	2.27117	4.48336	8.63841	4	3	1	2
2.10465	2.19592	4.48336	8.63841	1	2	3	4
2.10544	2.13166	4.48336	8.63841	1	2	3	4
2.10422	2.28857	4.48336	8.63841	1	2	3	4
2.11836	2.10539	4.48336	8.63841	2	1	3	4
6.43441	2.54172	4.48336	8.63841	3	1	2	4
6.43441	4.48336	4.48336	8.63841	3	2	1	4
4.48336	2.14514	4.48336	8.63841	3	1	2	4
4.48336	2.54172	4.48336	8.63841	2	1	3	4
2.12058	2.12934	4.48336	8.63841	1	2	3	4
2.10416	2.16238	4.48336	8.63841	1	2	3	4
2.10958	2.13533	4.48336	8.63841	1	2	3	4
2.66572	2.11322	4.48336	8.63841	2	1	3	4
2.11165	2.29893	4.48336	8.63841	1	2	3	4
2.10506	2.17233	4.48336	8.63841	1	2	3	4
6.43441	2.45338	4.48336	8.63841	3	1	2	4
2.10783	2.11097	4.48336	8.63841	1	2	3	4
2.10783	2.14728	4.48336	8.63841	1	2	3	4
2.10544	2.16238	4.48336	8.63841	1	2	3	4
2.10644	2.18284	4.48336	8.63841	1	2	3	4
2.10599	2.11632	4.48336	8.63841	1	2	3	4
2.11186	2.10908	4.48336	8.63841	2	1	3	4
2.14514	2.10419	4.48336	8.63841	2	1	3	4
2.10544	2.15016	4.48336	8.63841	1	2	3	4
2.11186	2.13286	4.48336	8.63841	1	2	3	4
2.10745	2.46586	4.48336	8.63841	1	2	3	4
2.10644	2.13407	4.48336	8.63841	1	2	3	4

Table 5 (Part 2 of 2). The ΣMSE and their Rankings for Run of No Leverage Points & Mild Collinearity

CK	PRESS	DF-TRACE (GRAPH.)	DF-TRACE (ANAL.)	RANKING OF THE MSE FOR EACH ITERATION			
2.10599	2.15311	4.48336	8.63841	1	2	3	4
2.10599	2.17401	4.48336	8.63841	1	2	3	4
2.11836	2.14312	4.48336	8.63841	1	2	3	4
2.11474	2.14177	4.48336	8.63841	1	2	3	4
2.11065	2.11808	4.48336	8.63841	1	2	3	4
2.10451	2.17749	4.48336	8.63841	1	2	3	4
6.43441	2.54172	4.48336	8.63841	3	1	2	4
2.11186	2.15766	4.48336	8.63841	1	2	3	4
2.10644	2.16892	4.48336	8.63841	1	2	3	4
2.84743	2.14022	4.48336	8.63841	2	1	3	4
2.10707	2.15766	4.48336	8.63841	1	2	3	4

for the fifty sets of ranks. Table 6 lists the value of W for each run of the Monte Carlo for both the values of k produced by the methods and for the $\Sigma MSE(\hat{y}_{R,i})$ produced by the corresponding values of k. It can be seen that the ranks for the fifty trials in most runs are highly correlated. This means that the ranking of the methods using the averaged value of k is very consistent with the rankings throughout the fifty trials.

Table 6 (Part 1 of 1). A Table of Kendall's W Values for k and MSE for Each Run of the Monte Carlo

NUMBER OF LEVERAGE POINTS	SEVERITY OF THE COLLINEARITY	KENDALL'S COEFFICIENT OF CONDCORDANCE (W) (FOR AVERAGE K)&(FOR VALUES OF MSE)	
None	Mild	.75504	.70464
None	Moderate	.76384	.72096
None	Heavy	.68848	.83104
1 Mild	Mild	.47728	.54592
1 Mild	Moderate	.68688	.77200
1 Mild	Heavy	.71344	.72208
1 Moderate	Mild	.65296	.79776
1 Moderate	Moderate	.60976	.78736
1 Moderate	Heavy	.64912	.79264
1 Heavy	Mild	.50848	.62224
1 Heavy	Moderate	.58848	.65184
1 Heavy	Heavy	.67104	.84528
1 Mild & 1 Moderate	Mild	.69712	.67600
2 Mild & 1 Moderate	Mild	.71248	.76368
3 Mild	Mild	.73040	.65152

10.0 CONCLUSIONS

Some very definite conclusions can be drawn from this study about the prediction capabilites of the PRESS, C_k, and DF-trace methods. First of all, each method provides the ridge regression model with a considerable improvement in prediction for all points over the capabilites of ordinary least squares model when multicollinearity is present. Even in the case of multiple leverage points in the data and mild collinearity, the worst method brought a reduction of nearly one-half in the $\Sigma MSE(\hat{y}_{R,i})$ in comparison to that of ordinary least squares.

Secondly, all methods performed well, i.e., near the optimum of the criteria, for several of the situations in the study. The DF-trace (analytic) method performed well when multiple leverage points were found in the data, even though it was the worst in relationship to the other three for the remaining runs. DF-trace (graphical) worked well when heavy or severe collinearity was present. But, overall, the performances of the PRESS and C_k methods were quite good and seemed only to deviate when multicollinearity and leverage points were present in the data. Though this difference was never found to be statistically significant, there is reason to believe that the deviation would widen if more variables or leverage points were present. The reason for this is that

the PRESS method with its use of the PRESS residuals from the RR model are more sensitive to the high leverage points. This quality, though thought to be favorable, is actually somewhat detrimental as it causes the method to produce a k that is too large. The C_k method, though, seems to continue to perform very well even in the presence of multiple high leverage points. Therefore, the C_k is considered to perform the "best" according to the study.

A third point can be made about the variability in the range of the values of k that brought favorable results in the $\Sigma MSE(\hat{y}_{R,i})$. An example of this would be the run that had 1 moderate leverage point and high collinearity. In this case, the values of k ranged from .0002 to .0017, a difference in magnitude of 8.5. Yet, the values of the $\Sigma MSE(\hat{y}_{R,i})$ for these k were 2.7769 and 2.6553, respectively. This result was particularly true in the cases of high collinearity and multiple leverage points.

Another point to be made about the values of k was that they decreased as the severity of the collinearity increased. This, at first, seems counter-intuitive. However, a close look at the criteria will make the result more reasonable. The sum of the mean square errors of prediction is made up of two components: $\Sigma var(\hat{y}_{R,i})$ and $\Sigma bias^2(\hat{y}_{R,i})$. As collinearity becomes more severe, the $\Sigma var(\hat{y}_{R,i})$ component is much more sensitive to the addition of k, even the smallest of values. This is seen in Table 5 in all the heavy

collinearity cases. The $\Sigma MSE(\hat{y}_{R,i})$ for ordinary least squares is very large. Yet, the addition of a k as small as .000006 brings a reduction of about one hundredfold. The reduction of this small of a k value is less when the collinearity is less severe. The conclusion then is this: When collinearity is mild the value of k needed to bring significant deflation in the $\Sigma var(\hat{y}_{R,i})$ and, hence, in $\Sigma MSE(\hat{y}_{R,i})$, is larger than when the collinearity is more severe. Therefore, this result, though at first seemingly counter-intuitive, is very reasonable.

Finally, some comments need to be made about future studies that could be derived from this one. As mentioned earlier, more work could be done involving models with more than two variables. This would probably allow more flexibility with the algorithm used to produce the data matrices. Also, more could be done to determine if the PRESS and C_k methods do significantly differ as the number of leverage points and collinearity increase. A more expanded study could be done on the two variable model to include the coefficient estimation based techniques mentioned in the introduction. This might determine if the prediction based methods are any better in improving the prediction capabilites of an ill-conditioned model.

APPENDIX A. PROGRAM TO GENERATE DATA MATRICES

```
DATA Z;
INPUT X0 X1 X2;
CARDS;

(Input the random nxp matrix Z)

DATA W;
INPUT W1 W2;
CARDS;

(Input the random pxp matrix W)

PROC MATRIX ;
   FETCH W0 DATA=Z(KEEP=X0);      *A HOUSEHOLDER TRANS-
                                  *              FORMATION OF Z;
   A0=J(1,1,1);                   *TRANSFORMATION OF COLUMN 1;
   B0=J(24,1,1);
   C=A0//B0;
     S0=SQRT(W0'*W0);
     U0=W0#C;
     IF W0(1,1)>0 THEN S0=-S0;
     U0(1,1)=W0(1,1)-S0;
     BETA0=-S0#U0(1,1);
     ID=I(25);
     H1=ID-((U0*U0')#/BETA0);
   FETCH X0 DATA=Z;
     X1=H1*X0;
     W1=X1( ,2);                  *TRANSFORMATION OF COLUMN 2;
     W1(1,1)=0;
   A1=J(1,1,0);
   B1=J(24,1,1);
   C1=A1//B1;
     S1=SQRT(W1'*W1);
     U1=W1#C1;
     IF W1(2,1)>0 THEN S1=-S1;
     U1(2,1)=W1(2,1)-S1;
     BETA1=-S1#U1(2,1);
     ID1=I(25);
     H2=ID1-((U1*U1')#/BETA1);
   X2=H2*X1;
   W2=X2( ,3);                    *TRANSFORMATION OF COLUMN 3;
   W2(1,1)=0; W2(2,1)=0;
   A2=J(2,1,0);
   B2=J(23,1,1);
   C2=A2//B2;
```

```
    S2=SQRT(W2'*W2);
    U2=W2#C2;
    IF W2(3,1)>0 THEN S2=-S2;
    U2(3,1)=W2(3,1)-S2;
    BETA2=-S2#U2(3,1);
    ID2=I(25);
    H3=ID2-((U2*U2')#/BETA2);
  X3=H3*X2;
 H=H3*H2*H1;
 HX=H*X0;                        *UPPER TRIANGULAR MATRIX R;
 R=J(3,3,0);
 R(1, )=HX(1, );
 R(2, )=HX(2, );
 R(3, )=HX(3, );
 INVR=INV(R);                    *INVERSE OF MATRIX R;
 UNOT=X0*INVR;                   *COMPUTATION OF U-NOT;
 IDENU=UNOT'*UNOT;
 FETCH W DATA=W;                 *RANDOM MATRIX W;
 P=W( ,1);
 S=SQRT(P'*P);                   *HOUSEHOLDER TRANSFORMATION TO;
 IF W(1,1)>0 THEN S=-S;          *FORM ORTHOGONAL MATRIX U;
 P(1,1)=W(1,1)-S;
 BETA=-S#P(1,1);
 ID3=I(2);  U=ID3-((P*P')#/BETA);
 IDEN=U*U'; U=U'; T=U*W;
 DUM=J(2,1,0); DUM2= 1 0 0;
 U=DUM||U; VPRIME=DUM2//U;          *CREATION OF V';
 DUM3=J(25,1,0);DUM4=J(25,1,20);
 DUM5=J(25,1,25.27);
 E=DUM3||DUM4||DUM5;                *CREATION OF ERROR MATRIX, E;
 DUM6=70.000  0.0030 70.0000;
 D=DIAG(DUM6); A=UNOT*D*VPRIME;     *SINGULAR VALUE DECOMPOSITION;
 X=A+E;                             *DATA MATRIX X;
 PRINT X;
//
```

APPENDIX B. PROGRAM FOR MONTE CARLO STUDY

(FORTRAN program to generate normal random errors.)

```
      DIMENSION E(25,50)
      ISEED=47583403
      DO 100 J=1,25
         DO 110 I=1,50
            CALL NORMKR(X,ISEED)
            E(J,I)=45.*X
  110    CONTINUE
      WRITE(9,120) (E(J,I), I=1,50)
  120 FORMAT(1X,7F14.8/1X,7F14.8)
  100 CONTINUE
      STOP
      END
C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C
      SUBROUTINE NORMKR(X,ISEED)
C
C  THIS SUBROUTINE GENERATES NORMALS BY THE KINDERMAN-RAMAGE
C  COMPOSITION REJECTION ALGORITHM.
C
      CALL RANDM(U,ISEED)
      IF(U.GT..884070402298758)GOTO 100
C
C  INNER TRIANGULAR REGION 88% OF THE TIME
C
      CALL RANDM(V,ISEED)
      X=2.2160358671664471*(1.131131635444180*U+V-1.)
      RETURN
 100  IF(U.LT..973310954173898)GOTO 120
C
C  TAIL REGION GENERATED HERE
C
 110  CALL RANDM(V,ISEED)
      CALL RANDM(W,ISEED)
      T=2.45540748228413-ALOG(W)
      IF(V*V*T.GT.2.45540748228413)GOTO 110
      X=SQRT(2.*T)
      IF(U.LT..986655477086949)X=-X
      RETURN
 120  IF(U.LT..958720824790463)GOTO 150
C
C  FIRST TRIANGULAR REGION
C
```

```
 130  CALL RANDM(V,ISEED)
      CALL RANDM(W,ISEED)
      Z=V-W
      IF(V.LT.W)GOTO 140
      TEMP=V
      V=W
      W=TEMP
 140  X=2.216035867166471-.630834801921960*V
      IF(W.LE..755591531667601)GOTO 210
      IF(.085828214837637*(W-V).LE.EXP(-.5*X*X))GOTO 210
      GOTO 130
 150  IF(U.LT..911312780288703)GOTO 180
C
C  SECOND TRIANGULAR REGION
C
 160  CALL RANDM(V,ISEED)
      CALL RANDM(W,ISEED)
      Z=V-W
      IF(V.LT.W)GOTO 170
      TEMP=V
      V=W
      W=TEMP
 170  X=.479727404222441+1.105473661022070*V
      IF(W.LT..872834976671790)GOTO 210
      IF(.123487779544339*(W-V).LE.EXP(-.5*X*X))GOTO 210
      GOTO 160
C
C  THIRD TRIANGULAR REGION
C
 180  CALL RANDM(V,ISEED)
      CALL RANDM(W,ISEED)
      Z=V-W
      IF(V.LT.W)GOTO 190
      TEMP=V
      V=W
      W=TEMP
 190  X=.479727404222441-.595507138015940*V
      IF(W.LE..805577924423817)GOTO 210
      IF(.133797674824476*(W-V).LE.EXP(-.5*X*X))GOTO 210
      GOTO 180
C
C  THE SWITCH IN SIGN IS MADE HERE FOR ALL THREE
C              TRIANGULAR REGIONS
C
 210  IF(Z.LT.0.)X=-X
      RETURN
      END
/*
```

(The uniform number generator in machine language.)

```
//LKED.SYSIN DD *
//          DD *
βESD       ≡  αRANDM    h
βTXT
        αT}νqΞ≡‹⁺ ⁻→ΠΟΛ!&Od→}Ο¤ω⁻ω⁻‹⁻0&&0‹0
βTXT
      δ  αOh&αω₊→\O∝!&O≈&⁺h&!&O-&₌qT}ν
βTXT   Π  αc+$ζ²4.Tι-c+c+≈α
βEND                                  15741SC103 020185027
//GO.FT09F001 DD UNIT=VIO,DISP=(NEW,PASS),
//      SPACE=(CYL,5),DCB=(LRECL=100,BLKSIZE=1000),
//      DSN=&&PHIL
//GO.SYSIN DD *
//    EXEC SAS
//ERROR DD DSN=&&PHIL,DISP=(OLD,PASS)
//SYSIN DD *
DATA ONE;
INFILE ERROR;
INPUT E1-E50;
DATA XDATA;
INPUT X1 X2 ;
CARDS;

(Input data matrix with particular level of
 collinearity and leverage.)

DATA NEW;
INPUT  X1 X2;
CARDS;

(Input 35 new data points to be predicted.)

PROC MATRIX ;

* * * * * * * * * * * * BEGINNING OF MONTE CARLO * * * * * * * *;

  FETCH E DATA=ONE;
  FETCH X DATA=XDATA (KEEP= X1 X2);
  X=J(25,1,1)||X;
  N=NROW(X);
  BETA = 1.76/
         7.45/
         3.745;
  FETCH X1 DATA=XDATA (KEEP= X1 X2);
  MEAN=X1(+, )#/N;
  XCEN=X1-J(N,1,1)*MEAN;
  SS=X1(##, );
  XCENSC=XCEN*DIAG(1#/SQRT(SS));
  XCENSC=J(25,1,1)||XCENSC;
  FETCH XF DATA=NEW;
  XFC=XF-J(35,1,1)*MEAN;
  XFC=XFC*DIAG(1#/SQRT(SS));
  XFC=J(35,1,1)||XFC;
```

```
   XF=J(35,1,1)||XF;
   XPX=XCENSC'*XCENSC;
   CON=J(50,4); CON2=J(50,4); MSE=J(50,4);
   CHART=J(50,3,0); CHART2=J(50,2,0);

* * * * * * * * * * * COMPUTATION OF THE Y VECTOR * * * * * * *;

   DO L=1 TO 50;
      Y=X*BETA+E( ,L);

* * * * * * * * * * * * * COMPUTATION OF R SQUARE * * * * * * * *;

      SSREG=Y'*(XCENSC*(INV(XCENSC'*XCENSC))*XCENSC')*Y;
      A=J(25,1,1);
      SSTOT=(Y'*Y)-(Y'*A*(INV(A'*A))*A'*Y);
      RSQ=SSREG#/SSTOT;

* * * * * * * * * * COMPUTATION OF THE CK STATISTIC * * * * * *;

      K=-.0001; P=0; CKPREV=5.5;
      YBAR=Y(+, )#/25;
      Y=Y-YBAR#J(25,1,1);
      SSRES=Y'*(I(25)-(XCENSC*INV(XCENSC'*XCENSC)
                        *XCENSC'))*Y;
      DO WHILE (P=0);
         K=K+.0001;
         ID=I(2); ID=K#ID; ID=J(2,1,0)||ID; ID=J(1,3,0)//ID;
         HK=XCENSC*INV((XCENSC'*XCENSC)+ID)*XCENSC';
         SSRESR=Y'*(I(25)-HK)*(I(25)-HK)*Y;
         CK=(2#TRACE(HK))+(22#(SSRESR#/SSRES))-25+2;
         IF CK>CKPREV THEN P=1;
         ELSE CKPREV=CK;
      END;
      KCP=K;
      CHART(L, )=RSQ||KCP||CKPREV;

* * * * * * * * * * COMPUTATION OF THE PRESS STATISTIC * * * * *;

      K=-.0001; PRPREV=700000000000; P=0;
      DO WHILE (P=0);
         PRESS=J(25,1,0);
         XX=J(24,2); YY=J(24,1);
         K=K+.0001;

* * * * * * * * * * * SETTING ASIDE THE ITH POINT * * * * * * * ;

         DO PT=1 TO 25;
            LA=0;
            DO IR=1 TO 25;
               IF IR NE PT THEN DO;
               LA=LA+1;
```

```
            XX(LA,1)=X1(IR,1);
            XX(LA,2)=X1(IR,2);
            YY(LA,1)=Y(IR,1);
            END;
         END;

* * * * * * OLS ANALYSIS ON DATA WITHOUT THE ITH POINT * * * * *;

         MEAN2=XX(+, )#/24;
         XX=XX-J(24,1,1)*MEAN2;
         SS2=XX(##, ); STD=SQRT(SS2);
         XX=XX*DIAG(1#/STD);
         BNEGI=INV(XX'*XX+(K#I(2)))*XX'*YY;
         XOBS=J(1,1,1)||X1(PT, );
         YYBAR=(YY'*J(24,1,1))#/24; S=1.#/DIAG(STD);
         T=VECDIAG(S);
         TBNEGI=T#BNEGI; BO=TBNEGI'*MEAN';
         INTCPT=YYBAR-BO; TBNEGI=INTCPT//TBNEGI;
         PRESS(PT,1)=Y(PT,1)-XOBS*TBNEGI;
      END;
      PRESS=PRESS'*PRESS;
      IF PRESS>PRPREV THEN P=1;
      ELSE PRPREV=PRESS;
   END;
   KPRESS=K;
   CHART2(L, )=KPRESS||PRPREV;

* * * * * * * * * * * RANKING OF THE VALUES OF K * * * * * * * * *;

   METHODS=J(1,4,0);
   METHODS(1,1)=KCP;  METHODS(1,2)=KPRESS;
   METHODS(1,3)=.0003; METHODS(1,4)=.006199;
   R=RANK(METHODS);
   CON(L, )=R;

* * * * * * * * * * * RANKING OF THE MEAN SQUARE ERRORS * * * * * ;

   MSEYHAT=J(1,4,0);
   IDEN=I(2); IDEN=J(2,1,0)||IDEN; IDEN=J(1,3,0)//IDEN;
   IDEN=J(1,3,0)//IDEN;
      SIGMA=2025.0;
   DO M=1 TO 4;
      K=METHODS(1,M); XPXKI=INV(XCENSC'*XCENSC+(K#IDEN));
      VAR=XFC*XPXKI*XCENSC'*XCENSC*XPXKI*XFC';
      SUMVAR=TRACE(VAR);
      SSQBIAS=((BETA'*XF'*XF*BETA)-(2#BETA'*XF'*XFC
              *XPXKI*XCENSC'*X*BETA)+(BETA'*X'*XCENSC
              *XPXKI*XFC'*XFC*XPXKI*XCENSC'*X*BETA))
         #/SIGMA;
      SUMMSE=SUMVAR+SSQBIAS;
      MSEYHAT(1,M)=SUMMSE;
```

```
      END;
      MSE(L, )=MSEYHAT;
      R2=RANK(MSEYHAT);
      CON2(L, )=R2;
   END;

* * * * * * * * KENDALL'S COEFFICIENT OF CONCORDANCE * * * * *;

   CONTOT=CON(+, );
   CON2TOT=CON2(+, );
   CONSQ=CONTOT#CONTOT;
   CON2SQ=CON2TOT#CON2TOT;
   Z=J(4,1,1);
   SUMR1=CONTOT*Z; SUMR1SQ=CONSQ*Z;
   SUMR2=CON2TOT*Z; SUMR2SQ=CON2SQ*Z;
   SSR1=SUMR1SQ-((SUMR1#SUMR1)#/4);
   SSR2=SUMR2SQ-((SUMR2#SUMR2)#/4);
   W1=(12#SSR1)#/150000;
   W2=(12#SSR2)#/150000;

* * * CALCULATION OF AVG K FOR EACH METHOD OVER 50 RUNS * * *;

   AVGK=J(1,4,1);
   CHART3=CHART||CHART2;
   Z1=J(1,50,1);
   MEANKCP=(Z1*CHART3( ,2))#/50;
   MEANKPR=(Z1*CHART3( ,4))#/50;
   AVGK(1,1)=MEANKCP;  AVGK(1,2)=MEANKPR;
   AVGK(1,3)=.0003   ;  AVGK(1,4)=.006199;

* * * * * MEAN SQUARE ERROR FOR ORDINARY LEAST SQUARES * * * *;

   MSEOLS=TRACE(XFC*INV(XCENSC'*XCENSC)*XFC');

* * * * CALCULATION OF MEAN SQUARE ERROR FOR EACH AVG K * * * ;

   MSEAVG=J(1,4,1);
   DO N=1 TO 4;
      K=AVGK(1,N); XPXKI=INV(XPX+(K#IDEN));
      VAR=XFC*XPXKI*XCENSC'*XCENSC*XPXKI*XFC';
      SUMVAR=TRACE(VAR);
      SSQBIAS=((BETA'*XF'*XF*BETA)-(2#BETA'*XF'*XFC*XPXKI
               *XCENSC'*X*BETA)+(BETA'*X'*XCENSC
               *XPXKI*XFC'*XFC*XPXKI*XCENSC'*X*BETA))
               #/SIGMA;
      SUMMSE=SUMVAR+SSQBIAS;
      MSEAVG(1,N)=SUMMSE;
   END;
   MSEAVG=MSEAVG||MSEOLS;

* * * * * * * * * PRINT OUT OF THE DATA * * * * * * * * * * ;
```

```
C1= 'RSQ' 'K_FOR_CK' 'CK' 'K_PRESS' 'PRESS';
PRINT CHART3 COLNAME=C1;
C2= 'MSE_CK' 'MSE_PRSS' 'MSE_DF1' 'MSE_DF2';
PRINT MSE COLNAME=C2;
RANKS=CON||CON2;
C3= 'K_CK' 'K_PRESS' 'K_DF1' 'K_DF2' 'MSE_CK' 'MSE_PRSS'
      'MSE_DF1' 'MSE_DF2';
PRINT RANKS COLNAME=C3;
KENDLW=W1||W2;
C6= 'K' 'MSE'; R6= 'W';
PRINT KENDLW COLNAME=C6 ROWNAME=R6;
C4= 'CK' 'PRESS' 'DF1' 'DF2'; R4= 'K';
PRINT AVGK COLNAME=C4 ROWNAME=R4;
C5='CK' 'PRESS' 'DF1' 'DF2' 'OLS'; R5= 'MSE';
PRINT MSEAVG COLNAME=C5 ROWNAME=R5;
```

BIBLIOGRAPHY

Allen, D.M. (1976), "The Prediction Sum of Squares as a Criterion for Selection of Prediction Variables," Technical Report No. 23, Dept. of Statistics, Univ. of Kentucky.

Belsley, D.A., E. Kuh, and R.E. Welsch (1980), Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, New York: Wiley.

Graybill, F.A. (1976), Theory and Application of the Linear Model, North Scituate, MA: Duxbury Press.

Hoaglin, D.C. and R.E. Welsch (1978), "The Hat Matrix in Regression and ANOVA," American Statistician, 32, 1, 17-22.

Hoerl, A.E. and R.W. Kennard (1970a), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, 12, 55-67.

Hoerl, A.E. and R.W. Kennard (1976), "Ridge Regression: Iterative Estimation of the Biasing Parameter," Commun. Statist., A5, 77-88.

Hoerl, A.E., R.W. Kennard, and K.F. Baldwin (1975), "Ridge Regression: Some Simulations," Commun. Statist., 4, 105-123.

Householder, A.S. (1958), "Unitary Triangularization of a Nonsymmetric Matrix," JACM, 5, 339-342.

Kendall, M.G. and B. Babington-Smith (1939), "The Problem of m Rankings," Ann. Math. Stat., 10, 275-287.

Kennedy, W.J., Jr. and J.E. Gentle (1980), Statistical Computing, New York: Dekker.

Kinderman, A.J. and J.G. Ramage (1976), "Computer Generation of Normal Random Variables," JASA, 71, 893-896.

Lawson, C.L. and R.J. Hanson (1974), Solving Least Squares Problems, Englewood Cliffs, NJ: Prentice-Hall.

Mallows, C.L. (1973), "Some Comments on Cp" Technometrics, 15, 661-675.

Marquardt, D.W. (1970), "Generalized Inverses, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation," Technometrics, 12, 591-612.

Montgomery, D.C. and E.A. Peck (1982), Introduction to Linear Regression Analysis, New York: Wiley.

Myers, R.H. ( ), Classical and Modern Regression with Applications, To be published in December 1985 by Duxbury Press, North Scituate, MA.

Procedures and Analyses for Staffing Standards Development: Data/Regression Analysis Handbook,(1979), San Diego, CA: Navy Manpower and Material Analysis Center.

Rao, C.R. (1967), Linear Statistical Inference, New York: Wiley.

Tripp, R.E. (1983), "Non-Stochaistic Ridge Regression and Effective Rank of the Regressors Matrix," Dissertation, Virginia Polytechnic Institute and State University.

The vita has been removed from the scanned document