Ridge Regression

Dave Mikelson
Department of Mathematical sciences
Montana State University

July 15, 1997

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Dave Mikelson

This writing project has been read by the writing project director and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date

John J. Borkowski
Writing Project Director

1 Introduction

Ridge Regression (RR) is a method of estimating the parameter β in the linear model $y^* = X^*\beta + \varepsilon$ when there are near linear dependencies among the columns of X^*. Introduced by Hoerl and Kennard(1970), RR produces a biased estimate of β with a smaller variance than the Ordinary Least Squares (OLS) estimator. As an alternative to OLS, the primary advantage of RR over OLS are more stable and more interpretable parameter estimates.

Terminology will be introduced followed by a short review of multicollinearity, the problems associated with it, and some of the diagnostics used to detect it. Next, the concept of RR will be introduced as well as some of its mathematical development. Although there are a number of different techniques used to find the "best" biasing constant used in RR, this paper will concentrate on just one of these techniques, the Ridge Trace (RT). The RT is a graphical technique and is the easiest method of finding a biasing constant that reduces the variance of the estimator. Finally, an example will be presented.

2 Terminology

Before reviewing multicollinearity, it is necessary to introduce the following terminology. The assumed linear model that will be used throughout this paper is $y^* = X^*\beta + \varepsilon$ where:

n: Number of observations

p: Number of measurements taken on each observation

y^*: nx1 random vector of measured centered and scaled responses

$$y^* \sim (X^*\beta, \sigma^2 I_p)$$

ε: nx1 random vector of unknown error terms

$\varepsilon \sim (0, \sigma^2 I_p)$ where 0 is a nx1 vector of zeros

X: nxp matrix of known uncentered and unscaled explanatory values

X^*: nxp matrix of known centered and scaled explanatory values

(also referred to as regressors)

β: px1 vector of unknown coefficients

$\hat{\beta}$: px1 vector, OLS estimator of β

$\hat{\beta_R}$: px1 vector, RR estimator of β.

The OLS estimator for the parameter β is $\hat{\beta} = ((X^*)'X^*)^{-1}(X^*)'y^*$. This estimator can also be shown to be the best linear unbiased estimator (BLUE) of β. *Best* refers to the fact that among linear unbiased estimators, $\hat{\beta}$ has the minimum variance. A potential problem with this estimator is no upper bound on its variance. A biased estimator with a corresponding smaller variance may exist.

The RR estimator is $\hat{\beta_R} = ((X^*)'X^* + kI_p)^{-1}(X^*)'y^*$ where k is the biasing constant, often referred to as the shrinkage parameter. This estimator is derived by modifying the normal equation from $(X^*)'X^*\beta = (X^*)'y^*$ to $((X^*)'X^* + kI_p)\beta = (X^*)'y^*$. The main objective of RR is to find an appropriate value for k such that the MSE($\hat{\beta_R}$) is less than the MSE($\hat{\beta}$).

The nxp matrix of explanatory values, X^*, is assumed to be centered and scaled. Centering implies the mean of each column vector of the original explanatory matrix X is subtracted from each element of the corresponding column vector. Next the column vector is scaled by dividing each element by the square root of the sum of squared deviations from the mean.

Let x_{ij} be the i^{th} element in the j^{th} column from the original explanatory matrix X. Then,

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

for $i = \{1, 2, 3, ..., n\}$ and $j = \{1, 2, 3, ..., p\}$. Centering the data is not always recommended or required. Brown (1977) and Myers (1990) give further details. The response vector y is centered and scaled in the same manner as the explanatory variables. There are other methods of scaling besides the one described above. When the regressors and response variables are scaled in this way there is no constant term in the model. In other words, the y intercept is zero. In addition, $(X^*)'X^*$ can be interpreted as the "correlation matrix" for the regressor variables. All elements of $(X^*)'X^*$ take on values in the interval [-1,1]. Although the explanatory variables are not random, the structure of $(X^*)'X^*$ does measure linear dependence among the regressor variables (e.g. ones along the diagonal). One advantage of computing this "correlation matrix" is that it can be used for the detection of multicollinearity which will be discussed later. Another advantage for centering and scaling the data is that the magnitude of the regression coefficients are comparable. Without centering and scaling the magnitude of the coefficients are not necessarily related to their importance in the estimated model (Hoerl and Kennard, 1988).

3 Multicollinearity

Multicollinearity in a regression setting is defined to be a linear or near linear dependence between two or more column vectors of the X^* matrix. For example, if the three column vectors x_1^*, x_2^*, and x_3^* are linearly dependent, then there exists a set of nonzero constants, say

$\{a_1, a_2, a_3\}$, such that $\sum a_i x_i = 0$. If the three vectors are approximately linearly dependent then $\sum a_i x_i \approx 0$. In either case multicollinearity is said to exist in OLS.

The degree to which multicollinearity is present is subjective and depends on the strength of the linear relationships among the regressor vectors of X^*. The explanatory vectors are often treated as independent. In reality, however, these vectors are rarely independent. Often the regression model is used to estimate or predict changes in a response when there is a unit change in a particular regressor. Changing a particular regressor value while all remaining regressors are held constant is easy to do mathematically, but in actuality, when one regressor value is changed the value of the other regressors related to it will also change.

The degree of linear dependence between the regressors can effect the stability of the OLS estimator. In the extreme case of absolute linear dependence between two or more column vectors, the matrix $(X^*)'X^*$ is less than full rank and therefore not invertible. Thus, no unique solution to the normal equation exists. In the case of approximate linear dependence, problems still occur depending on the severity of the multicollinearity. One of the major concerns when multicollinearity is present is the increase in variance of the OLS regression coefficients:

$$\begin{aligned} \mathrm{Var}(\hat{\beta}) &= \mathrm{Var}[((X^*)'X^*)^{-1}(X^*)'y] \\ &= ((X^*)'X^*)^{-1}(X^*)'\mathrm{Var}(y)X^*((X^*)'X^*)^{-1} \\ &= \sigma^2((X^*)'X^*)^{-1}. \end{aligned}$$

Because $((X^*)'X^*)^{-1}$ is nearly singular, inverting it is similar to reciprocating a scalar that is close to zero. The result is some of the elements along the diagonal of $((X^*)'X^*)^{-1}$

are large (Birkes and Dodge, 1993). For this reason the variance of $\hat{\beta}$ is said to be inflated. Another way to display the problem associated with multicollinearity is to look at the squared distance between $\hat{\beta}$ and β, (Hoerl and Kennard, 1970). Defining this value as

$$L^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta).$$

Its expectation,

$$E(L^2) = E(\hat{\beta}'\hat{\beta}) - 2\beta' E(\hat{\beta}) + \beta'\beta$$

where

$$E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 tr[((X^*)'X^*)^{-1}]$$

and tr $\equiv$ trace operator. Therefore,

$$E(L^2) = \sigma^2 \mathrm{tr}[((X^*)'X^*)^{-1}] \tag{1}$$

Because $(X^*)'X^*$ is symmetric it is also diagonalizable (Searle, 1982). Diagonalizable implies the matrix $(X^*)'X^*$ can be written as $(X^*)'X^* = P\Lambda P'$ where P is an orthogonal matrix. Hence, $P'P = PP' = I_p$, and Λ is a diagonal matrix composed of eigenvalues of $((X^*)'X^*)$ along the diagonal. Also from Searle (1982),

$$\begin{aligned} ((X^*)'X^*)^{-1} &= (P\Lambda P')^{-1} \\ &= P'\Lambda^{-1}P. \end{aligned}$$

Because $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, when multiplication is conformable,

$$\begin{aligned} \mathrm{tr}((X^*)'X^*)^{-1} &= \mathrm{tr}(P'\Lambda^{-1}P) \\ &= \mathrm{tr}(P'P\Lambda^{-1}) \\ &= \mathrm{tr}(\Lambda^{-1}) \\ &= \sum \lambda_i^{-1} \end{aligned} \tag{2}$$

where λ_i is the i^{th} diagonal element in Λ. As the severity of multicollinearity increases some of the eigenvalues of $(X^*)'X^*$ decrease, and from equations 1 and 2 the expected squared distance from $\hat{\beta}$ to β increases. The precision of the estimator $\hat{\beta}$ decreases as multicollinearity increases. Because the variance of $\hat{\beta}$ is inflated, the regression coefficients may appear with signs opposite than what were expected from a subject matter expert (Mullet, 1976). This poor precision in estimating individual model parameters does not necessarily imply that the estimated model is a poor predictor (Gunst, Mason & Weber, 1975). Despite the fact that the elements of β are poorly estimated, a linear combination of these estimates may provide satisfactory prediction. But, poor prediction can occur when predictors do not follow the same collinearity pattern that were used in the computation of regression coefficients.

These are some of the problems caused by multicollinearity. Some diagnostics for multicollinearity will now be presented.

4 Detection of Multicollinearity

As stated earlier, when the explanatory variables are centered and scaled (resulting in x^*), the matrix $(X^*)'X^*$ can be interpreted as a correlation matrix. Recall that correlation is

a measure of linear dependence between two quantitative variables. Therefore, $(X^*)'X^*$ may be useful in identifying if multicollinearity is present. If collinearity exists between two regressors, the elements in $(X^*)'X^*$ corresponding to these two regressors will be "large", i.e. close to one. The disadvantage with this method occurs when the collinearity exist between three or more regressors, because this diagnostic may not detect it.

A second technique that can be used as a diagnostic of multicollinearity are the F-statistics and $p-values$ given in most computer packages' regression output. If the F-statistic for the overall regression is significant but the individual $p-values$ for the regression coefficients are not significant, multicollinearity may be present (Gunst, Mason & Weber 1975).

Another technique suggested by Gunst, Mason & Weber (1975) for detection of multicollinearity is computing the determinant of the $(X^*)'X^*$ matrix. Because $(X^*)'X^*$ is in correlation form, the determinant is bounded between zero and one. The closer the value of the determinent is to zero the stronger the evidence that multicollinearity exists. When the determinant is equal to zero there is an exact linear relationship between two or more of the regressors. If the determinant equals one, the column vectors of $(X^*)'X^*$ are orthogonal and therefore independent.

The coefficient of determinantion, R_j^2, is calculated by regressing the j^{th} column vector of X^*, (x_j^*), on the $p-1$ remaining column vectors. If x_j is close to a linear combination of the remaining $p-1$ vectors then R_j^2 will be large, close to one. Therefore, R_j^2 values for $j = \{1...p\}$ can be used to diagnose multicollinearity. A different coefficient of determination, $R_{(j)}^2$, which is calculated by regressing the response y^* on all regressor variables except x_j^*, can also be used as a multicollinearity diagnostic. This method is performed by comparing $R_{(j)}^2$ against R^2 (the coefficient of determination of y^* regressed on X^*). If $R_{(j)}^2$ is close to

R^2, a problem with multicollinearity may exist. The problem with this method occurs when x_j^* is just a poor explanatory variable of y^* and should not be included in the model.

Another diagnostic of multicollinearity are the eigenvalues of $(X^*)'X^*$. The closer any one of the eigenvalues is to zero the more severe the problem of multicollinearity. According to Myers (1990), the severity of multicollinearity can be measured in terms of the ratio of largest to smallest eigenvalues. This ratio, c, is called the condition number.

$$c = \frac{\lambda_{Max}}{\lambda_{Min}}$$

Large values of c indicate multicollinearity may be a problem. Myers (1990) states a condition number exceeding 1,000 should cause concern about the effects of multicollinearity. In addition, the number of eigenvalues close to zero indicates how many regressor variables are involved in the collinearity.

The last diagnostic covered is the Variance Inflation factor (VIF). Each of the p elements of $\hat{\beta}$ has an associated VIF. Recall the variance of $\hat{\beta}$ is $\sigma^2((X^*)'X^*)^{-1}$. The j^{th} diagonal element of $((X^*)'X^*)^{-1}$ is equal to the factor by which the variance of the j^{th} element of $\hat{\beta}$ is inflated over σ^2. The name VIF comes from the fact that the variance of the elements are inflated by VIF over the amount if there was no collinearity within the columns of the X^* matrix. In the ideal case $((X^*)'X^*)^{-1} = I_p$ and all VIF's are one. The coefficient of variation R_j^2 is related to the VIF as follows:

$$VIF(\hat{\beta}_j) = \frac{1}{(1 - R_j^2).}$$

As linear dependence between x_j^* and the remaining $p-1$ regressors increases, R_j^2 increases

and likewise VIF($\hat{\beta}_j$) increases. Similar to the condition number diagnostic, there is no exact cut off value in which one can say with absolute certainty that multicollinearity is going to be a significant problem. Myers (1990) gives a "rule of thumb" of VIF values in excess of 10 as an indication that multicollinearity is causing problems with the estimation of the regression coefficients.

5 Ridge Regression (RR)

Once multicollinearity is identified as a problem, RR provides an alternative to OLS regression. Recall the RR estimator is $\hat{\beta}_R = ((X^*)'X^* + kI_p)^{-1}(X^*)'y^*$ where $k \geq 0$ and when $k = 0$ the RR estimator is equivalent to the OLS estimator. Hoerl and Kennard (1970) provide some motivation for why RR works. In this section their work is outlined and supplemented by a few more mathematical details.

The RR estimator can be written as a function of the OLS estimator.

$$\begin{aligned}
\hat{\beta}_R &= ((X^*)'X^* + kI_p)^{-1}(X^*)'y^* \\
&= (I_p + k((X^*)'X^*)^{-1})^{-1}((X^*)'X^*)^{-1}(X^*)'y^* \\
&= (I_p + k((X^*)'X^*)^{-1})^{-1}\hat{\beta} \\
&= Z\hat{\beta}.
\end{aligned} \tag{3}$$

To simplify notation we define W as:

$$W = ((X^*)'X^* + kI_p)^{-1}. \tag{4}$$

The relationship between Z and W is:

$$\begin{aligned} Z &= W(X^*)'X^* \\ W^{-1}Z &= (X^*)'X^* \\ W^{-1}Z &= (X^*)'X^* + kI_p - kI_p \\ W^{-1}Z &= W^{-1} - kI_p \\ Z &= I_p - kW. \end{aligned} \tag{5}$$

Next the squared distance between $\hat{\beta}_R$ and β is calculated. This value is a measure of how good $\hat{\beta}_R$ is as an estimator. The expected value of this squared distance is the mean square error (MSE) of the RR estimator.

$$\begin{aligned} L_R^2 &= (\hat{\beta}_R - \beta)'(\hat{\beta}_R - \beta) \\ E[L_R^2] &= E[(Z\hat{\beta} - \beta)'(Z\hat{\beta} - \beta)] \text{ substituting equation 3} \\ &= E[\hat{\beta}'Z'Z\hat{\beta}] - 2\beta'ZE[\hat{\beta}] + \beta'\beta \\ &= E[\hat{\beta}'Z'Z\hat{\beta}] - 2\beta'Z\beta + \beta'\beta \end{aligned}$$

where

$$E[\hat{\beta}'Z'Z\hat{\beta}] = E[(y^*)'X^*((X^*)'X^*)^{-1}Z'Z((X^*)'X^*)^{-1}(X^*)'y^*].$$

To help reduce the above equation an identity from Boik (1995) is used.

Let $y \sim (\mu, \sum)$ and A : nxn matrix of constants. Then,

$$E[y'Ay] = \mathrm{tr}(A\sum) + \mu' A\mu.$$

Note: This identity does not require normality.

So,

$$\begin{aligned} E[\hat{\beta}'Z'Z\beta] &= \mathrm{tr}(X^*((X^*)'X)^{-1}Z'Z((X^*)'X)^{-1}(X^*)'\sigma^2 I_p) \\ &\quad +\beta'(X^*)'X^*((X^*)'X^*)^{-1}Z'Z((X^*)'X^*)^{-1}(X^*)'X^*\beta \\ &= \sigma^2\mathrm{tr}(((X^*)'X^*)^{-1}Z'Z) + \beta'Z'Z\beta. \end{aligned}$$

By substitution,

$$\begin{aligned} E[L_R^2] &= \sigma^2\mathrm{tr}(((X^*)'X^*)^{-1}Z'Z) + \beta'Z'Z\beta - 2\beta'Z\beta + \beta'\beta \\ &= \sigma^2\mathrm{tr}(((X^*)'X^*)^{-1}Z'Z) + \beta'(Z'Z - 2Z + I_p)\beta \\ &= \sigma^2\mathrm{tr}(((X^*)'X^*)^{-1}Z'Z) + \beta'(Z - I_p)^2\beta \text{ since } Z = Z' \\ &= \sigma^2\mathrm{tr}(((X^*)'X^*)^{-1}Z'Z) + \beta'(I_p - kW - I_p)^2\beta \text{ substituting 5} \\ &= \sigma^2\mathrm{tr}(((X^*)'X^*)^{-1}Z'Z + \beta'(-kW)^2\beta \\ &= \sigma^2\mathrm{tr}(((X^*)'X^*)^{-1}Z'Z) + k^2\beta'((X^*)'X^* + kI_p)^{-2}\beta. \end{aligned} \tag{6}$$

Now define the two terms in equation 6 as follows:

$$\gamma_1 = \sigma^2\mathrm{tr}(((X^*)'X^*)^{-1}Z'Z) \tag{7}$$

$$\gamma_2 = k^2\beta'((X^*)'X^* + kI_p)^{-2}\beta. \tag{8}$$

Next examine equation 7.

$$\begin{aligned}
\gamma_1 &= \sigma^2 \text{tr}[((X^*)'X^*)^{-1}(I_p + k((X^*)'X^*)^{-1})^{-1}Z] \\
&= \sigma^2 \text{tr}[(((X^*)'X^*) + kI_p)^{-1}Z] \\
&= \sigma^2 \text{tr}[WZ] \\
&= \sigma^2 \text{tr}[W(I_p - kW)] \text{ substituting equation 5} \\
&= \sigma^2 \text{tr}[W - kW^2] \\
&= \sigma^2 \{\text{tr}[((X^*)'X^* + kI_p)^{-1}] - k\text{tr}[(((X^*)'X^*) + kI_p)^{-2}]\}.
\end{aligned}$$

Recall, λ_i represents the i^{th} eigenvalue of $(X^*)'X^*$. Since W and W^2 are symmetric they are both diagonalizable. This allows for easy computation of their eigenvalues.

The i^{th} eigenvalue of $((X^*)'X^* + kI_p)^{-1}$ is $\frac{1}{(\lambda_i+k)}$.

The i^{th} eigenvalue of $((X^*)'X^* + kI_p)^{-2}$ is $\frac{1}{(\lambda_i+k)^2}$ for $i = \{1, 2, 3, ..., p\}$.

Now write equation 7 in terms of eigenvalues:

$$\gamma_1 = \sigma^2 \sum \frac{\lambda_i}{(\lambda_i + k)^2}. \tag{9}$$

Equation 9 is related to the variance of β_R as demonstrated below.

$$\begin{aligned}
\text{Var}(\hat{\beta}_R) &= \text{Var}(Z\hat{\beta}) \\
&= \text{Var}(Z((X^*)'X^*)^{-1}(X^*)'y^*) \\
&= \sigma^2 Z((X^*)'X^*)^{-1}(X^*)'X^*((X^*)'X^*)^{-1}Z' \\
&= \sigma^2 Z((X^*)'X^*)^{-1}Z'.
\end{aligned}$$

The diagonal elements of $Z((X^*)'X^*)^{-1}Z'$ are the RR VIF's. Next applying the trace operator:

$$\begin{aligned} \text{tr}[\text{Var}(\hat{\beta}_R)] &= \sigma^2 \text{tr}(Z((X^*)'X^*)^{-1}Z') \\ &= \sigma^2 \text{tr}(((X^*)'X^*)^{-1}Z'Z) \\ &= \gamma_1. \end{aligned}$$

Therefore, γ_1 is the sum of the diagonals of $\text{Var}(\hat{\beta}_R)$ and is referred to as the total variance (Hoerl and Kennard 1970). The expected value of L_R^2 is the mean square error of $\hat{\beta}_R$ where γ_2 can be considered a squared bias term.

$$\text{MSE}(\hat{\beta}_R) = \gamma_1 + \gamma_2.$$

Returning to equation 9 it is apparent that γ_1 is a decreasing function of k. As k increases the variance of the RR estimator decreases. In the next section some theorems for γ_1 and γ_2 are presented.

6 Ridge Trace (RT)

It can be shown that equation 9, the squared bias of $\hat{\beta}_R$ is a continuous, monotonically increasing function of k with an upper bound $\beta'\beta$. See Hoerl and Kennard (1970) for proofs. The MSE of $\hat{\beta}_R$ is the sum of two expressions γ_1 and γ_2, where γ_1 is a decreasing function of k and γ_2 is an increasing function of k. The goal is to find a value of k such that the $\text{MSE}(\hat{\beta}_R)$ is less than $\text{MSE}(\hat{\beta})$. It is not obvious that a value of k satisfying these conditions

even exists. Hoerl and Kennard (1970) give a proof of the following theorem:

There exists a $k > 0$ such that $E[L_R^2(k)] < E[L_R^2(0)] = \sigma^2 \text{tr}((X^*)'X^*) = \text{Var}(\hat{\beta})$.

This is a very powerful theorem because it guarantees the existence of a k such that the MSE of the RR estimator is smaller than the MSE of the OLS estimator. It is important to emphasize that this theorem only states that a k exists, but it does not provide any clues on how to find a value of k satisfying this condition. Furthermore, Birkes and Dodge (1993) state that there is no explicit formula for k. Because the MSE($\hat{\beta}_R$) depends on the unknown parameter β, one can not be absolutely certain the value of k chosen is best.

There are many methods for choosing reasonable values of k (Myers 1990). One of the easiest and most intuitive methods is the ridge trace (RT). The RT is a two dimensional graph of $\hat{\beta}_R$ versus k which assists the data analyst with choosing a better estimator (in terms of variance) than OLS. The domain of k is usually in the interval [0...1], but can be larger. In the RT, there is a separate curve for each element of $\hat{\beta}$. When k equals zero the RR estimator reduces to the OLS estimator. When the data are seriously affected by multicollinearity the OLS estimator is unstable since it has a large variance. As the value of k increases, the variances of the regression coefficients decrease, allowing the estimates to become more stable. A value of k is usually chosen when the coefficients have stabilized.

RT is a subjective method because it is dependent upon the analyst to choose the value k that reduces the variance by the desired amount at the expense of introducing bias into the estimate. When choosing k one should try to minimize the variance while maintaining the interpretability of the regression coefficients. In other words, one should not choose a k at the expense of ending up with RR coefficients that do not make sense to a subject matter expert.

An important point worth restating is that the MSE($\hat{\beta}_R$) can not be calculated explicitly because the parameter β is unknown. The variance of the estimator is being reduced at the expense of introducing bias. This is why it is critical that the estimates make sense or "agree with" a subject matter expert. Hoerl and Kennard (1970) list a set of guidelines which will assist the user in choosing k:

- At a certain value of k the system will stabilize and have the general characteristics of an orthogonal system.
- Coefficients will not have unreasonable absolute values with respect to the factor for which they represent rates of change.
- Coefficients with apparently incorrect signs at $k = 0$ will have changed to have the proper sign.
- The residual sum of squares will not have been inflated to an unreasonable value. It will not be large relative to the minimum residual sum of squares or large relative to what would be reasonable variances for the process generating the data.

7 Numerical Example

The sample data in Table 1 was taken from Longley (1967). SAS REG (version 6.11) and MATLAB (version 4.2a) were used to perform the analysis. The explatory data consists of the following six variables:

x_1: GNP Implicit Price Deflator

x_2: GNP

x_3: Unemployment

x_4: Size of Armed Forces

x_5: Non-Institutional Population 14 Years of Age and Over

x_6: Year

The response variable y is total derived employment, in thousands. Prior to any analysis the data were centered and scaled. The six variables were regressed on the response using OLS, the results are shown below.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob> F
Model	6	0.99548	0.16591	366.984	0.0001
Error	10	0.00452	0.00045		
U Total	16	1.00000			

Root MSE	0.02126	R-square	0.9955
Dep Mean	0.00000	Adj R-sq	0.9928
C.V.	3.4047425 E15		

Paramter Estimates:

Variable	DF	Parameter Estimate	Standard Error	T for H_o: Parameter= 0	Prob> $\|T\|$	Variance Inflation
x_1^*	1	0.046282	0.2475	0.187	0.8554	135.5324
x_2^*	1	-1.013746	0.8992	-1.127	0.2859	1788.5135
x_3^*	1	-0.537543	0.1233	-4.360	0.0014	33.6189
x_4^*	1	-0.204741	0.0403	-5.083	0.0005	3.5889
x_5^*	1	-0.101221	0.4248	-0.238	0.8165	399.1510
x_6^*	1	2.479664	0.5858	4.233	0.0017	758.9806

Note the negative signs on the coefficients for x_2^* (GNP) and x_5^* (population). One would expect as these variables increased the derived employment would also increase, but with this model just the opposite happens. The data were then checked for the presence of any collinearity. Five of the six VIF's exceed the upper limit value of 10, which is an indication that multicollinearity is adversely affecting the OLS coefficient estimates. VIF's this large reduce our confidence that the regression estimates are correct. Standard errors for x_1^* and x_5^* are large compared to the corresponding coefficients and neither is significant. A possible reason for this are the small eigenvalues of the $(X^*)'X^*$ matrix. The eigenvalues for the $(X^*)'X^*$ matrix are:

4.60338

1.17534

0.20343

0.01493

0.00255

0.00038

Notice the last three eigenvalues are extremely small compared to the others, which causes the variances of the estimates to be inflated. The condition number

$$c = \frac{4.60338}{0.00038} \approx 12,000$$

is over ten times larger than the recommended upper bound. Consistent with the VIF's this condition number is another strong indication that multicollinearity is present in the data. The correlation matrix, $(X^*)'X^*$ shown below is also consistent with the condition number and VIF diagnostics by indicating there is a strong pairwise linear relationship between some of the regressors.

	x_1^*	x_2^*	x_3^*	x_4^*	x_5^*	x_6^*
x_1^*	1.0000					
x_2^*	0.9916	1.0000				
x_3^*	0.6206	0.6043	1.0000			
x_4^*	0.4647	0.4464	-0.1774	1.0000		
x_5^*	0.9792	0.9911	0.6877	0.3644	1.0000	
x_6^*	0.9911	0.9953	0.6683	0.4172	0.9940	1.0000

Because of the obvious multicollinearity problem in this data set, RR is a viable alternative to OLS regression. Figure 1 is a RT graph of the standardized data over the interval [0,1]. It is apparent most of the action within the graph occurs for values of k less than 0.1. For k values greater than 0.1 RR estimates seem to stabilize. In order to get a more detailed graph the RT is plotted again in Figure 2 but the k axis is magnified. From this graph the regression coefficients stabilize between 0.008 and 0.02. Table 2 lists coefficient estimates, VIF's, standard error eigenvalues and condition numbers for eight k values. At $k = 0.018$ all VIF's are below the recommend upper bound and the algebraic sign on coefficients for x_2^* and x_5^* are positive. So $k = 0.018$ would be a reasonable choice for the shrinkage parameter.

This example has demonstrated that when multicollinearity is present in the data, the RR gave a better estimate of the unknown parameter than the OLS estimate.

8 References

1. Birkes, D. and Dodge, Y. (1993) *Alternative Methods of Regression.* John Wiley and Sons, Inc., New York, New York.

2. Boik, R.J. (Fall Semester 1995) STAT 505, Linear Models Lecture Notes, at MSU.

3. Brown, P.J. 1977) Centering and Scaling in Ridge Regression. *Technometrics* 19, 35-36.

4. Gunst, R.F., Mason, R. L. and Webster, J. T. (1975) Regression Analysis and Problems of Multicollinearity. *Communications in Statistics*, 4, 277-291.

5. Hoerl, A. E. and Kennard R. W. Ridge Regression (1988) *Encyclopedia of Statistical Sciences* 8, 129-136. John Wiley and Sons, Inc., New York, New York.

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

7. Longley, J. W. (1967) An appraisal of Least Squares Programs for the Electronic Computer from the point of View of the User. *Journal of the American Statistical Association* 62, 819-841.

8. Mullet, G. M. (1977) Why Regression Coefficients Have the Wrong Sign. *Journal of Quality Technology.* 8, 121-126

9. Myers, R. A. (1990) *Classical and Modern Regression with Applications.* Wadsworth, Inc. Belmont California.

10. Searle, S. (1982) *Matrix Algebra Useful for Statistics.* John Wiley and Sons, Inc., New York, New York.

Figure 1

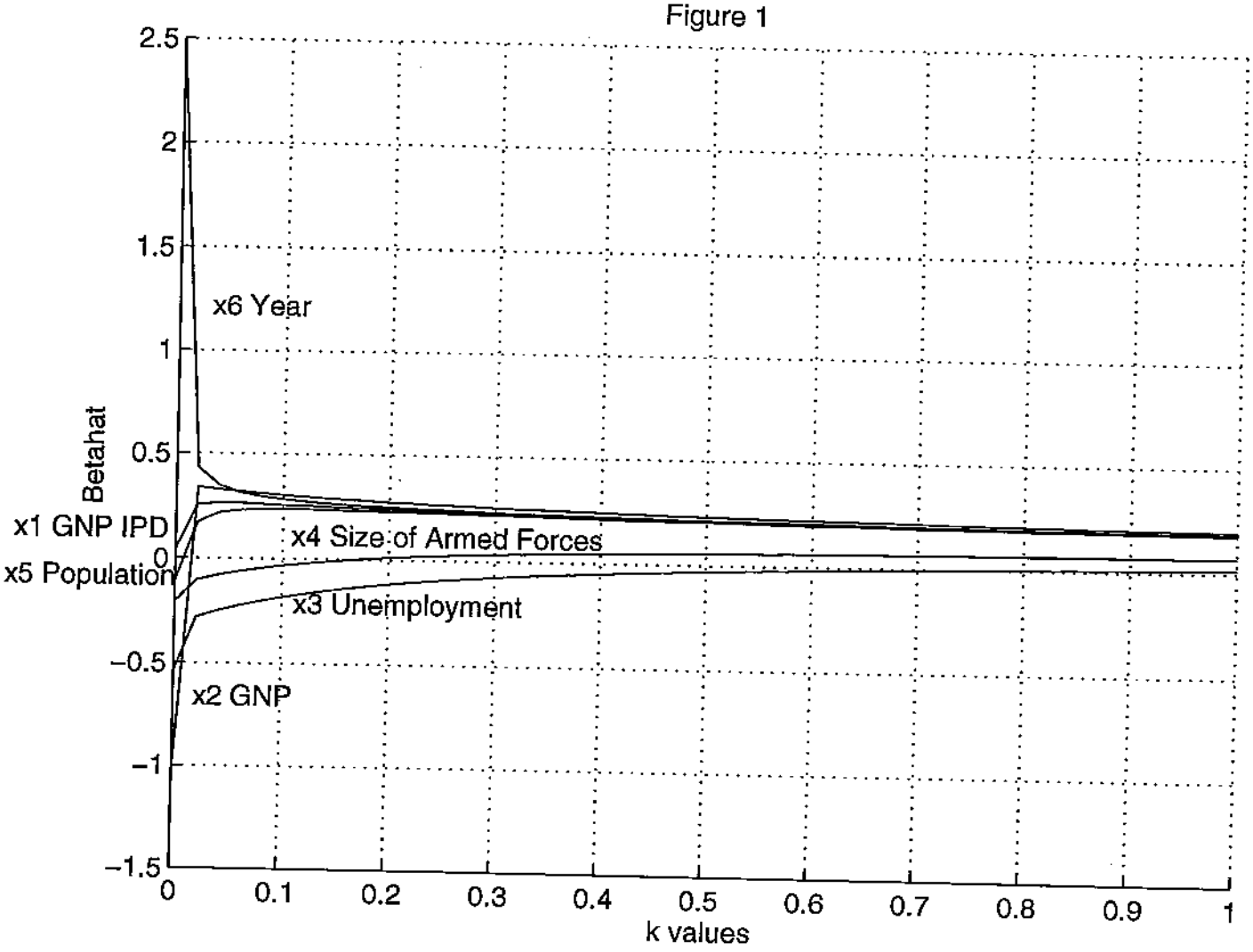
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
Betahat
x6 Year
x1 GNP IPD
x5 Population
x4 Size of Armed Forces
x3 Unemployment
x2 GNP
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
k values

Figure 2

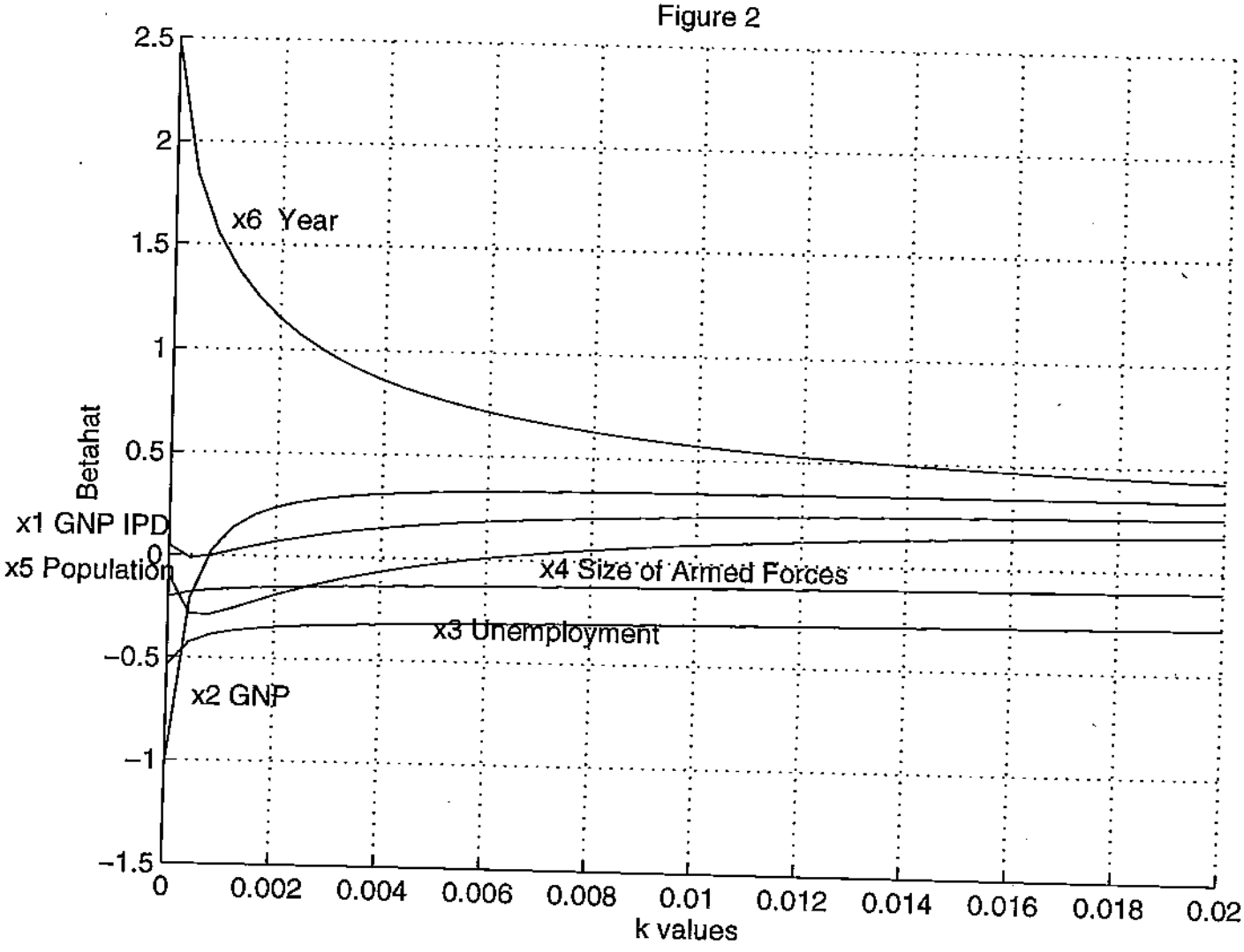
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
Betahat
x6 Year
x1 GNP IPD
x5 Population
x4 Size of Armed Forces
x3 Unemployment
x2 GNP
0
0.002
0.004
0.006
0.008
0.01
0.012
0.014
0.016
0.018
0.02
k values

Table 1

x_1	x_2	x_3	x_4	x_5	x_6	y
83	234289	2356	1590	107608	1947	60323
88.5	259426	2325	1456	108632	1948	61122
88.2	258054	3682	1616	109773	1949	60171
89.5	284599	3351	1650	110929	1950	61187
96.2	328975	2099	3099	112075	1951	63221
98.1	346999	1932	3594	113270	1952	63639
99	365385	1870	3547	115094	1953	64989
100	363112	3578	3350	116219	1954	63761
101.2	397469	2904	3048	117388	1955	66019
104.6	419180	2822	2857	118734	1956	67857
108.4	442769	2936	2798	120445	1957	68169
110.8	444546	4681	2637	121950	1958	66513
112.6	482704	3813	2552	123366	1959	68655
114.2	502601	3931	2514	125368	1960	69564
115.7	518173	4806	2572	127852	1961	69331
116.9	554894	4007	2827	130081	1962	70551

Table 2

	$k = 0$				$k = 0.006$			
Variable	Coefficient	Standard Error	VIF	Eigenvalue	Coefficient	Standard Error	VIF	Eigenvalue
x_1^*	0.0463	0.2475	135.5324	0.0149	0.1849	0.1586	25.7329	0.0209
x_2^*	-1.0137	0.8992	1788.514	0.0026	0.3279	0.0878	7.8972	0.0086
x_3^*	-0.5375	0.1233	33.6189	0.0004	-0.3147	0.0532	2.8932	0.0064
x_4^*	-0.2047	0.0403	3.5889	0.2034	-0.1311	0.0467	2.2352	0.2094
x_5^*	-0.1012	0.4248	399.151	1.1753	0.0076	0.1515	23.4806	1.1813
x_6^*	2.4797	0.5858	758.9806	4.6034	0.7148	0.1456	21.6995	4.6094
	Condition Number = 11509				Condition Number = 720			

Table 2 continued

	$k = 0.008$				$k = 0.010$			
Variable	Coefficient	Standard Error	VIF	Eigenvalue	Coefficient	Standard Error	VIF	Eigenvalue
x_1^*	0.2089	0.1475	20.7173	0.0229	0.2244	0.1377	17.164	0.0249
x_2^*	0.3353	0.0719	4.9287	0.0106	0.3384	0.0618	3.4542	0.0126
x_3^*	-0.3072	0.0542	2.7995	0.0084	-0.3013	0.0549	2.731	0.0104
x_4^*	-0.1252	0.0477	2.1636	0.2114	-0.1203	0.0482	2.1043	0.2134
x_5^*	0.0561	0.1356	17.5178	1.1833	0.09	0.1234	13.7889	1.1853
x_6^*	0.6274	0.122	14.1651	4.6114	0.5685	0.1051	9.9957	4.6134
	Condition Number = 549				Condition Number = 444			

Table 2 continued

	$k = 0.012$				$k = 0.016$			
Variable	Coefficient	Standard Error	VIF	Eigenvalue	Coefficient	Standard Error	VIF	Eigenvalue
x_1^*	0.235	0.1291	14.5096	0.0269	0.2481	0.1147	10.826	0.0309
x_2^*	0.3397	0.0547	2.608	0.0146	0.34	0.0456	1.7118	0.0186
x_3^*	-0.2962	0.0554	2.6737	0.0124	-0.2873	0.0559	2.5748	0.0164
x_4^*	-0.1161	0.0486	2.0532	0.2154	-0.1089	0.0489	1.9669	0.2194
x_5^*	0.115	0.1136	11.2419	1.873	0.1492	0.0986	8.0057	1.1913
x_6^*	0.526	0.0925	7.4446	4.6154	0.4686	0.0748	4.6094	4.6194
	Condition Number = 372				Condition Number = 282			

Table 2 continued

	$k = 0.018$				$k = 0.02$			
Variable	Coefficient	Standard Error	VIF	Eigenvalue	Coefficient	Standard Error	VIF	Eigenvalue
x_1^*	0.2523	0.1087	9.5066	0.0329	0.2554	0.1033	8.4214	0.0349
x_2^*	0.3396	0.0425	1.4543	0.0206	0.339	0.04	1.2635	0.0226
x_3^*	-0.2833	0.0561	2.5299	0.0184	-0.2795	0.0561	2.4871	0.0204
x_4^*	-0.1057	0.049	1.9293	0.2214	-0.1027	0.049	1.8944	0.2234
x_5^*	0.1614	0.0927	6.9222	1.1933	0.1714	0.0876	6.0582	1.1953
x_6^*	0.4482	0.0685	3.7731	4.6214	0.4314	0.0632	3.1501	4.6234
	Condition Number = 251				Condition Number = 227			